INTRODUCTION TO THE SERIES

The aim of the *Handbooks in Economics* series is to produce Handbooks for various branches of economics, each of which is a definitive source, reference, and teaching supplement for use by professional researchers and advanced graduate students. Each Handbook provides self-contained surveys of the current state of a branch of economics in the form of chapters prepared by leading specialists on various aspects of this branch of economics. These surveys summarize not only received results but also newer developments, from recent journal articles and discussion papers. Some original material is also included, but the main goal is to provide comprehensive and accessible surveys. The Handbooks are intended to provide not only useful reference volumes for professional collections but also possible supplementary readings for advanced courses for graduate students in economics.

KENNETH J. ARROW and MICHAEL D. INTRILIGATOR

CONTENTS OF THE HANDBOOK

VOLUME 1

PART I: PRIMARY ELEMENTS OF THE LEGAL SYSTEM

PREFACE

Law can be viewed as a body of rules and legal sanctions that channel behavior in socially desirable directions—for example, by encouraging individuals to take proper precautions to prevent accidents or by discouraging competitors from colluding to raise prices. The incentives created by the legal system are thus a natural subject of study by economists. Moreover, given the importance of law to the welfare of societies, the economic analysis of law merits prominent treatment as a subdiscipline of economics. Our hope is that this two volume Handbook will foster the study of the legal system by economists.

The origins of law and economics may be traced to eighteenth century writings on crime by Beccaria (1767) and Bentham (1789). The modern incarnation of the field dates from the 1960s: Coase (1960) on property rights, externalities, and bargaining; Calabresi (1961, 1970) on liability rules and accident law; Demsetz (1967) on the emergence of property rights; and Becker (1968) on crime. Of great significance was Posner (1972), the first application of economic analysis to the body of law as a whole (Posner also authored numerous influential articles on specific legal topics).

This early writing in law and economics was mainly informal and emphasized basic subject areas of law. Later scholarship began to include formal work, notably by Brown (1973), Diamond (1974), Spence (1977), and Shavell (1980a) on liability rules and accidents, Polinsky (1979) on property rights and liability rules, Barton (1972) and Shavell (1980b) on contract law, Landes (1971) and Gould (1973) on litigation behavior, and, following Becker (1968), Polinsky and Shavell (1979) on law enforcement. Law and economics scholarship also expanded into other subject areas, with corporate law receiving the most attention—see, for example, early contributions by Bebchuk (1985), Easterbrook and Fischel (1991) (synthesizing previously written articles), Gilson (1981), and Manne (1965). Additionally, empirical research was undertaken, initially mostly in the area of crime, and subsequently in many other fields as well, especially corporate law.

The purpose of this Handbook is to provide economists with a systematic introduction to and survey of research in the field of law and economics. The Handbook contains 22 chapters and is organized into three main parts. Part I deals with the building blocks of the legal system: property law; contract law; accident law (torts); litigation (including aspects of civil procedure); and public enforcement of law (including criminal law). Part II treats other prominent areas of law: corporate law; bankruptcy law; antitrust law; regulation (of externalities, natural monopolies, and network industries); employment and labor law; antidiscrimination law; intellectual property law; environmental law; and international law. Part III addresses three additional topics: norms and the law; the experimental study of law; and political economy and the law. Most of the chapters are

theoretically-oriented, but many mention relevant empirical work and three focus on empirical research (on civil law, public law enforcement, and corporate law).

The first volume of the Handbook includes all of Part I and several chapters from Part II. The second volume contains the remaining chapters of Part II and all of Part III.

We are grateful to Kenneth Arrow and Michael Intriligator for encouraging the development of the Handbook and for their substantive suggestions about it; to Valerie Teng and Mark Newson of Elsevier for their able assistance with the administrative tasks associated with its production; and to the John M. Olin Foundation, through our respective institutions' law and economics programs, for supporting our preparation of it.

A. Mitchell Polinsky & Steven Shavell

References

Barton, J.H. (1972). "The economic basis of damages for breach of contract". Journal of Legal Studies 1, 277–304.

Bebchuk, L.A. (1985). "Toward undistorted choice and equal treatment in corporate takeovers". Harvard Law Review 98, 1695–1808.

Beccaria, C. (1767). On Crimes and Punishments, and Other Writings. Bellamy, R. (Ed.), Translator Davies, R. et al. Cambridge University Press, New York (1995).

Becker, G.S. (1968). "Crime and punishment: an economic approach". Journal of Political Economy 76, 169–217.

Bentham, J. (1789). "An Introduction to the Principles of Morals and Legislation". In: *The Utilitarians*. Anchor Books, Garden City, N.Y. (1973).

Brown, J.P. (1973). "Toward an economic theory of liability". Journal of Legal Studies 2, 323–349.

Coase, R.H. (1960). "The problem of social cost". Journal of Law and Economics 3, 1–44.

Calabresi, G. (1961). "Some thoughts on risk distribution and the law of torts". Yale Law Journal 70, 499–553.

Calabresi, G. (1970). The Costs of Accidents: A Legal and Economic Analysis. Yale University Press, New Haven.

Demsetz, H. (1967). "Toward a theory of property rights". American Economic Review: Papers and Proceedings 57, 347–359.

Diamond, P.A. (1974). "Single activity accidents". Journal of Legal Studies 3, 107–164.

Easterbrook, F.H., Fischel, D.R. (1991). The Economic Structure of Corporate Law. Harvard University Press, Cambridge.

Gilson, R. (1981). "A structural approach to corporations: the case against defensive tactics in tender offers". Stanford Law Review 33, 819–891.

Gould, J.P. (1973). "The economics of legal conflicts". Journal of Legal Studies 2, 279–300.

Landes, W.M. (1971). "An economic analysis of the courts". Journal of Law and Economics 14, 61–107.

Manne, H.G. (1965). "Mergers and the market for corporate control". Journal of Political Economy 73, 110–120.

Polinsky, A.M. (1979). "Controlling externalities and protecting entitlements: property right, liability rule, and tax subsidy approaches". Journal of Legal Studies 8, 1–48.

Polinsky, A.M., Shavell, S. (1979). "The optimal tradeoff between the probability and magnitude of fines". American Economic Review 69, 880–891.

Posner, R.A. (1972). Economic Analysis of Law, 1st edn. Little, Brown and Company, Boston.

Shavell, S. (1980a). "Strict liability versus negligence". Journal of Legal Studies 9, 1–25.

Shavell, S. (1980b). "Damage measures for breach of contract". Bell Journal of Economics 11, 466–490.

Spence, A.M. (1977). "Consumer misperceptions, product failure and producer liability". Review of Economic Studies 44, 561–572.

CONTENTS OF VOLUME 2

Chapter 12

CORPORATE LAW AND GOVERNANCE*

MARCO BECHT

ECARES, Université Libre de Bruxells and European Corporate Governance Institute (ECGI)

PATRICK BOLTON

Graduate School of Business and Department of Economics, Columbia University

AILSA RÖELL

School of International and Public Affairs, Columbia University

Contents

* An earlier version of this chapter appeared under the title *Corporate Governance and Control* in the **Handbook of the Economics of Finance**, edited by G.M. Constantinides, M. Harris and R. Stulz, 2003 Elsevier B.V. Substantive new material is confined to Section 8.

Handbook of Law and Economics, Volume 2
Edited by A. Mitchell Polinsky and Steven Shavell

DOI: 10.1016/S1574-0730(07)02012-9

Abstract

This chapter surveys the theoretical and empirical research on the main mechanisms of corporate law and governance, discusses the main legal and regulatory institutions in different countries, and examines the comparative governance literature. Corporate governance is concerned with the reconciliation of conflicts of interest between various corporate claimholders and the resolution of collective action problems among dispersed investors. A fundamental dilemma of corporate governance emerges from this overview: large shareholder intervention needs to be regulated to guarantee better small investor protection; but this may increase managerial discretion and scope for abuse. Alternative methods of limiting abuse have yet to be proven effective.

Keywords

Corporate governance, ownership, takeovers, block holders, boards

JEL classification: G32, G34

1. Introduction

At the most basic level a corporate governance problem arises whenever an outside investor wishes to exercise control differently from the manager in charge of the firm. Dispersed ownership magnifies the problem by giving rise to conflicts of interest between the various corporate claimholders and by creating a collective action problem among investors.[1]

Most research on corporate governance has been concerned with the resolution of this collective action problem. Five alternative mechanisms may mitigate it: (i) partial concentration of ownership and control in the hands of one or a few large investors, (ii) hostile takeovers and proxy voting contests, which concentrate ownership and/or voting power temporarily when needed, (iii) delegation and concentration of control in the board of directors, (iv) alignment of managerial interests with investors through executive compensation contracts, and (v) clearly defined fiduciary duties for CEOs together with class-action suits that either block corporate decisions that go against investors' interests, or seek compensation for past actions that have harmed their interests.

In this survey we review the theoretical and empirical research on these five main mechanisms and discuss the main legal and regulatory institutions of corporate governance in different countries. We discuss how different classes of investors and other constituencies can or ought to participate in corporate governance. We also review the comparative corporate governance literature.[2]

The favoured mechanism for resolving collective action problems among shareholders in most countries appears to be partial ownership and control concentration in the hands of large shareholders.[3] Two important costs of this form of governance have been emphasised: (i) the potential collusion of large shareholders with management against smaller investors and, (ii) the reduced liquidity of secondary markets. In an attempt to boost stock market liquidity and limit the potential abuse of minority shareholders some countries' corporate law drastically curbs the power of large shareholders.[4] These countries rely on the board of directors as the main mechanism for co-ordinating shareholder actions. But boards are widely perceived to be ineffective.[5] Thus, while minority shareholders get better protection in these countries, managers may also have greater discretion.

[1] See Zingales (1998) for a similar definition.

[2] We do not cover the extensive strategy and management literature; see Pettigrew, Thomas, and Whittington (2002) for an overview, in particular Davis and Useem (2002).

[3] See ECGN (1997), La Porta, Lopez-de-Silanes, and Shleifer (1999), Claessens, Djankov, and Lang (2000) and Barca and Becht (2001) for evidence on control concentration in different countries.

[4] Black (1990) provides a detailed description of the various legal and regulatory limits on the exercise of power by large shareholders in the U.S. Wymeersch (2003) discusses legal impediments to large shareholder actions outside the U.S.

[5] Gilson and Kraakman (1991) provide analysis and an agenda for board reform in the U.S. against the background of a declining market for corporate control and scattered institutional investor votes.

In a nutshell, the fundamental issue concerning governance by shareholders today seems to be how to regulate large or active shareholders so as to obtain the right balance between managerial discretion and small shareholder protection. Before exploring in greater detail the different facets of this issue and the five basic mechanisms described above, it is instructive to begin with a brief overview of historical origins and early writings on the subject.

2. Historical origins: A brief sketch

The term "corporate governance" derives from an analogy between the government of cities, nations or states and the governance of corporations.[6] The early corporate finance textbooks saw "representative government" (Mead, 1928, p. 31) as an important advantage of the corporation over partnerships but there has been and still is little agreement on how representative corporate governance really is, or whom it should represent.

2.1. How representative is corporate government?

The institutional arrangements surrounding corporate elections and the role and fiduciary duties of the board have been the central themes in the corporate governance literature from its inception. The dilemma of how to balance limits on managerial discretion and small investor protection is ever present. Should one limit the power of corporate plutocrats (large shareholders or voting trusts) or should one tolerate concentrated voting power as a way of limiting managerial discretion?

The concern of early writers of corporate charters was the establishment of "corporate suffrage", where each member (shareholder) had one vote (Dunlavy, 1998). The aim was to establish "democracy" by eliminating special privileges of some members and by limiting the number of votes each shareholder could cast, irrespective of the number of shares held.[7] However, just as "corporate democracy" was being established it was already being transformed into "plutocracy" by moving towards "one-share-one-vote" and thus allowing for concentrated ownership and control (Dunlavy, 1998).[8]

[6] The analogy between corporate and political voting was explicit in early corporate charters and writings, dating back to the revolutionary origins of the American corporation and the first railway corporations in Germany (Dunlavy, 1998). The precise term "corporate governance" itself seems to have been used first by Richard Eells (1960, p. 108), to denote "the structure and functioning of the corporate polity".

[7] Frequently voting scales were used to achieve this aim. For example, under the voting scale imposed by a Virginia law of 1836 shareholders of manufacturing corporations cast "one vote for each share up to 15, one vote for every five shares from 15 to 100, and one vote for each increment of 20 shares above 100 shares" (Dunlavy, 1998, p. 18).

[8] Voting right restrictions survived until very recently in Germany (Franks and Mayer, 2001). They are still in use in Denmark, France, Spain and other European countries (Becht and Mayer, 2001).

In the U.S. this was followed by two distinct systems of "corporate feudalism": first, to the voting trusts[9] and holding companies[10] (Cushing, 1915; Mead, 1903; Liefmann, 1909, 1920) originating in the "Gilded Age" (Twain and Warner, 1873)[11] and later to the managerial corporation.[12] The "captains of industry" in the trusts and hierarchical groups controlled the majority of votes in vast corporate empires with relatively small(er) amounts of capital, allowing them to exert product market power and leaving ample room for self-dealing.[13] In contrast, the later managerial corporations were controlled mainly by professional managers and most of their shareholders were too small and numerous to have a say. In these firms control was effectively separated from ownership.[14]

Today corporate feudalism of the managerial variety in the U.S. and the "captain of industry" kind elsewhere is challenged by calls for more "shareholder democracy", a global movement that finds its roots with the "corporate Jacksonians" of the 1960s in the U.S.[15]

[9] Under a typical voting trust agreement shareholders transfer their shares to a trust and receive certificates in return. The certificate holders elect a group of trustees who vote the deposited shares. Voting trusts were an improvement over pooling agreements and designed to restrict product market competition. They offered two principal advantages: putting the stock of several companies into the voting trust ensured that the trustees had permanent control over the management of the various operating companies, allowing them to enforce a common policy on output and prices; the certificates issued by the voting trust could be widely placed and traded on a stock exchange.

[10] Holding companies have the purpose of owning and voting shares in other companies. After the passage of the Sherman Antitrust Act in 1890 many of the voting trusts converted themselves into New Jersey registered holding companies ("industrial combinations") that were identical in function, but escaped the initial round of antitrust legislation, for example the Sugar Trust in 1891 (Mead, 1903, p. 44) and Rockefeller's Standard Oil in 1892 (Mead, 1903, p. 35).

[11] The "captains of industry" of this era, also referred to as the "Robber Barons" (Josephson, 1934; DeLong, 1998), were the target of an early anti-trust movement that culminated in the election of Woodrow Wilson as U.S. President in 1912. Standard Oil was broken up even before (in 1911) under the Sherman Act of 1890 and converted from a corporation that was tightly controlled by the Rockefeller clan to a managerial corporation. Trust finance disappeared from the early corporate finance textbooks (for example Mead, 1912 versus Mead, 1928). In 1929 Rockefeller Jr. (14.9%) ousted the scandal ridden Chairman of Standard Oil of Indiana, who enjoyed the full support of his board, only by a small margin, an example that was widely used for illustrating how much the balance of power had swung from the "Robber Barons" to management (Berle and Means, 1932, pp. 82–83, cited in Galbraith, 1967), another type of feudal lord.

[12] For Berle and Means (1930): "[the] "publicly owned" stock corporation in America... constitutes an institution analogous to the feudal system in the Middle Ages".

[13] They also laid the foundations for some of the World's finest arts collections, philanthropic foundations and university endowments.

[14] This "separation of ownership and control" triggered a huge public and academic debate of "the corporate problem"; see, for example, the Berle and Means symposia in the Columbia Law Review (1964) and the Journal of Law and Economics (1983). Before Means (1931a, 1931b) and Berle and Means (1930, 1932) the point was argued in Lippmann (1914), Veblen (1923), Carver (1925), Ripley (1927) and Wormser (1931); see Hessen (1983).

[15] Non-Americans often consider shareholder activism as a free-market movement and associated calls for more small shareholder power as a part of the conservative agenda. They are puzzled when they learn that

As an alternative to shareholder activism some commentators in the 1960s proposed for the first time that hostile takeovers might be a more effective way of disciplining management. Thus, Rostow (1959) argued, "the raider persuades the stockholders for once to act as if they really were stockholders, in the black-letter sense of the term, each with the voice of partial ownership and a partial owner's responsibility for the election of directors" (1959, p. 47). Similarly, Manne (1964) wrote, "vote selling [. . .] negatives many of the criticisms often levelled at the public corporation" [1964, p. 1445]. As we shall see, the abstract "market for corporate control" has remained a central theme in the corporate governance literature.

2.2. Whom should corporate government represent?

The debate on whether management should run the corporation solely in the interests of shareholders or whether it should take account of other constituencies is almost as old as the first writings on corporate governance. Berle (1931) held the view that corporate powers are powers in trust for shareholders and nobody else.[16] But, Dodd (1932) argued that: "[business] is private property only in the qualified sense, and society may properly demand that it be carried on in such a way as to safeguard the interests of those who deal with it either as employees or consumers even if the proprietary rights of its owners are thereby curtailed" (Dodd, 1932, p. 1162). Berle (1932) disagreed on the grounds that responsibility to multiple parties would exacerbate the separation of ownership and control and make management even less accountable to shareholders.[17]

There is nowadays a voluminous literature on corporate governance. On many key issues our understanding has improved enormously since the 1930s. Remarkably though, some of the main issues over which the early writers have been debating remain central today.

3. Why corporate governance is currently such a prominent issue

Why has corporate governance become such a prominent topic in the past two decades or so and not before? We have identified, in no particular order, the following reasons:

shareholder activism today has its roots in part of the anti-Vietnam War, anti-apartheid and anti-tobacco movements and has close links with the unions. In terms of government (of corporations) there is no contradiction. The "corporate Jacksonians", as a prominent critic called them (Manning, 1958, p. 1489), are named after the 7th U.S. President (1829–1937) who introduced universal male suffrage and organised the U.S. Democratic Party that has historically represented minorities, labour and progressive reformers (Encyclopaedia Britannica, *Jackson, Andrew; Democratic Party*).

[16] Consequently "all powers granted to a corporation or to the management of a corporation, or to any group within the corporation, whether derived from statute or charter or both, are necessarily and at all times exercisable only for the ratable benefit of all the shareholders as their interest appears", Berle (1931).

[17] He seems to have changed his mind some twenty years later as he wrote that he was "squarely in favour of Professor Dodd's contention" [Berle (1954)]. For a comprehensive account of the Berle-Dodd dialogue see Weiner (1964) and for additional papers arguing both points of view Mason (1959). Galbraith (1967) in his influential "The New Industrial State" took Dodd's position.

(i) the world-wide wave of privatisation of the past two decades, (ii) pension fund reform and the growth of private savings, (iii) the takeover wave of the 1980s, (iv) deregulation and the integration of capital markets, (v) the 1998 East Asia crisis, which has put the spotlight on corporate governance in emerging markets (vi) a series of recent U.S. scandals and corporate failures that built up but did not surface during the bull market of the late 1990s.

3.1. The world-wide privatisation wave

Privatisation has been an important phenomenon in Latin America, Western Europe, Asia and (obviously) the former Soviet block, but not in the U.S. where state ownership of enterprises has always been very small. On average, since 1990 OECD privatisation programmes have generated proceeds equivalent to 2.7% of total GDP, and in some cases up to 27% of country GDP. The privatisation wave started in the U.K., which was responsible for 58% of OECD and 90% of European Community privatisation proceeds in 1991. Since 1995 Australia, Italy, France, Japan and Spain alone have generated 60% of total privatisation revenues.

Inevitably, the privatisation wave has raised the issue of how the newly privatised corporations should be owned and controlled. In some countries, most notably the U.K., part of the agenda behind the massive privatisation program was to attempt to recreate a form of "shareholder democracy"[18] (see Biais and Perotti, 2002). In other countries great care was taken to ensure the transfer of control to large shareholders. The issues surrounding the choice of privatisation method rekindled interest in governance issues; indeed Shinn (2001) finds that the state's new role as a public shareholder in privatised corporations has been an important source of impetus for changes in corporate governance practices worldwide. In general, privatisations have boosted the role of stock markets as most OECD sales have been conducted via public offerings, and this has also focused attention on the protection of small shareholders.

3.2. Pension funds and active investors

The growth in defined contribution pension plans has channelled an increasing fraction of household savings through mutual and pension funds and has created a constituency of investors that is large and powerful enough to be able to influence corporate governance. Table 1 illustrates how the share of financial assets controlled by institutional investors has steadily grown over the 1990s in OECD countries. It also highlights the disproportionately large institutional holdings in small countries with large financial centres, like Switzerland, the Netherlands and Luxembourg. Institutional investors in

[18] A state-owned and -controlled company is indirectly owned by the citizens via the state, which has a say in the affairs of the company. In a "shareholder democracy" each citizen holds a small share in the widely held company, having a direct interest and—theoretically—say in the affairs of the company.

Table 1
Financial assets of institutional investors in OECD countries

	Value assets billion U.S.$		Asset growth 1990–1996	% Total OECD assets 1996	Assets as % GDP		% Pension funds 1996	% Insurance companies 1996	% Invest. companies 1996	% of Assets in equity 1996	% OECD equity 1996
	1990	1996			1990	1996					
Australia	145.6	331.1	127.4	1.3	49.3	83.8	36.3	46.0	14.1	52	1.9
Austria	38.8	90.1	132.2	0.3	24.3	39.4	3.0	53.3	43.7	8	0.1
Belgium	87.0	169.1	94.4	0.7	44.4	63	6.5	49.0	41.0	23	0.4
Canada	332.8	560.5	68.4	2.2	58.1	94.6	43.0	31.4	25.7	9	0.6
Czech Republic	–	(1994) 7.3	–	–	–	–	–	–	–		<0.1
Denmark	74.2	123.5	66.4	0.5	55.6	67.1	25.2	67.2	7.6	31	0.4
Finland	44.7	71.2	59.3	0.3	33.2	57	–	24.6	3.4	23	0.2
France	655.7	1,278.1	94.9	4.9	54.8	83.1		55.2	44.8	26	3.7
Germany	599.0	1,167.9	95.0	4.5	36.5	49.9	5.5	59.2	35.3	14	1.8
Greece	5.4	35.1	550.0	0.1	6.5	28.5	41.6	12.3	46.2	6	<0.1
Hungary	–	2.6	–	<0.1		5.7	–	65.4	26.9	6	<0.1
Iceland	2.9	5.8	100.0	<0.1	45.7	78.7	79.3	12.1	8.6	6	<0.1
Italy	146.6	484.6	230.6	1.9	13.4	39.9	8.1	30.1	26.6	12	0.6
Japan	2,427.9	3,563.6	46.8	13.7	81.7	77.6	–	48.9	12.6	21	8.3
Korea	121.9	277.8	127.9	1.1	48	57.3	4.9	43.4	51.7	12	0.4
Luxembourg	95.9	392.1	308.9	1.5	926.8	2139.1	0.8	–	99.2		<0.1
Mexico	23.1	14.9	−35.5	0.1	8.8	4.5		32.9	67.1	17	<0.1
Netherlands	378.3	671.2	77.4	2.6	133.4	169.1	55.2	33.5	9.9	28	2.1

Table 1
(*Continued*)

	Value assets billion U.S.$		Asset growth 1990–1996	% Total OECD assets 1996	Assets as % GDP		% Pension funds 1996	% Insurance companies 1996	% Invest. companies 1996	% of Assets in equity 1996	% OECD equity 1996
	1990	1996			1990	1996					
New Zealand	–	24.9	–	0.1	–	38.1	–	31.7	17.3	37	0.1
Norway	41.5	68.6	65.3	0.3	36	43.4	14.9	70.1	15.0	20	0.2
Poland	–	2.7	–	<0.1	–	2	–	81.5	18.5	23	<0.1
Portugal	6.2	37.5	504.8	0.1	9	34.4	26.4	27.2	45.1	9	<0.1
Spain	78.9	264.5	235.2	1.0	16	45.4	4.5	41.0	54.5	6	0.2
Sweden	196.8	302.9	53.9	1.2	85.7	120.3	2.0	47.3	19.8	40	1.4
Switzerland	271.7	449.8	65.6	1.7	119	77.3	49.3	40.2	10.5	24	1.2
Turkey	0.9	2.3	155.6	<0.1	0.6	1.3	–	47.8	52.2	8	<0.1
U.K.	1,116.8	2,226.9	99.4	8.6	114.5	193.1	40.1	45.9	14.0	67	16.6
U.S.	6,875.7	13, 382.1	94.6	51.5	123.8	181.1	35.6	22.6	25.2	40	59.7
Total OECD	15,758.3	26,001.4									
Mean OECD			94.6		49.3	83.8	26.3	33.6	24.9	22	

Source: OECD (1999), *Institutional Investors Statistical Yearbook 1998*, Tables S.1., S.2., S.3., S.4., S.6., S.11 and own calculations.

the U.S. alone command slightly more than 50% of the total assets under management and 59.7% of total equity investment in the OECD, rising to 60.1% and 76.3% respectively when U.K. institutions are added. A significant proportion is held by pension funds (for U.S. and U.K. based funds, 35.1% and 40.1% of total assets respectively). These funds are playing an increasingly active role in global corporate governance. In the U.S. ERISA[19] regulations oblige pension funds to cast the votes in their portfolio responsibly. This has led to the emergence of a service industry that makes voting recommendations and exercises votes for clients. The largest providers now offer global services.

Japanese institutional investors command 13.7% of total institutional investor assets in the OECD but just 8.3% of the equities. These investors are becoming more demanding and they are one of the forces behind the rapid transformation of the Japanese corporate governance system. As a percentage of GDP, the holdings of Italian and German institutional investors are small (39.9% and 49.9% in 1996) and well below the OECD average of 83.8%. The ongoing reform of the pension systems in both countries and changing savings patterns, however, are likely to change this picture in the near future.[20]

3.3. Mergers and takeovers

The hostile takeover wave in the U.S. in the 1980s and in Europe in the 1990s, together with the recent merger wave, has also fuelled the public debate on corporate governance. The successful $199 billion cross-border hostile bid of Vodafone for Mannesmann in 2000 was the largest ever to take place in Europe. The hostile takeovers in Italy (Olivetti for Telecom Italia; Generali for INA) and in France (BNP-Paribas; Elf Aquitaine for Total Fina) have spectacularly shaken up the sleepy corporate world of continental Europe. Interestingly, these deals involve newly privatised giants. It is also remarkable that they have not been opposed by the social democratic administrations in place at the time. Understandably, these high profile cases have moved takeover regulation of domestic and cross-border deals in the European Union to the top of the political agenda.

3.4. Deregulation and capital market integration

Corporate governance rules have been promoted in part as a way of protecting and encouraging foreign investment in Eastern Europe, Asia and other emerging markets.

[19] ERISA stands for the Employee Retirement Income Security Act of 1974.

[20] One note of caution. The figures for Luxemburg and Switzerland illustrate that figures are compiled on the basis of the geographical location of the fund managers, not the origin of the funds under management. Judging from the GDP figures, it is very likely that a substantial proportion of the funds administered in the U.K., the U.S., Switzerland and the Netherlands belong to citizens of other countries. For governance the location of the fund managers matters. They make the investment decisions and have the power to vote the equity in their portfolios and the sheer size of the numbers suggests that fund governance is a topic in its own right.

The greater integration of world capital markets (in particular in the European Union following the introduction of the Euro) and the growth in equity capital throughout the 1990s have also been a significant factor in rekindling interest in corporate governance issues. Increasingly fast growing corporations in Europe have been raising capital from different sources by cross listing on multiple exchanges (Pagano, Röell, and Zechner, 2002). In the process they have had to contend more with U.S. and U.K. pension funds. This has inevitably contributed to the spread of an 'equity culture' outside the U.S. and U.K.

3.5. The 1998 East Asia/Russia/Brazil crisis

The East Asia crisis has highlighted the flimsy protections investors in emerging markets have and put the spotlight on the weak corporate governance practices in these markets. The crisis has also led to a reassessment of the Asian model of industrial organisation and finance around highly centralised and hierarchical industrial groups controlled by management and large investors. There has been a similar reassessment of mass insider privatisation and its concomitant weak protection of small investors in Russia and other transition economies.

The crisis has led international policy makers to conclude that macro-management is not sufficient to prevent crises and their contagion in an integrated global economy. Thus, in South Korea, the International Monetary Fund has imposed detailed structural conditions that go far beyond the usual Fund policy. It is no coincidence that corporate governance reform in Russia, Asia and Brazil has been a top priority for the OECD, the World Bank and institutional investor activists.

3.6. Scandals and failures at major U.S. corporations

A series of scandals and corporate failures surfaced in the United States, a market where the other factors we highlighted played a less important role.[21] Many of these cases concern accounting irregularities that enabled firms to vastly overstate their earnings. Such scandals often emerge during economic downturns: as John Kenneth Galbraith once remarked, recessions catch what the auditors miss.

[21] Prominent failures include undetected off-balance sheet loans to a controlling family (Adelphia) combined with alleged self-dealing by CEOs and other company employees (Computer Associates, Dynegy, Enron, Global Crossing, Qwest, Tyco), deliberate misleading of investors (Kmart, Lucent Technologies, WorldCom), insider trading (ImClone Systems) and/or fraud (Rite Aid) ("Accounting Scandals Spread Across Wall Street", *Financial Times*, 26 June 2002).

4. Conceptual framework

4.1. Agency and contracting

At a general level corporate governance can be described as a problem involving an agent—the CEO of the corporation—and multiple principals—the shareholders, creditors, suppliers, clients, employees, and other parties with whom the CEO engages in business on behalf of the corporation. Boards and external auditors act as intermediaries or representatives of these different constituencies. This view dates back to at least Jensen and Meckling (1976), who describe a firm in abstract terms as "a nexus of contracting relationships". Using more modern language the corporate governance problem can also be described as a "common agency problem", that is an agency problem involving one agent (the CEO) and multiple principals (shareholders, creditors, employees, clients [see Bernheim and Whinston (1985, 1986a, 1986b)].[22]

Corporate governance rules can be seen as the outcome of the contracting process between the various principals or constituencies and the CEO. Thus, the central issue in corporate governance is to understand what the outcome of this contracting process is likely to be, and how corporate governance deviates in practice from the efficient contracting benchmark.

4.2. Ex-ante and ex-post efficiency

Economists determine efficiency by two closely related criteria. The first is ex-ante efficiency: a corporate charter is ex-ante efficient if it generates the highest possible joint payoff for all the parties involved, shareholders, creditors, employees, clients, tax authorities, and other third parties that may be affected by the corporation's actions. The second criterion is Pareto efficiency: a corporate charter is Pareto efficient if no other charter exists that all parties prefer. The two criteria are closely related when the parties can undertake compensating transfers among themselves: a Pareto efficient charter is also a surplus maximizing charter when the parties can make unrestricted side transfers. As closely related as these two notions are it is still important to distinguish between them, since in practice side transfers are often constrained by wealth or borrowing constraints.

[22] A slightly different, sometimes broader perspective, is to describe corporate governance as a multi-principal-multi-agent problem, where both managers and employees are seen as agents for multiple classes of investors. The labelling of employees as "agent" or "principal" is not just a matter of definition. If they are defined as "principal" they are implicitly seen as participants in corporate governance. When and how employees should participate in corporate governance is a delicate and politically sensitive question. We discuss this issue at length in Section 5.6 below. For now, we shall simply take the view that employees are partly "principal" when they have made firm specific investments, which require protection.

4.3. Shareholder value

An efficiency criterion that is often advocated in finance and legal writings on corporate governance is "shareholder value", or the stock market valuation of the corporation. An important basic question is how this notion is related to Pareto efficiency or surplus maximization. Is maximisation of shareholder value synonymous with either or both notions of efficiency?

One influential view on this question (articulated by Jensen and Meckling, 1976) is the following. If (a) the firm is viewed as a nexus of complete contracts with creditors, employees, clients, suppliers, third and other relevant parties, (b) only contracts with shareholders are open-ended; that is, only shareholders have a claim on residual returns after all other contractual obligations have been met, and (c) there are no agency problems, then maximisation of (residual) shareholder value is tantamount to economic efficiency. Under this scenario, corporate governance rules should be designed to protect and promote the interests of shareholders exclusively. [23]

As Jensen and Meckling point out, however, managerial agency problems produce inefficiencies when CEOs act only in the interest of shareholders. There may be excess risk-taking when the firm is highly levered, or, as Myers (1977) has shown, debt overhang may induce underinvestment. Either form of investment inefficiency can be mitigated if managers do not exclusively pursue shareholder value maximisation.

4.4. Incomplete contracts and multiple constituencies

Contracts engaging the corporation with parties other than shareholders are generally incomplete, so that there is no guarantee that corporate governance rules designed to maximise shareholder value are efficient. To guarantee efficiency it is then necessary to take into account explicitly the interests of other constituencies besides shareholders. Whether to take into account other constituencies, and how, is a central issue in corporate governance. Some commentators have argued that shareholder value maximisation is the relevant objective even if contracts with other constituencies are incomplete. Others maintain that board representation should extend beyond shareholders and include other constituencies. There are major differences across countries on this issue, with at one extreme U.K. and U.S. rules designed mainly to promote shareholder value, and at the other German rules designed to balance the interests of shareholders and employees.

One line of argument in favour of shareholder value maximisation in a world of incomplete contracts, first articulated by Oliver Williamson (1984, 1985b), is that shareholders are relatively less well protected than other constituencies. He argues that most workers are not locked into a firm specific relation and can quit at reasonably low cost.

[23] Jensen and Meckling's argument updates an older observation formally articulated by Arrow and Debreu (see Debreu, 1959), that in a competitive economy with complete markets the objective of the firm—unanimously espoused by all claimholders—is profit (or value) maximization.

Similarly, creditors can get greater protection by taking collateral or by shortening the maturity of the debt. Shareholders, on the other hand, have an open-ended contract without specific protection. They need protection the most. Therefore, corporate governance rules should primarily be designed to protect shareholders' interests.

In addition, Hansmann (1996) has argued that one advantage of involving only one constituency in corporate governance is that both corporate decision-making costs and managerial discretion will be reduced. Although Hansmann argues in favour of a governance system by a single constituency he allows for the possibility that other constituencies besides shareholders may control the firm. In some situations a labour-managed firm, a customer co-operative, or possibly a supplier co-operative may be a more efficient corporate governance arrangement. In his view, determining which constituency should govern the firm comes down to identifying which has the lowest decision making costs and which has the greatest need of protection.

An obvious question raised by Williamson's argument is that if it is possible to get better protection by signing debt contracts, why not encourage all investors in the firm to take out debt contracts. Why worry about protecting shareholders when investors can find better protection by writing a debt contract? Jensen (1986, 1989) has been a leading advocate of this position, arguing that the best way to resolve the agency problem between the CEO and investors is to have the firm take on as much debt as possible. This would limit managerial discretion by minimising the "free cash-flow" available to managers and, thus, would provide the best possible protection to investors.

The main difficulty with Jensen's logic is that highly levered firms may incur substantial costs of financial distress. They may face direct bankruptcy costs or indirect costs in the form of debt-overhang (see Myers, 1977 or Hart and Moore, 1995 and Hennessy and Levy, 2002). To reduce the risk of financial distress it may be desirable to have the firm rely partly on equity financing. And to reduce the cost of equity capital it is clearly desirable to provide protections to shareholders through suitably designed corporate governance rules.

Arguably it is in the interest of corporations and their CEOs to design efficient corporate governance rules, since this would minimise their cost of capital, labour and other inputs. It would also maximise the value of their products or services to their clients. Firms may want to acquire a reputation for treating shareholders or creditors well, as Kreps (1990) and Diamond (1989) have suggested.[24] If reputation building is effective then mandatory regulatory intervention seems unnecessary.

[24] Interestingly, although reputation building is an obvious way to establish investor protection, this type of strategy has been somewhat under-emphasised in the corporate governance literature. In particular, there appears to be no systematic empirical study on reputation building, even if there are many examples of large corporations that attempt to build a reputation by committing to regular dividend payments, disclosing information, and communicating with analysts [see however Carleton, Nelson, and Weisbach (1998) for evidence on voluntary communications between large U.S. corporations and institutional investors]. For a recent survey of the disclosure literature, including voluntary disclosure by management, see Healy and Palepu (2001).

4.5. Why do we need regulation?

A natural question to ask then is why regulations imposing particular governance rules (required by stock exchanges, legislatures, courts or supervisory authorities) are necessary.[25] If it is in the interest of firms to provide adequate protection to shareholders, why mandate rules, which may be counterproductive? Even with the best intentions regulators may not have all the information available to design efficient rules.[26] Worse still, regulators can be captured by a given constituency and impose rules favouring one group over another.

There are at least two reasons for regulatory intervention. The main argument in support of mandatory rules is that even if the founder of the firm or the shareholders can design and implement any corporate charter they like, they will tend to write inefficient rules since they cannot feasibly involve all the parties concerned in a comprehensive bargain. By pursuing their interests over those of parties missing from the bargaining table they are likely to write inefficient rules. For example, the founder of the firm or shareholders will want to put in place anti-takeover defences in an attempt to improve the terms of takeovers and they will thereby tend to limit hostile takeover activity excessively.[27] Alternatively, shareholders may favour takeovers that increase the value of their shares even if they involve greater losses for unprotected creditors or employees.[28]

Another argument in support of mandatory rules is that, even if firms initially have the right incentives to design efficient rules, they may want to break or alter them later. A problem then arises when firms do not have the power to commit not to change (or break) the rules down the road. When shareholders are dispersed and do not take an active interest in the firm it is possible, indeed straightforward, for management to change the rules to their advantage *ex post*. Dispersed shareholders, with small interests in the corporation, are unlikely to incur the large monitoring costs that are sometimes required to keep management at bay. They are more likely to make management their proxy, or to abstain.[29] Similarly, firms may not be able to build credible reputations for treating shareholders well if dispersed shareholders do not take an active interest in the firm and if important decisions such as mergers or replacements of CEOs are infrequent. Shareholder protection may then require some form of concentrated ownership or a regulatory intervention to overcome the collective action problem among dispersed shareholders.

[25] Compliance with corporate governance "codes" is mostly voluntary.

[26] On the other hand, if the identification and formulation of efficient corporate governance rules is a costly process it makes sense to rely on courts and corporate law to formulate default rules, which corporations could adopt or opt out of [see Ayres and Gertner (1989)].

[27] We shall return to this observation, articulated in Grossman and Hart (1980) and Scharfstein (1988), at greater length in Section 5.

[28] Shleifer and Summers (1988) discuss several hostile takeover cases where the value for target and bidding shareholders came apparently at the expense of employees and creditors.

[29] Alternatively, limiting managerial discretion ex ante and making it harder to change the rules by introducing supermajority requirements into the corporate charter would introduce similar types of inefficiency as with debt.

4.6. Dispersed ownership

Since dispersed ownership is such an important source of corporate governance problems it is important to inquire what causes dispersion in the first place. There are at least three reasons why share ownership may be dispersed in reality. First, and perhaps most importantly, individual investors' wealth may be small relative to the size of some investments. Second, even if a shareholder can take a large stake in a firm, he may want to diversify risk by investing less. A related third reason is investors' concern for liquidity: a large stake may be harder to sell in the secondary market.[30] For these reasons it is not realistic or desirable to expect to resolve the collective action problem among dispersed shareholders by simply getting rid of dispersion.

4.7. Summary and conclusion

In sum, mandatory governance rules (as required by stock exchanges, legislatures, courts or supervisory authorities) are necessary for two main reasons: first, to overcome the collective action problem resulting from the dispersion among shareholders, and second, to ensure that the interests of all relevant constituencies are represented. Indeed, other constituencies besides shareholders face the same basic collective action problem. Corporate bondholders are also dispersed and their collective action problems are only imperfectly resolved through trust agreements or consortia or in bankruptcy courts. In large corporations employees and clients may face similar collective action problems, which again are imperfectly resolved by unions or consumer protection organisations.

Most of the finance and corporate law literature on corporate governance focuses only on collective action problems of shareholders. Accordingly, we will emphasize those problems in this survey. As the literature on representation of other constituencies is much less developed we shall only touch on this issue in Sections 5 to 7.

We distinguish five main ways to mitigate shareholders' collective action problems:

(1) Election of a board of directors representing shareholders' interests, to which the CEO is accountable.
(2) When the need arises, a takeover or proxy fight launched by a corporate raider who temporarily concentrates voting power (and/or ownership) in his hands to resolve a crisis, reach an important decision or remove an inefficient manager.
(3) Active and continuous monitoring by a large blockholder, who could be a wealthy investor or a financial intermediary, such as a bank, a holding company or a pension fund.
(4) Alignment of managerial interests with investors through executive compensation contracts.

[30] A fourth reason for the observed dispersion in shareholdings may be securities regulation designed to protect minority shareholders, which raises the cost of holding large blocks. This regulatory bias in U.S. corporate law has been highlighted by Black (1990), Roe (1990, 1991, 1994) and Bhide (1993).

(5) Clearly defined fiduciary duties for CEOs and the threat of class-action suits that either block corporate decisions that go against investors' interests, or seek compensation for past actions that have harmed their interests.

As we shall explain, a potential difficulty with the first three approaches is the old problem of who monitors the monitor and the risk of collusion between management (the agent) and the delegated monitor (director, raider, blockholder). If dispersed shareholders have no incentive to supervise management and take an active interest in the management of the corporation why should directors—who generally have equally small stakes—have much better incentives to oversee management? The same point applies to pension fund managers. Even if they are required to vote, why should they spend the resources to make informed decisions when the main beneficiaries of those decisions are their own principals, the dispersed investors in the pension fund? Finally, it might appear that corporate raiders, who concentrate ownership directly in their hands, are not susceptible to this delegated monitoring problem. This is only partially true since the raiders themselves have to raise funds to finance the takeover. Typically, firms that are taken over through a hostile bid end up being substantially more highly levered. They may have resolved the shareholder collective action problem, but at the cost of significantly increasing the expected cost of financial distress.

Enforcement of fiduciary duties through the courts has its own shortcomings. First, management can shield itself against shareholder suits by taking out appropriate insurance contracts at the expense of shareholders.[31] Second, the "business judgement" rule (and similar provisions in other countries) severely limits shareholders' ability to prevail in court.[32] Finally, plaintiffs' attorneys do not always have the right incentives to monitor management. Managers and investment bankers often complain that contingency fee awards (which are typically a percentage of damages awarded in the event that the plaintiff prevails) can encourage them to engage in frivolous suits, a problem that is likely to be exacerbated by the widespread use of director and officer (D&O) liability insurance. This is most likely to be the case in the U.S. In other countries fee awards (which mainly reflect costs incurred) tend to increase the risk of lawsuits for small shareholders and the absence of D&O insurance makes it harder to recover damages.[33]

[31] Most large U.S. corporations have taken out director and officer liability (D&O) insurance policies (see Danielson and Karpoff, 1998). See Gutiérrez (2003) for an analysis of fiduciary duties, liability and D&O insurance.

[32] The "director's business judgement cannot be attacked unless their judgement was arrived at in a negligent manner, or was tainted by fraud, conflict of interest, or illegality" (Clark, 1986, p. 124). The business judgement rule give little protection to directors for breaches of form (e.g. for directors who fail to attend meetings or read documents) but can extend to conflict of interest situations, provided that a self-interested decision is approved by disinterested directors (Clark, 1986, pp. 123, 138).

[33] See Fischel and Bradley (1986), Romano (1991) and Kraakman, Park, and Shavell (1994) for an analysis of distortions of litigation incentives in shareholder suits.

5. Models

5.1. Takeover models

One of the most radical and spectacular mechanisms for disciplining and replacing managers is a hostile takeover. This mechanism is highly disruptive and costly. Even in the U.S. and the U.K. it is relatively rarely used. In most other countries it is almost non-existent. Yet, hostile takeovers have received a great deal of attention from academic researchers. In a hostile takeover the raider makes an offer to buy all or a fraction of outstanding shares at a stated tender price. The takeover is successful if the raider gains more than 50% of the voting shares and thereby obtains effective control of the company. With more than 50% of the voting shares, in due course he will be able to gain majority representation on the board and thus be able to appoint the CEO.

Much research has been devoted to the mechanics of the takeover process, the analysis of potentially complex strategies for the raider and individual shareholders, and to the question of ex-post efficiency of the outcome. Much less research has been concerned with the ex-ante efficiency of hostile takeovers: the extent to which takeovers are an effective disciplining device on managers.

On this latter issue, the formal analysis by Scharfstein (1988) stands out. Building on the insights of Grossman and Hart (1980), he considers the ex-ante financial contracting problem between a financier and a manager. This contract specifies a state contingent compensation scheme for the manager to induce optimal effort provision. In addition the contract allows for ex-post takeovers, which can be efficiency enhancing if either the raider has information about the state of nature not available to the financier or if the raider is a better manager. In other words, takeovers are useful both because they reduce the informational monopoly of the incumbent manager about the state of the firm and because they allow for the replacement of inefficient managers. The important observation made by Scharfstein is that even if the firm can commit to an ex-ante optimal contract, this contract is generally inefficient. The reason is that the financier and manager partly design the contract to try and extract the efficiency rents of future raiders. Like a non-discriminating monopolist, they will design the contract so as to "price" the acquisition above the efficient competitive price. As a result, the contract will induce too few hostile takeovers on average.

Scharfstein's observation provides an important justification for regulatory intervention limiting anti-takeover defences, such as super-majority amendments[34], staggered

[34] These amendments raise the majority rule above 50% in the event of an hostile takeover.

boards[35], fair price amendments (ruling out two-tier tender offers)[36], and poison pills[37] (see Section 7.1.4 for a more detailed discussion). These defences are seen by many to be against shareholders' interests and to be put in place by managers of companies with weak corporate governance structures (see, for example, Gilson, 1981 and Easterbrook and Fischel, 1981). Others, however, see them as an important weapon enabling the target firm to extract better terms from a raider (see Baron, 1983; Macey and McChesney, 1985; Shleifer and Vishny, 1986; Hirshleifer and Titman, 1990; Hirshleifer and Thakor, 1994; and Hirshleifer, 1995). Even if one takes the latter perspective, however, Scharfstein's argument suggests that some of these defences should be regulated or banned.

A much larger literature exists on the issue of ex-post efficiency of hostile takeovers. The first formal model of a tender offer game is due to Grossman and Hart (1980). They consider the following basic game. A raider can raise the value per share from $v = 0$ under current management to $v = 1$. He needs 50% of the voting shares and makes a conditional tender offer of p per share.[38] Share ownership is completely dispersed; indeed to simplify the analysis they consider an idealised situation with an infinite number of shareholders. It is not difficult to see that a dominant strategy for each shareholder is to tender if $p \geq 1$ and to hold on to their shares if $p < 1$. Therefore the lowest price at which the raider is able to take over the firm is $p = 1$, the post-takeover value per share. In other words, the raider has to give up all the value he can generate to existing shareholders. If he incurs costs in making the offer or in undertaking the management changes that produce the higher value per share he may well be discouraged from attempting a takeover. In other words, there may be too few takeover attempts ex-post.

Grossman and Hart (1980) suggest several ways of improving the efficiency of the hostile takeover mechanism. All involve some dilution of minority shareholder rights. Consistent with their proposals for example is the idea that raiders be allowed to "squeeze (freeze) out" minority shareholders that have not tendered their shares[39], or

[35] Staggered boards are a common defence designed to postpone the time at which the raider can gain full control of the board after a takeover. With only a fraction y of the board renewable every x years, the raider would have to wait up to $x/2y$ years before gaining over 50% of the seats.

[36] Two-tier offers specify a higher price for the first n shares tendered than for the remaining ones. They tend to induce shareholders to tender and, hence, facilitate the takeover. Such offers are generally illegal in the U.S., but when they are not companies can ban them by writing an amendment into the corporate charter.

[37] Most poison pills give the right to management to issue more voting shares at a low price to existing shareholders in the event that one shareholder owns more than a fraction x of outstanding shares. Such clauses, when enforced, make it virtually impossible for a takeover to succeed. When such a defence is in place the raider has to oust the incumbent board in a proxy fight and remove the pill. When the pill is combined with defences that limit the raider's ability to fight a proxy fight—for example a staggered board—the raider effectively has to bribe the incumbent board.

[38] A conditional offer is one that binds only if the raider gains control by having more than a specified percentage of the shares tendered.

[39] A squeeze or freeze out forces minority shareholders to sell their shares to the raider at (or below) the tender offer price. When the raider has this right it is no longer a dominant strategy to hold on to one's shares when $p < 1$.

to allow raiders to build up a larger "toehold" before they are required to disclose their stake.[40]

Following the publication of the Grossman and Hart article a large literature has developed analysing different variants of the takeover game, with non-atomistic share ownership (e.g. Kovenock, 1984; Bagnoli and Lipman, 1988; and Holmstrom and Nalebuff, 1992), with multiple bidders (e.g. Fishman, 1988; Burkart, 1995; and Bulow, Huang, and Klemperer, 1999), with multiple rounds of bidding (Dewatripont, 1993), with arbitrageurs (e.g. Cornelli and Li, 2002), asymmetric information (e.g. Hirshleifer and Titman, 1990; and Yilmaz, 2000), etc. Much of this literature has found Grossman and Hart's result that most of the gains of a takeover go to target shareholders (because of "free riding" by small shareholders) to be non-robust when there is only one bidder. With either non-atomistic shareholders or asymmetric information their extreme "free-riding" result breaks down. In contrast, empirical studies have found again and again that on average all the gains from hostile takeovers go to target shareholders [see Jensen and Ruback (1983) for a survey of the early literature]. While this is consistent with Grossman and Hart's result, other explanations have been suggested, such as (potential) competition by multiple bidders, or raiders' hubris leading to over-eagerness to close the deal (Roll, 1986).

More generally, the theoretical literature following Grossman and Hart (1980) is concerned more with explaining bidding patterns and equilibrium bids given existing regulations than with determining which regulatory rules are efficient. A survey of most of this literature can be found in Hirshleifer (1995). For an extensive discussion of empirical research on takeovers see also the survey by Burkart (1999).

Formal analyses of optimal takeover regulation have focused on four issues: (1) whether deviations from a "one-share-one vote" rule result in inefficient takeover outcomes; (2) whether raiders should be required to buy out minority shareholders; (3) whether takeovers may result in the partial expropriation of other inadequately protected claims on the corporation, and if so, whether some anti-takeover amendments may be justified as basic protections against expropriation; and (4) whether proxy contests should be favored over tender offers.

From 1926 to 1986 one of the requirements for a new listing on the New York Stock Exchange was that companies issue a single class of voting stock (Seligman, 1986).[41] That is, companies could only issue shares with the same number (effectively one) of votes each. Does this regulation induce efficient corporate control contests? The analysis of Grossman and Hart (1988) and Harris and Raviv (1988a, 1988b) suggests that the

[40] A toehold is the stake owned by the raider before he makes a tender offer. In the U.S. a shareholder owning more than 5% of outstanding shares must disclose his stake to the SEC. The raider can always make a profit on his toehold by taking over the firm. Thus the larger his toehold the more likely he is to make a takeover attempt (see Shleifer and Vishny, 1986 and Kyle and Vila, 1991).

[41] A well-known exception to this listing rule was the Ford Motor Company, listed with a dual class stock capitalisation in 1956, allowing the Ford family to exert 40% of the voting rights with 5.1% of the capital (Seligman, 1986).

answer is a qualified "yes". They point out that under a "one-share-one-vote" rule inefficient raiders must pay the highest possible price to acquire control. In other words, they face the greatest deterrent to taking over a firm under this rule. In addition, they point out that a simple majority rule is most likely to achieve efficiency by treating incumbent management and the raider symmetrically.

Deviations from "one-share-one-vote" may, however, allow initial shareholders to extract a greater share of the efficiency gain of the raider in a value-increasing takeover. Indeed, Harris and Raviv (1988a), Zingales (1995) and Gromb (1993) show that maximum extraction of the raider's efficiency rent can be obtained by issuing two extreme classes of shares, votes-only shares and non-voting shares. Under such a share ownership structure the raider only purchases votes-only shares. He can easily gain control, but all the benefits he brings go to the non-voting shareholders. Under their share allocation scheme all non-voting shareholders have no choice but to "free-ride" and thus appropriate most of the gains from the takeover.

Another potential benefit of deviations from "one-share-one-vote" is that they may induce more listings by firms whose owners value retaining control of the company. Family-owned firms are often reluctant to go public if they risk losing control in the process. These firms might go public if they could retain control through a dual-class share structure. As Hart (1988) argues, deviations from one-share-one-vote would benefit both the firm and the exchange in this case. They are also unlikely to hurt minority shareholders, as they presumably price in the lack of control rights attached to their shares at the IPO stage.

Burkart, Gromb, and Panunzi (1998) extend this analysis by introducing a post-takeover agency problem. Such a problem arises when the raider does not own 100% of the shares ex post, and is potentially worse, the lower the raider's post-takeover stake. They show that in such a model initial shareholders extract the raider's whole efficiency rent under a "one-share-one-vote" rule. As a result, some costly takeovers may be deterred. To reduce this inefficiency they argue that some deviations from "one-share-one-vote" may be desirable.

The analysis of mandatory bid rules is similar to that of deviations from "one-share-one-vote". By forcing a raider to acquire all outstanding shares, such a rule maximises the price an inefficient raider must pay to acquire control. On the other hand, such a rule may also discourage some value increasing takeovers (see Bergstrom, Hogfeldt, and Molin, 1997).

In an influential article Shleifer and Summers (1988) have argued that some takeovers may be undesirable if they result in a "breach of trust" between management and employees. If employees (or clients, creditors and suppliers) anticipate that informal relations with current management may be broken by a new managerial team that has taken over the firm they may be reluctant to invest in such relations and to acquire firm specific human capital. They argue that some anti-takeover protections may be justified at least for firms where specific (human and physical) capital is important. A small formal literature has developed around this theme (see e.g. Knoeber, 1986; Schnitzer, 1995; and Chemla, 2005). One lesson emerging from this research is that efficiency de-

pends critically on which type of anti-takeover protection is put in place. For example, Schnitzer (1995) shows that only a specific combination of a poison pill with a golden parachute would provide adequate protection for the manager's (or employees') specific investments. The main difficulty from a regulatory perspective, however, is that protection of specific human capital is just too easy an excuse to justify managerial entrenchment. Little or no work to date has been devoted to the question of identifying which actions or investments constitute "entrenchment behaviour" and which do not. It is therefore impossible to say conclusively whether current regulations permitting anti-takeover amendments, which both facilitate managerial entrenchment and provide protections supporting informal agreements, are beneficial overall.

Another justification for poison pills that has recently been proposed by Bebchuk and Hart (2001) is that poison pills make it impossible to remove an incumbent manager through a hostile takeover unless the tender offer is accompanied by a proxy fight over the redemption of the poison pill.[42] In other words, Bebchuk and Hart argue that the presence of a poison pill requires a mechanism for removing incumbent managers that combines both a tender offer and a proxy contest. In their model such a mechanism dominates both straight proxy contests and straight tender offers. The reason why straight proxy contests are dominated is that shareholders tend to be (rationally) sceptical of challengers. Challengers may be worse than incumbents and only seek control to gain access to large private benefits of control. A tender offer accompanying a proxy fight mollifies shareholder scepticism by demonstrating that the challenger is ready to "put his money where his mouth is". In general terms, the reason why straight tender offers are dominated is that a tender offer puts the decision in the hands of the marginal shareholder while majority voting effectively puts the control decision in the hands of the average shareholder (or median voter). The average shareholder always votes in

[42] Bebchuk and Hart's conclusions rest critically on their view for why straight proxy fights are likely to be ineffective in practice in removing incumbent management. Alternative reasons have been given for why proxy fights have so often failed, which would lead to different conclusions. For example, it has often been argued that management has an unfair advantage in campaigning for shareholder votes as they have access to shareholder lists as well as the company coffers (for example, Hewlett-Packard spent over $100 mn to convince shareholders to approve its merger with Compaq). In addition they can pressure institutional investors to vote for them (in the case of Hewlett-Packard, it was alleged that the prospect of future corporate finance business was implicitly used to entice Deutsche Bank to vote For the merger). If it is the case that institutional and other affiliated shareholders are likely to vote for the incumbent for these reasons then it is imperative to ban poison pills to make way for a possible hostile takeover as Shleifer and Vishny (1986), Harris and Raviv (1988a), Gilson (2001, 2002) and Gilson and Schwartz (2001) have argued among others. Lipton and Rowe (2002) take yet another perspective. They question the premise in most formal analyses of takeovers that financial markets are efficient. They point to the recent bubble and crash on NASDAQ and other financial markets as evidence that stock valuations are as likely to reflect fundamental value as not. They argue that when stock valuations deviate in this way from fundamental value they can no longer be taken as a reliable guide for the efficient allocation of control or for that matter as a reliable mechanism to discipline management. In such inefficient financial markets poison pills are necessary to protect management from the vagaries of the market and from opportunistic bids. They maintain that this is the doctrine underlying Delaware law on takeover defenses.

favour of a value increasing control change, while the marginal shareholder in a tender offer only decides to tender if she is better off tendering than holding on to her shares assuming that the takeover will succeed. Such behaviour can result in excessive free-riding and inefficient control allocations.

5.2. Blockholder models

An alternative approach to mitigating the collective action problem of shareholders is to have a semi-concentrated ownership structure with at least one large shareholder, who has an interest in monitoring management and the power to implement management changes. Although this solution is less common in the U.S. and U.K.—because of regulatory restrictions on blockholder actions—some form of concentration of ownership or control is the dominant form of corporate governance arrangement in continental Europe and other OECD countries.

The first formal analyses of corporate governance with large shareholders point to the benefits of large shareholders in facilitating takeovers (see Grossman and Hart, 1980, and Shleifer and Vishny, 1986). A related theme is the classic tradeoff underlying the standard agency problem with moral hazard: the tradeoff between optimal risk diversification, which is obtained under a fully dispersed ownership structure, and optimal monitoring incentives, which require concentrated ownership. Thus, Leland and Pyle (1977) have shown that it may be in the interest of a risk-averse entrepreneur going public to retain a large stake in the firm as a signal of quality, or as a commitment to manage the firm well. Later, Admati, Pfleiderer, and Zechner (1994) and Huddart (1993) have considered the monitoring incentives of a large risk-averse shareholder. They show that in equilibrium the large shareholder has too small a stake and under-invests in monitoring, because the large shareholder prefers to diversify his holdings somewhat even if this reduces his incentives to monitor. They also point out that ownership structures with one large block may be unstable if the blockholder can gradually erode his stake by selling small quantities of shares in the secondary market. The main regulating implication of these analyses is that corporate governance might be improved if blockholders could be subsidised to hold larger blocks. Indeed, the main problem in these models is to give greater incentives to monitor to the blockholder.[43]

A related set of models further pursues the issue of monitoring incentives of firms with liquid secondary markets. An influential view generally attributed to Hirschman (1970) is that when monitors can easily 'exit' the firm they tend not to exercise their 'voice'. In other words, blockholders cannot be relied upon to monitor management actively if they have the option to sell their stake instead.[44] Indeed, some commentators (most notably Mayer, 1988; Black, 1990; Coffee, 1991; Roe, 1994; and Bhide, 1993)

[43] Demsetz (1986) points out that insider trading makes it easier for a shareholder to build a toehold and thus facilitates monitoring.

[44] The idea that blockholders would rather sell their stake in mismanaged firms than try to fix the management problem is known as the "Wall Street rule" (see Black, 1990).

have argued that it is precisely the highly liquid nature of U.S. secondary markets that makes it difficult to provide incentives to large shareholders to monitor management.

This issue has been analysed by Kahn and Winton (1998) and Maug (1998) among others. Kahn and Winton show how market liquidity can undermine large shareholders' incentives to monitor by giving them incentives to trade on private information rather than intervene. They argue, however, that incentives to speculate may be small for blue-chip companies, where the large shareholder is unlikely to have a significant informational advantage over other market participants. Similarly, Maug points out that in liquid markets it is also easier to build a block. This gives large shareholders an added incentive to invest in information gathering.

To summarise, this literature emphasizes the idea that if the limited size of a block is mainly due to the large shareholder's desire to diversify risk then under-monitoring by the large shareholder is generally to be expected.

An entirely different perspective is that the large investor may want to limit his stake to ensure minimum secondary market liquidity. This is the perspective taken by Holmstrom and Tirole (1993). They argue that share prices in the secondary market provide valuable information about the firm's performance. To obtain accurate valuations, however, the secondary market must be sufficiently liquid. Indeed, liquidity raises speculators' return to acquiring information and thus improves the informativeness of the secondary market price. The more informative stock price can then be included in compensation packages to provide better incentives to managers. According to this view it is the market that does the monitoring and the large shareholder may only be necessary to act on the information produced by the market.[45]

In other words, there may be a natural complementarity between speculation in secondary markets and monitoring by large shareholders. This idea is pursued further in Faure-Grimaud and Gromb (2004) and Aghion, Bolton, and Tirole (2000). These models show how large shareholders' monitoring costs can be reduced through better pricing of shares in the secondary market. The basic idea is that more accurate pricing provides not only greater liquidity to the large shareholder, but also enhances his incentives to monitor by reflecting the added value of his monitoring activities in the stock price. The latter paper also determines the optimal degree of liquidity of the large shareholder's stake to maximize his incentives to monitor. This theory finds its most natural application for corporate governance in start-ups financed with venture capital. It is well known that venture capitalists not only invest large stakes in individual start-ups but also participate in running the firm before it goes public. Typical venture capital contracts can

[45] Strictly speaking, in their model the large shareholder is only there by default, because in selling to the secondary market he has to accept a discount reflecting the information-related trading costs that investors anticipate incurring. Thus, the large shareholder can achieve the desired amount of information acquisition in the market by adjusting the size of his stake.

be seen as incentive contracts aimed in part at regulating the venture capitalist's exit options so as to provide the best incentives for monitoring.[46,47]

Just as with takeovers, there are obvious benefits from large shareholder monitoring but there may also be costs. We pointed out earlier that hostile takeovers might be undesirable if their main purpose is to expropriate employees or minority shareholders. Similarly, large shareholder monitoring can be too much of a good thing. If the large shareholder uses his power to hold up employees or managers, the latter may be discouraged from making costly firm specific investments. This point has been emphasized in a number of theoretical studies, most notably in Aghion and Tirole (1997), Burkart, Gromb, and Panunzi (1997), and Pagano and Röell (1998). Thus, another reason for limiting a large shareholder's stake may be to prevent over-monitoring and ex-post opportunism. As privately held firms tend to have concentrated ownership structures they are more prone to over-monitoring. Pagano and Röell argue that one important motive for going public is that the manager may want to free himself from an overbearing owner or venture capitalist.[48]

There is only a short step from over-monitoring to downright expropriation, self-dealing or collusion with management at the expense of minority shareholders. Indeed, an important concern of many commentators is the conflict of interest among shareholders inherent in blockholder ownership structures. This conflict is exacerbated when in addition there is separation between voting rights and cash-flow rights, as is common in continental Europe. Many commentators have argued that such an arrangement is particularly vulnerable to self-dealing by the controlling shareholder (see e.g. Zingales, 1994; Bianco, Casavola, and Ferrando, 1997; Burkart, Gromb, and Panunzi, 1997; La Porta,

[46] See Bartlett (1994), Gompers and Lerner (1999), Levin (1995) and Kaplan and Strömberg (2003) for discussions of contractual provisions governing the venture capitalist's "exit". See also Berglöf (1994) and Hellman (1997) for models of corporate governance of venture capital financed firms.

[47] Another form of complementarity is considered in a recent paper by Chidambaran and John (2000). They argue that large shareholder monitoring can be facilitated by managerial cooperation. However, to achieve such cooperation managers must be given an equity stake in the firm. With sufficient equity participation, the authors show that managers have an incentive to disclose information that brings market valuations closer to fundamental values of the business. They argue that this explains why greater institutional holdings are associated with larger stock option awards but lower compensation levels for CEOs (see Hartzell and Starks, 2003).

[48] Most of the theoretical literature on large shareholders only considers ownership structures where all but one shareholder are small. Zwiebel (1995) is a recent exception. He considers ownership structures where there may be more than one large shareholder and also allows for alliances among small block-holders. In such a setting he shows that one of the roles of a large block-holding is to fend off alliances of smaller block-holders that might compete for control (see also Gomes and Novaes, 2000 and Bloch and Hege, 2000 for two other recent formal analyses of ownership structures with multiple large shareholders). An entirely different perspective on the role of large outside shareholders is given in Muller and Warneryd (2001) who argue that outside owners can reduce inefficient rent seeking of insiders and managers by inducing them to join forces to fight the outsider's own rent seeking activities. This story fits well the situation of many second generation family-owned firms, who decide to open up their ownership to outsiders in an attempt to stop feuding among family members.

Lopez-de-Silanes, and Shleifer, 1998; Wolfenzon, 1999; Bebchuk, 1999; and Bebchuk, Kraakman, and Trianis, 2000).[49] Most of these commentators go as far as arguing that existing blockholder structures in continental Europe are in fact likely to be inefficient and that U.S.-style regulations restricting blockholder rights should be phased in.

The analyses of Aghion and Tirole (1997), Burkart, Gromb, and Panunzi (1997), and Pagano and Röell (1998), however, suggest that if there is a risk of over-monitoring or self-dealing it is often possible to design the corporate ownership structure or charter to limit the power of the blockholder. But Bebchuk (1999) and Bebchuk and Roe (1999) retort that although it is theoretically possible to design corporate charters that restrain self-dealing, in practice the Coase theorem is likely to break down and therefore regulations limiting blockholder rights are called for. Bebchuk (1999) develops a model where dispersed ownership is unstable when large shareholders can obtain rents through self-dealing since there is always an incentive to grab and protect control rents. If a large shareholder does not grab the control rents then management will. Bebchuk's extreme conclusion, however, is based on the assumption that a self-dealing manager cannot be disciplined by a takeover threat.[50] His general conclusion—that if self-dealing is possible under a lax corporate law it will inevitably lead to concentrated ownership—is a particular version of the general argument outlined in the introduction that under dispersed ownership management may not be able to commit to an ex-ante efficient corporate governance rule. Bebchuk and Roe (1999) make a complementary point, arguing that inefficiencies can persist if there is a collective action problem in introducing better corporate governance arrangements.

So far we have discussed the costs and benefits of takeovers and large shareholder monitoring respectively. But what are the relative advantages of each approach? One comparative analysis of this question is proposed by Bolton and Von Thadden (1998a, 1998b). They argue that one potential benefit of blockholder structures is that monitoring will take place on an ongoing basis. In contrast, a system with dispersed shareholders can provide monitoring and intervention only in crisis situations (if at all), through a hostile takeover. The benefit of dispersed ownership, on the other hand is enhanced liquidity in secondary markets. They show that depending on the value of monitoring, the need for intervention and the demand for liquidity either system can dominate the other. The comparison between the two systems obviously also depends on the regulatory structure in place. If, as Black (1990) has forcefully argued, regulations substantially

[49] Most commentators point to self-dealing and "private benefits" of control of the large shareholder. Perhaps, equally worrying, however is collusion between management and the blockholder. This aspect of the problem has not received much attention. For two noteworthy exceptions see Tirole (1986) and Burkart and Panunzi (2006).

[50] The issue of competition for control rents between a large shareholder and the CEO is analysed in Burkart and Panunzi (2006). They argue that access to control rents has positive incentive effects on the CEO. It also has positive effects on the blockholder's incentive to monitor. However, competition for these rents between the CEO and the blockholder may undermine the incentives of either party.

increase the costs of holding blocks[51] (as is the case in both the U.S. and the U.K.) then a system with dispersed shareholders relying on hostile takeovers might be best. On the other hand, if regulations which mainly increase the costs of hostile takeovers but do not otherwise substantially restrict blockholder rights (as in continental Europe) are in place then a system based on blockholder monitoring may arise.

Another comparative analysis is proposed by John and Kedia (2000). They draw the distinction between "self-binding" mechanisms (like bank or large shareholder monitoring) and "intervention" mechanisms (like hostile takeovers). They let underlying conditions vary according to two parameters: the costs of bank monitoring and the effectiveness of hostile takeovers. Depending on the values of these parameters the optimal governance mechanism is either: (i) concentrated ownership (when bank monitoring is costly and takeovers are not a threat), (ii) bank monitoring (when monitoring costs are low and takeovers are ineffective), or (iii) dispersed ownership and hostile takeovers (when anti-takeover defences are low and monitoring is costly). One implication of their analysis is that corporate governance in Europe and Japan may not converge to U.S. practice simply by introducing the same takeover regulations. If banks are able to maintain a comparative advantage in monitoring these countries may continue to see a predominance of bank monitoring.[52]

5.3. Delegated monitoring and large creditors

One increasingly important issue relating to large shareholder or investor monitoring concerns the role of institutional shareholder activism by pension funds and other financial intermediaries. Pension funds, mutual funds and insurance companies (and banks outside the U.S.) often buy large stakes in corporations and could take an active role in monitoring management. Generally, however, because of regulatory constraints or lack of incentives they tend to be passive (see Black, 1990; Coffee, 1991; Black and Coffee, 1994). One advantage of greater activism by large institutional investors is that fund managers are less likely to engage in self-dealing and can therefore be seen as almost ideal monitors of management. But a major problem with institutional monitoring is that fund managers themselves have no direct financial stake in

[51] Among U.S. rules discouraging shareholder action are disclosure requirements, prohibitions on insider trading and short-swing trading, rules imposing liability on "controlling shareholders", limits on institutional shareholdings in a single company and fiduciary duty rules; a detailed account is given by Black (1990). One of the most striking restrictions is the rule governing shareholder proposals (Rule 14a-8): a shareholder "can offer only one proposal per year, ... must submit the proposal ... 5 months before the next annual meeting .... A proposal cannot relate to ordinary business operations or the election of directors ... and not conflict with a manager proposal" (Black, 1990, p. 541).

[52] Yet another comparative analysis is given in Ayres and Cramton (1994). They emphasise two benefits of large shareholder structures. First, better monitoring and second less myopic market pressure to perform or fend off a hostile takeover (see also Narayanan, 1985; Shleifer and Vishny, 1989; and Stein, 1988, 1989 for a formal analysis of myopic behaviour induced by hostile takeovers). It is debatable, however, whether less market pressure is truly a benefit (see Romano, 1998 for a discussion of this point).

the companies they invest in and therefore have no direct or adequate incentives for monitoring.[53]

The issue of institutional investor incentives to monitor has been analysed mainly in the context of bank monitoring. The first formal analysis of the issue of who monitors the monitor (in the context of bank finance) is due to Diamond (1984). He shows that, as a means of avoiding duplication of monitoring by small investors, delegated monitoring by a banker may be efficient.[54] He resolves the issue of "who monitors the monitor" and the potential duplication of monitoring costs for depositors, by showing that if the bank is sufficiently well diversified then it can almost perfectly guarantee a fixed return to its depositors. As a result of this (almost safe) debt-like contract that the bank offers to its depositors, the latter do not need to monitor the bank's management continuously.[55] They only need to inspect the bank's books when it is in financial distress, an event that is extremely unlikely when the bank is well diversified. As Calomiris and Kahn (1991) and Diamond and Rajan (2001) have emphasized more recently, however, preservation of the banker's incentives to monitor also requires a careful specification of deposit contracts. In particular, banks' incentives are preserved in their model only if there is no deposit insurance and the first-come first-served feature of bank deposit contracts is maintained. In other words, bankers' incentives to monitor are preserved only if banks are disciplined by the threat of a bank run by depositors.[56]

One implication of these latter models is that under a regime of deposit insurance banks will not adequately monitor firms and will engage in reckless lending. The greater incidence of banking crises in the past 20 years is sometimes cited as corroborating evidence for this perspective. Whether the origin of these crises is to be found in deposit insurance and inadequate bank governance is a debated issue. Other commentators argue that the recent banking crises are just as (or more) likely to have resulted from exchange rate crises and/or a speculative bubble. Many commentators put little faith in depositors' abilities (let alone incentives) to monitor banks and see bank regulators as better placed to monitor banks in the interest of depositors (see Dewatripont and Tirole, 1994). Consistent with this perspective is the idea that deposit insurance creates adequate incentives for bank regulators to monitor banks, as it makes them residual claimants on banks' losses. However, these incentives can be outweighed by a lack of commitment to close down insolvent banks and by regulatory forbearance. It is often argued that bank bailouts and the expectation of future bailouts create a "moral hazard"

[53] As Romano (2001) has argued and as the empirical evidence to date suggests (see Karpoff, 2001), U.S. institutional activism can be ineffective or misplaced.

[54] More generally, banks are not just delegated monitors but also delegated renegotiators; that is they offer a lending relationship; see Bolton and Freixas (2000) and Petersen and Rajan (1994).

[55] See also Krasa and Villamil (1992) and Hellwig (2000a) for generalizations of Diamond's result.

[56] Pension fund managers' incentives to monitor are not backed with a similar disciplining threat. Despite mandatory requirements for activism (at least in the U.S.) pension fund managers do not appear to have strong incentives to monitor managers (see Black, 1990 for a discussion of U.S. regulations governing pension funds' monitoring activities and their effects).

problem in the allocation of credit (see Gorton and Winton, 2003 for an extended survey of these issues).[57]

To summarize, the theoretical literature on bank monitoring shows that delegated monitoring by banks or other financial intermediaries can be an efficient form of corporate governance. It offers one way of resolving collective action problems among multiple investors. However, the effectiveness of bank monitoring depends on bank managers' incentives to monitor. These incentives, in turn, are driven by bank regulation. The existing evidence on bank regulation and banking crises suggests that bank regulation can at least be designed to work when the entire banking system is healthy, but it is often seen to fail when there is a system-wide crisis (see Gorton and Winton, 1998). Thus, the effectiveness of bank monitoring can vary with the aggregate state of the banking industry. This can explain the perception that Japanese banks have played a broadly positive role in the 1970s and 1980s, while in the 1990s they appear to have been more concerned with covering up loan losses than with effectively monitoring the corporations they lend to.

5.4. Board models

The third alternative for solving the collective action problem among dispersed shareholders is monitoring of the CEO by a board of directors. Most corporate charters require that shareholders elect a board of directors, whose mission is to select the CEO, monitor management, and vote on important decisions such as mergers & acquisitions, changes in remuneration of the CEO, changes in the firm's capital structure like stock repurchases or new debt issues, etc. In spirit most charters are meant to operate like a "shareholder democracy", with the CEO as the executive branch of government and the board as the legislative branch. But, as many commentators have argued, in firms with dispersed share ownership the board is more of a "rubber-stamp assembly" than a truly independent legislature checking and balancing the power of the CEO. One important reason why boards are often "captured" by management is that CEOs have considerable influence over the choice of directors. CEOs also have superior information. Even when boards have achieved independence from management they are often not as effective as they could be because directors prefer to play a less confrontational "advisory" role than a more critical monitoring role. Finally, directors generally only have a very limited financial stake in the corporation.

Most regulatory efforts have concentrated on the issue of independence of the board. In an attempt to reduce the CEO's influence over the board many countries have introduced requirements that a minimum fraction of the board be composed of so-called

[57] The moral hazard problem is exacerbated by bank managers' incentives to hide loan losses as Mitchell (2000) and Aghion, Bolton, and Fries (1999) have pointed out. A related problem, which may also exacerbate moral hazard, is banks' inability to commit ex ante to terminate inefficient projects (see Dewatripont and Maskin, 1995). On the other hand, as senior (secured) debt-holders banks also have a bias towards liquidation of distressed lenders (see Zender, 1991 and Dewatripont and Tirole, 1994).

"independent" directors.[58] The rationale behind these regulations is that if directors are not otherwise dependent on the CEO they are more likely to defend shareholders' interests. It is not difficult to find flaws in this logic. For one thing, directors who are unrelated to the firm may lack the knowledge or information to be effective monitors. For another, independent directors are still dependent on the CEO for reappointment. Perhaps the biggest flaw in this perspective is that it does not apply well to concentrated ownership structures. When a large controlling shareholder is in place what may be called for is not only independence from the CEO, but also independence from the controlling shareholder. In corporations with concentrated ownership independent directors must protect the interests of minority shareholders against both the CEO's and the blockholder's actions.

Many commentators view these regulations with much scepticism. To date, most research on boards and the impact of independent directors is empirical, and the findings concerning the effects of independent directors are mixed. Some evidence supporting the hypothesis that independent directors improve board performance is available, such as the higher likelihood that an independent board will dismiss the CEO following poor performance (Weisbach, 1988), or the positive stock price reaction to news of the appointment of an outside director (Rosenstein and Wyatt, 1990). But other evidence suggests that there is no significant relation between firm performance and board composition (e.g. Hermalin and Weisbach, 1991; Byrd and Hickman, 1992; and Mehran, 1995; see Romano, 1996; John and Senbet, 1998; and Hermalin and Weisbach, 2003 for surveys of the empirical literature on boards).

In contrast to the large empirical literature on the composition of boards, formal analysis of the role of boards of directors and how they should be regulated is almost non-existent. An important contribution in this area is by Hermalin and Weisbach (1998). They consider a model where the firm's performance together with monitoring by the board reveals information over time about the ability of the CEO. The extent of monitoring by the board is a function of the board's "independence" as measured by directors' financial incentives as well as their distaste for confronting management. Board independence is thus an endogenous variable. Board appointments in their model are determined through negotiations between the existing board and the CEO. The latter's bargaining power derives entirely from his perceived superior ability relative to alternative managers that might be available. Thus, as the firm does better the CEO's power grows and the independence of the board tends to diminish. As a result CEOs tend to be less closely monitored the longer they have been on the job. Their model highlights an important insight: the gradual erosion of the effectiveness of boards over time. It suggests that regulatory responses should be targeted more directly at the selection process of directors and their financial incentives to monitor management.

[58] A director is defined as "independent" if he or she is not otherwise employed by the corporation, is not engaged in business with the corporation, and is not a family member. Even if the director is a personal friend of the CEO, (s)he will be considered independent if (s)he meets the above criteria.

The model by Hermalin and Weisbach is an important first step in analysing how directors get selected and how their incentives to monitor management are linked to the selection process. Other formal analyses of boards do not explicitly model the selection process of directors. Warther (1998) allows for the dismissal of minority directors who oppose management, but newly selected members are assumed to act in the interest of shareholders.[59] Since directors prefer to stay on the board than be dismissed, his model predicts that directors will be reluctant to vote against management unless the evidence of mismanagement is so strong that they can be confident enough that a majority against management will form. His model thus predicts that boards are active only in crisis situations. One implication of his analysis is that limiting dismissal and/or introducing fixed term limits tends to improve the vigilance of the board.

Raheja (2005) does not model the selection process of directors either. He takes the proportion of independent directors as a control variable. A critical assumption in his model is that independent directors are not as well informed as the CEO and inside directors. He considers two types of board decisions: project choice and CEO succession. Competition for succession is used to induce insiders to reveal the private information they share about project characteristics. Raheja derives the board composition and size that best elicits insider information and shows how it may vary with underlying firm characteristics.

Hirshleifer and Thakor (1994) consider the interaction between inside monitoring by boards and external monitoring by corporate raiders. Takeover threats have a disciplining effect on both management and boards. They show that sometimes even boards acting in the interest of shareholders may attempt to block a hostile takeover.[60]

Adams (2001) focuses on the conflict between the monitoring and advisory functions of the board: the board's monitoring role can restrict its ability to extract information from management that is needed for its advisory role. Thus the model gives insight into the possible benefits of instituting a dual board system, as in Germany.

In sum, the formal literature on boards is surprisingly thin given the importance of the board of directors in policy debates. This literature mainly highlights the complexity of the issues. There is also surprisingly little common ground between the models. Clearly, much remains to be explored. The literature has mainly focused on issues relating to board composition and the selection of directors. Equally important, however, are issues relating to the functioning of the board and how board meetings can be structured to ensure more effective monitoring of management. This seems to be a particularly fruitful area for future research.

[59] See also Noe and Rebello (1996) for a similar model of the functioning of boards.

[60] See also Maug (1997) for an analysis of the relative strengths and weaknesses of board supervision, takeovers and leverage in disciplining management.

5.5. Executive compensation models

Besides monitoring and control of CEO actions another way of improving shareholder protection is to structure the CEO's rewards so as to align his objectives with those of shareholders. This is what executive compensation is supposed to achieve.

Most compensation packages in publicly traded firms comprise a basic salary component, a bonus related to short run performance (e.g. accounting profits), and a stock participation plan (most of the time in the form of stock options). The package also includes various other benefits, such as pension rights and severance pay (often described as "golden parachutes").

Executive compensation in the U.S. has skyrocketed in the past decade, in part as a result of the unexpectedly strong bull market, and in part because of the process of determining compensation packages for CEOs. In most U.S. corporations a compensation committee of the board is responsible for setting executive pay. These committees generally rely on "market standards" for determining the level and structure of pay.[61] This process tends to result in an upward creep in pay standards. U.S. corporations set by far the highest levels of CEO compensation in the world. Although U.S. executives were already the highest paid executives in the world by a wide margin at the beginning of the past decade—even correcting for firm size—the gap in CEO pay has continued to widen significantly over the past decade—largely due to the growing importance of stock options in executive compensation packages (see Murphy, 1999 for an extensive survey of empirical and theoretical work on executive compensation and Hallock and Murphy, 1999 for a reader).

There has always been the concern that although stock options may improve CEOs' incentives to raise share value they are also a simple and direct way for CEOs to enrich themselves and expropriate shareholders. Indeed, practitioners see a grant of an unusually large compensation package as a signal of poor corporate governance (Minow, 2000).

Despite this frequently voiced concern, however, there has been no attempt to analyse the determination of executive pay along the lines of Hermalin and Weisbach (1998), by explicitly modelling the bargaining process between the CEO, the remuneration committee and the Board, as well as the process of selection of committee and board members. Instead, most existing formal analyses have relied on the general theory of contracting under moral hazard of Mirrlees (1976, 1999), Holmstrom (1979) and Grossman and Hart (1983) to draw general conclusions about the structure of executive pay, such as the trade-off between risk-sharing and incentives and the desirability of basing compensation on all performance measures that are informative about the CEO's actions.

[61] Compensation committees often rely on the advice of outside experts who make recommendations based on observed average pay, the going rate for the latest hires, and/or their estimate of the pay expected by potential candidates.

The agency model of Holmstrom and Tirole (1993), which introduces stock trading in a secondary market, can rationalize the three main components of executive compensation packages (salary, profit related bonus, and stock participation), but that does not mean that in practice executive compensation consultants base the design of compensation contracts on fine considerations such as the relative informativeness of different performance measures. On the contrary, all existing evidence suggests that these are not the main considerations for determining the structure of the pay package (see again the extensive survey by Murphy, 1999).

Another complicating factor is that CEOs are driven by both implicit and explicit incentives. They are concerned about performance not only because their pay is linked to performance but also because their future career opportunities are affected. The formal analysis of Gibbons and Murphy (1992) allows for both types of incentives.[62] It suggests that explicit incentives should be rising with age and tenure, as the longer the CEO has been on the job the lower are his implicit incentives.

Finally, much of the agency theory that justifies executive compensation scheme unrealistically assumes that earnings and stock prices cannot be manipulated. This is a major weakness of the theory as brought to light in recent accounting scandals involving Enron, Global Crossing, WorldCom and others. To quote corporate governance expert Nell Minow: "Options are very motivational. We just have to be a little more thoughtful about what it is we're asking them to motivate." [63]

All in all, while the extensive literature on agency theory provides a useful framework for analysing optimal incentive contracts it is generally too far removed from the specifics of executive compensation. Moreover, the important link between executive compensation and corporate governance, as well as the process of determination of executive pay remain open problems to be explored at a formal level.

5.6. Multi-constituency models

The formal literature on boards and executive compensation takes the view that the board exclusively represents the interests of shareholders. In practice, however, this is not always the case. When a firm has a long-term relation with a bank it is not uncommon that a bank representative sits on the board (see Bacon and Brown, 1975). Similarly, it is not unusual for CEOs of firms in related businesses to sit on the board. In some countries, most notably Germany, firms are even required to have representatives of employees on the board. The extent to which boards should be mandated to have representatives of other constituencies besides shareholders is a hotly debated issue. In

[62] See also Holmstrom and Ricart i Costa (1986) and Zwiebel (1995) for an analysis of managerial compensation with implicit incentives. These papers focus on the issue of how career concerns can distort managers' incentives to invest efficiently. In particular they can induce a form of conservatism in the choice of investment projects.

[63] *New York Times*, 2/17/02.

the European Union in particular the issue of board representation of employees is a major stumbling block for the adoption of the European Company Statute (ECS).[64]

As important as this issue is there is only a small formal literature on the subject. What is worse, this literature mostly considers highly stylised models of multiple constituencies. Perhaps the biggest gap is the absence of a model that considers the functioning of a board with representatives of multiple constituencies. Existing models mainly focus on the issue of when and whether it is desirable for the firm to share control among multiple constituencies. These models are too stylised to address the issue of board representation.

5.6.1. Sharing control with creditors

A number of studies have considered the question of dividing control between managers, shareholders and creditors and how different control allocations affect future liquidation or restructuring decisions. A critical factor in these studies is whether share ownership is concentrated or not.

Aghion and Bolton (1992) consider a situation where ownership is concentrated and argue that family-owned firms want to limit control by outside investors because they value the option of being able to pursue actions in the future which may not be profit maximising. They may value family control so much that they may want to turn down acquisition bids even if they are worth more than the net present value of the current business. Or, they may prefer to keep the business small and under family control even if it is more profitable to expand the business. In some situations, however, they may have no choice but to relinquish some if not all control to the outside investor if they want to secure capital at reasonable cost. Aghion and Bolton show that under some conditions the efficient contractual arrangement is to have a state-contingent control allocation, as under debt financing or under standard venture capital arrangements.[65] Although their model only considers a situation of bilateral contracting with incomplete contracts it captures some basic elements of a multi-constituency situation and provides a rationale for extending control to other constituencies than shareholders.

Another rationale for dividing control with creditors (or more generally fixed claim holders) is given in Zender (1991), Diamond (1991, 1993), Dewatripont and Tirole (1994), Berglöf and von Thadden (1994) and Aoki (1990) and Aoki, Hugh, and Sheard (1994). All these studies propose that the threat of termination (or liquidation) if performance is poor may be an effective incentive scheme for management. But, in order to credibly commit to liquidate the firm if performance is poor, control must be transferred

[64] Either the ECS would allow German companies to opt out of mandatory codetermination or it would impose mandatory codetermination on all companies adopting the ECS.

[65] The analysis of venture capital contracts in terms of contingent control allocations has been pursued and extended by Berglöf (1994), Hellman (1997) and Neher (1999). More recently, Kaplan and Strömberg (2003) have provided a detailed analysis of control allocation in 100 venture capital contracts. Their analysis highlights the prevalence of contingent control allocations in venture capital contracts.

to fixed claimholders. As these investors get a disproportionate share of the liquidation value and only a fraction of the potential continuation value, they are more inclined to liquidate the firm than shareholders, who as the most junior claimholders often prefer to "gamble for resurrection". The commitment to liquidate is all the stronger the more dispersed debt is, as that makes debt restructuring in the event of financial distress more difficult (see Hart and Moore, 1995; Dewatripont and Maskin, 1995; and Bolton and Scharfstein, 1996).

Interestingly, Berkovitch and Israel (1996) have argued that when it comes to replacing managers, shareholders may be more inclined to be tough than creditors. The reason why a large shareholder is more likely to fire a poorly performing manager is that the shareholder effectively exercises a valuable option when replacing the manager, while the creditor does not. Sometimes the large shareholder may be too eager to replace management, in which case it may be desirable to let creditors have veto rights over management replacement decisions (or to have them sit on the board).

Another way of limiting shareholders' power to dismiss management is, of course, to have a diffuse ownership structure. This is the situation considered by Chang (1992). In his model the firm can only rely on creditors to dismiss management, since share ownership is dispersed. Chang shows that creditors are more likely to dismiss a poorly performing manager the higher the firm's leverage. Since a large shareholder would tend to dismiss poorly performing managers too easily, Chang shows that there is an efficient level of leverage, implementing a particular division of control rights.

5.6.2. Sharing control with employees

Models of corporate governance showing that some form of shared control between creditors and shareholders may be optimal can sometimes also be reinterpreted as models of shared control between employees and the providers of capital. This is the case of Chang's model, where the role of employee representatives on the board can be justified as a way of dampening shareholders' excessive urge to dismiss employees.

But for a systematic analysis of shared governance arrangements one has to turn to the general theory of property rights recently formulated by Grossman, Hart and Moore (see Grossman and Hart, 1986; Hart and Moore, 1990; and Hart, 1995). The central issue in their theory is the so-called "holdup" problem[66], which refers to the potential ex-post expropriation of unprotected returns from *ex ante* (specific)[67] human capital investment. Much of the property-rights theory is concerned with the protection of physical capital (as in Grossman and Hart, 1986), but it also deals with human capital

[66] See Goldberg (1976) and Klein, Crawford, and Alchian (1978) for an early informal definition and discussion of the holdup concept. See also Williamson (1971, 1975, 1979, 1985a) for a discussion of the closely related concept of opportunism.

[67] It is only when investment is specific to a relation, or a task, that concerns of ex-post expropriation arise. If investment is of a general purpose, then competition ex-post for the investment provides adequate protection to the investor.

investments. An extreme example of "holdup" problem for human capital investments is the case of a researcher or inventor, who cannot specify terms of trade for his invention before its creation. Once his machine or product is invented, however, the inventor can only extract a fraction of the total value of the invention to his clients (assuming there is limited competition among clients). What is worse, the ex-post terms of trade will not take into account the research and development costs, which are "sunk" at the time of negotiation. The terms of trade the inventor will be able to negotiate, however, will be greater if he owns the assets that are required to produce the invention, or if he sits on the board of directors of the client company.

As this example highlights, a general prediction of the theory of property rights is that some form of shared control with employees is efficient, whenever employees (like the inventor) make valuable firm-specific human-capital investments.[68]

Building on this property-rights theory, Roberts and Van den Steen (2000) and Bolton and Xu (2001) provide a related justification for employee representation on the board to Chang's. They consider firms in professional service or R&D intensive industries, where firm-specific human capital investment by employees adds significant value. As in Hart and Moore (1990), say, an important issue in these firms is how to protect employees against the risk of ex-post expropriation or hold-up by management or the providers of financial capital. More concretely, the issue is how to guarantee sufficient job security to induce employees to invest in the firm. Indeed, as with any provider of capital (financial or human), employees will tend to under-invest in firm-specific human capital if they do not have adequate protection against ex-post hold ups and expropriation threats. They show that in firms where (firm-specific) human capital is valuable it may be in the interest of the providers of capital to share control with employees, although generally the providers of financial capital will relinquish less control to employees than is efficient. Indeed, the providers of financial capital are concerned as much with extracting the highest possible share of profits as with inducing the highest possible creation of profits through human capital investments.[69]

[68] The property-rights theory also provides a useful analytical framework to assess the costs and benefits of privatisation of state-owned firms. Thus Hart, Shleifer, and Vishny (1997) have argued that privatised firms have a better incentive to minimize costs, but the systematic pursuit of profits may also lead to the provision of poorer quality service. They apply their analysis to the case of privatisation of prisons. Perhaps, a more apt application might have been to the privatisation of railways in the U.K. and the Netherlands, where quality of service has visibly deteriorated following privatisation. Schmidt (1996) and Shapiro and Willig (1990) emphasize a different trade-off. They argue that under state ownership the government has better information about the firm's management (that is the benefit), but the government also tends to interfere too much (that is the cost). Bolton (1995) looks at yet another angle. He argues that state ownership is actually a form of governance with extreme dispersion of ownership (all the citizens are owners). This structure tends to exacerbate problems of self-dealing. These problems, however, are not always best dealt with through privatisation, which may also involve shareholder dispersion. Pointing to the example of Chinese Township and Village enterprises, Bolton argues instead that state ownership at the community level may be another way of mitigating the inefficiencies of state-owned firms.

[69] Again, see Aghion and Bolton (1987) for a formal elaboration of this point.

Sharing control with employees can be achieved by letting employees participate in share ownership of the company, by giving them board representation, or by strengthening their bargaining power through, say, increased unionisation. An important remark made by Holmstrom (1999) and echoed by Roberts and Van den Steen (2000) is that when employees cannot participate in corporate decision-making a likely response may be unionisation and/or strikes. There are many examples in corporate history where this form of employee protection has proved to be highly inefficient, often resulting in extremely costly conflict resolutions.

Thus, in practice an important effect of employee representation on boards may be that employees' human capital investments are better protected and that shareholders' excessive urge to dismiss employees is dampened. Interestingly, there appears to be some empirical evidence of this effect of employee representation in the study of co-determination in German corporations by Gorton and Schmid (2004). However, their study also suggests that shareholders in Germany do not passively accept board representation by employees. In an effort to counteract employees' influence they tend to encourage the firm to be more highly levered [as Perotti and Spier (1993) have explained, creditors are likely to be tougher in liquidation decisions than shareholders]. Also, in some cases, shareholder representatives have gone as far as holding informal meetings on their own to avoid disclosing sensitive information or discussing delicate decisions with representatives of employees.

An extreme result highlighted by Roberts and Van den Steen (2000) is that it may even be efficient to have employee-dominated boards when only human capital investment matters. Examples of such governance structures are not uncommon in practice, especially in the professional services industry. Most accounting, consulting or law partnerships effectively have employee-dominated boards. Another example is universities, where academics not only have full job security (when they have tenure) but also substantial control rights.[70]

Hansmann (1996) and Hart and Moore (1996, 1998) are concerned with another aspect of governance by employees. They ask when it is best to have 'inside' ownership and control in the form of an employee cooperative or partnership, or when "outside" ownership in the form of a limited liability company is better. A central prediction of the property rights theory is that ownership and control rights should be given to the parties that make ex-ante specific investments. In other words, it should be given mainly to "insiders". Yet, as Hansmann and Hart and Moore observe, the dominant form of governance structure is "outside" ownership. Hansmann resolves this apparent paradox by arguing that often shareholders are the most homogenous constituency in a firm and therefore are generally the best placed group to minimize decision-making costs. He

[70] Bolton and Xu (2001) extend this analysis by considering how internal and external competition among employees can provide alternative or complementary protections to employee control [see also Zingales (1998) for a discussion of corporate governance as a mechanism to mitigate ex-post hold-up problems, and Rajan and Zingales (2000) for an analysis of when a shareholder-controlled firm wants to create internal competition among employees as an incentive scheme].

also accepts Williamson's argument that shareholders are the constituency in most need of protection due to the open-ended nature of their contracts. Hart and Moore (1996, 1998) also focus on distortions in decision-making that can arise in a member cooperative, where members have very diverse interests.[71] They compare these distortions to those that can arise under outside ownership. However, they only consider outside ownership by a single large shareholder and assume away all the governance issues related to dispersed ownership. Like Aghion and Tirole (1997), Burkart, Gromb, and Panunzi (1997), and Pagano and Röell (1998), they argue that a large shareholder will introduce distortions in his attempt to extract a larger share of the firm's value. At the margin he will do this even at the expense of greater value creation. The central observations of their analysis are that employee cooperatives are relatively worse governance structures the more heterogeneous employees are as a group, and outside ownership is relatively better the more the firm faces competition limiting the outside owner's ability to extract rents. They apply their analytical framework to explain why greater worldwide financial integration, which has resulted in increased competition among stock exchanges, has led to a move towards the incorporation of exchanges.

To summarize, the property rights theory of Grossman, Hart and Moore provides one basic rationale for sharing corporate control with employees and for employee representation on the board: protection of employees' firm-specific investments. But there may be others, like potentially better monitoring of management by employees. Indeed, the latter are likely to be better informed than shareholders about the management's actions, and they may be in a better position to monitor the management of, say, company pension plans. As persuasive as these reasons may be, however, it does not follow that rules mandating employee representation on the board, as in Germany, are necessarily desirable. As we have argued above, such rules can only be justified by appealing to a contractual failure of some kind. As we have already mentioned, one important potential source of contractual failure under sequential contracting, may arise when the providers of capital and the entrepreneur design the corporate charter partly as a means of extracting future potential rents from employees (see Aghion and Bolton, 1987; and Scharfstein, 1988). Another possible failure, as Aghion and Bolton (1987), Aghion and Hermalin (1990), Spier (1992) and Freeman and Lazear (1995) have argued, may be due to the firm's founders' concern that allowing for employee representation may send a bad signal to potential investors.

But, even if contractual failures exist, they must be weighed against other potential inefficiencies that may arise as a result of multi-constituency representation on the board, such as shareholder responses to weaken employee influence, greater board passivity or less disclosure of valuable but divisive information by management. One argument against multiple constituencies that is sometimes voiced is that when the firm's management is required to trade off the interests of different constituencies one important "side

[71] It has often been highlighted that an important source of conflict in member cooperatives is the conflict between old and young members. The former want to milk past investments, while the younger members want to invest more in the firm (see Mitchell, 1990).

effect" is that management gains too much discretion. When the stock tanks management can always claim that it was acting in the interest of employees (see, for example, Macey, 1992; Tirole, 2001; Hart, 1995; or Jensen, 2002). This argument is particularly relevant when defining the CEO's fiduciary duties (or "mission"). If these duties are too broadly defined to include the interests of multiple constituencies they are in danger of becoming toothless. The current narrow definition of fiduciary duties in the U.S. is already balanced by the "business judgement rule", which makes it difficult for plaintiffs to prevail. If one were to add a "protection of other constituencies rule" it is likely that winning a suit would be even harder.

However, note that as relevant as this argument is when applied to the definition of the fiduciary duties of the CEO, it is less so when applied to board representation. Having representatives of creditors, employees or related firms on the board does not *per se* increase the manager's discretion. The manager is still monitored by the board and will still have to deal with the majority of directors that control the board, just as in any democracy the power of the executive branch of government is held in check by the majority in control of the legislature, no matter how diverse the representation of the legislature is. Unfortunately, a systematic analysis of these issues remains to be done, as there are no formal models of the functioning of boards with representation of multiple constituencies. Nor are there comparative empirical studies analysing the differences in managerial accountability and discretion in Germany and other countries.

Finally, as the introduction of mandatory employee representation has both efficiency and distributive effects there must be a sufficiently strong political constituency supporting such rules. Although the link between politics and corporate governance regulation is clearly relevant there has been virtually no formal modelling of this link. A recent exception is Pagano and Volpin (2005a) who derive the degree of investor protection endogenously from a political equilibrium between "rentier", management and employees.[72] They show that depending on the relative political power of these constituencies, different laws on shareholder protection will be enacted. Thus, if the employee constituency is large and powerful as, say in Italy, then laws will be less protective of shareholder interests.[73]

6. Comparative perspectives and debates

As Sections 4 and 5 illustrate, the core issues of corporate governance: how to decide who should participate in corporate governance, how to solve the collective action problem of supervising management, how to regulate takeovers and the actions of large

[72] A second paper by Pagano and Volpin (2005b) shifts the focus to the internal politics of the firm, arguing that there is a natural alliance between management and employees in staving off hostile bids.

[73] As we discuss below, there has been substantially more systematic historical analysis of the link between politics and corporate governance, most notably by Roe (1994), who argues that weak minority shareholder protection is the expected outcome in social democracies.

investors, how boards should be structured, how managers' fiduciary duties should be defined, what are appropriate legal actions against managerial abuses, all these issues have no unique simple answer. Corporations have multiple constituencies and there are multiple and interlocking tradeoffs. Different solutions may be needed depending on the type of activity to be financed. Human capital-intensive projects may require different governance arrangements than capital-intensive projects[74]; projects with long implementation periods may require different solutions than projects with short horizons.[75] It is not possible to conclude on the basis of economic analysis alone that there is a unique set of optimal rules that are universally applicable to all corporations and economies, just as there is no single political constitution that is universally best for all nations.

The practical reality of corporate governance is one of great diversity across countries and corporations. An alternative line of research that complements the formal analyses described in the previous section exploits the great diversity of corporate governance rules across countries and firms, attempting to uncover statistical relations between corporate governance practice and performance or to gain insights from a comparative institutional analysis. A whole sub-field of research has developed comparing the strengths and weaknesses of corporate governance rules in different countries. In this section we review the main comparative perspectives on governance systems proposed in the literature.[76]

6.1. Comparative systems

Broadly speaking and at the risk of oversimplifying, two systems of corporate governance have been pitted against each other: the Anglo-American market based system and the long-term large investor models of, say, Germany and Japan. Which of these systems has been most favored by commentators has varied over time as a function of the relative success of each country's underlying economy, with two broad phases: the 1980s—when the Japanese and German long-term investor corporate governance perspective were seen as strengths relative to the Anglo-American market based short-termist perspective—and the 1990s—when greater minority shareholder protections and the greater reliance on equity financing in the Anglo-American systems were seen as major advantages.[77]

[74] See, for example, Allen and Gale (2000), Maher and Andersson (2000), Rajan and Zingales (2000) and Roberts and Van den Steen (2000) for discussions of how corporate governance may vary with underlying business characteristics.

[75] See Maher and Andersson (2000) and Carlin and Mayer (2003) for a discussion of corporate governance responses in firms with different investment horizons.

[76] For recent surveys of the comparative corporate governance literature see Roe (1996), Bratton and McCahery (1999) and Allen and Gale (2000);. see also the collections edited by Hopt et al. (1998), and Hopt and Wymeersch (2003).

[77] The comparative classifications proposed in the literature broadly fit this (over)simplification. Commentators have distinguished between "bank oriented" and "market oriented" systems (e.g. Berglöf, 1990) and

Japanese and German corporate governance looked good in the 1980s when Japan and Germany were growing faster than the U.S. In contrast, in the late 1990s, following nearly a decade of economic recession in Japan, a decade of costly post-unification economic adjustments in Germany, and an unprecedented economic and stock market boom in the U.S., the American corporate governance model has been hailed as the model for all to follow (see Hansmann and Kraakman, 2001). As we are writing sentiment is turning again in light of the stock market excesses on Nasdaq and the *Neuer Markt*, which have resulted in massive overinvestment in the technology sector, leading to some of the largest bankruptcies in corporate history, often accompanied by corporate governance scandals.[78]

Critics of U.S. governance in the 1980s have argued that Germany and Japan had a lower cost of capital because corporations maintained close relationships with banks and other long-term debt and equity holders. As a result Japan had a low cost of equity[79], Germany a low cost of bank debt and both could avoid the equity premium by sustaining high levels of leverage (see e.g. Fukao, 1995). Despite a convergence of the real cost of debt and equity during the 1980s (McCauley and Zimmer, 1994), they have enjoyed a lower cost of capital than the U.S. and the U.K. As a result, Japanese corporations had higher investment rates than their U.S. counterparts (Prowse, 1990). Interestingly, a revisionist perspective gained prominence in the early 90s according to which the low cost of capital in Japan was a sign of excesses leading to overinvestment (Kang and Stulz, 2000).

Following the stock market crash of 1990, Japan lost its relatively low cost of equity capital, while the U.S. gradually gained a lower cost of equity capital as the unprecedented bull market gained steam. This lower cost of equity capital in the U.S. has been seen by many commentators as resulting from superior minority shareholder protections (see e.g. La Porta, Lopez-de-Silanes, and Shleifer, 1998), and was often the stated reason why foreign firms increasingly chose to issue shares on Nasdaq and other U.S. exchanges and why the *Neuer Markt* was booming (see Coffee, 2002;

"insider" versus "outsider" systems (e.g. Franks and Mayer, 1995). These distinctions are based on a range of characteristics of governance and financial systems, such as the importance of long-term bank lending relations, share ownership concentration, stock market capitalisation and regulatory restrictions on shareholder power. More recently, commentators such as La Porta, Lopez-de-Silanes, and Shleifer (1998) attempt no such distinction and introduce a single ranking of countries' corporate governance systems according to the extent of minority shareholder protections as measured by an "anti-director rights index" based on six elements of corporate law. As we shall see, all attempts at objectively classifying country corporate governance systems have been criticised for overemphasising, leaving out or misunderstanding elements of each country's system. Thus, for example, the declining importance of the market for corporate control in the U.S. has generally been overlooked, as well as the lower anti-director rights in Delaware (see Hansmann and Kraakman, 2001; Hansmann et al., 2004). Similarly, bank influence in Germany has often been exaggerated (see Edwards and Fischer, 1994; Hellwig, 2000b), or the importance of stock markets in Japan (La Porta et al., 2000b).

[78] Enron is the landmark case, but there have been many smaller cases on *Neuer Markt* that have these characteristics.

[79] The cost of equity was significantly lower in Japan in the 1980s. This advantage has of course disappeared following the stock market crash.

La Porta et al., 2000b). Similarly the Asian crisis has been attributed to poor investor protections (see Johnson, 2000; and Claessens et al., 2002; and Shinn and Gourevitch, 2002 for the implications for U.S. policy to promote better governance worldwide). Exchanges that adopted NASDAQ-style IPO strategies and investor protections, like the *Neuer Market* in Germany have witnessed a similar boom (and bust) cycle. With the benefit of hindsight, however, it appears that the low cost of equity capital on these exchanges during the late 1990s had more to do with the technology bubble than with minority shareholder protection, just as the low cost of capital in Japan in the late 1980s had more to do with the real estate bubble than with Japanese corporate governance.

Another aspect of Japanese corporate governance that has been praised in the 1980s is the long run nature of relationships between the multiple constituencies in the corporation, which made greater involvement by employees and suppliers possible. It has been argued that this greater participation by employees and suppliers has facilitated the introduction of "just in time" or "lean production" methods in Japanese manufacturing firms (see Womack, Jones, and Roos, 1991). The benefits of these long-term relations have been contrasted with the costs of potential "breaches of trust" following hostile takeovers in the U.S. (Shleifer and Summers, 1988).[80]

One of the main criticisms of Anglo-American market-based corporate governance has been that managers tend to be obsessed with quarterly performance measures and have an excessively short-termist perspective. Thus, Narayanan (1985), Shleifer and Vishny (1989), Porter (1992a, 1992b) and Stein (1988, 1989), among others, have argued that U.S. managers are myopically "short-termist" and pay too much attention to potential takeover threats. Porter, in particular, contrasts U.S. corporate governance with the governance in German and Japanese corporations, where the long-term involvement of investors, especially banks, allowed managers to invest for the long run while, at the same time, monitoring their performance. Japanese *keiretsu* have also been praised for their superior ability to resolve financial distress or achieve corporate diversification (see e.g. Aoki, 1990; and Hoshi, Kashyap, and Scharfstein, 1990). This view has also been backed by critics in the U.S., who have argued that populist political pressures at the beginning of the last century have led to the introduction of financial regulations which excessively limit effective monitoring by U.S. financial institutions and other large investors, leading these authors to call for larger and more active owners (see Roe, 1990, 1991, 1994; Black, 1990).[81]

[80] As "lean production" methods have successfully been implemented in the U.S., however, it has become clear that these methods do not depend fundamentally on the implementation of Japanese-style corporate governance (Sabel, 1996).

[81] Interestingly, even the former chairman of the Securities and Exchange Commission argued against "over-regulation" and "short-termism" (Grundfest, 1993) and for "investors' ability to monitor corporate performance and to control assets that they ultimately own", an ability that the U.S. regulatory systems has "subordinated to the interests of other constituencies, most notable corporate management" (Grundfest, 1990, pp. 89–90). The call for more active (and larger) owners is also typical of U.S. shareholder activists (see Monks and Minow, 2001).

In the 1990s the positive sides of Anglo-American corporate governance have gradually gained greater prominence. Hostile takeovers were no longer criticised for bringing about short-termist behaviour. They were instead hailed as an effective way to break up inefficient conglomerates (Shleifer and Vishny, 1997b).[82] Most commentators praising the Anglo-American model of corporate governance single out hostile takeovers as a key feature of this model. Yet, starting in the early 1990s the market for corporate control in the U.S. has essentially collapsed.[83] Indeed, following the wave of anti-takeover laws and charter amendments introduced at the end of the 1980s, most U.S. corporations are now extremely well protected against hostile takeovers.[84] Their control is generally no longer contestable.[85] In contrast, in the U.K. the City Code prevents post-bid action that might frustrate the bid and few companies have put in place pre-bid defences, thus making the U.K. the only OECD country with an active and open market for corporate control.[86]

An influential recent classification of corporate governance systems has been provided by La Porta et al. (1997, 1998). The authors show that indices designed to capture the degree of investor protection in different countries correlate very strongly with a classification of legal systems based on the notion of "legal origin" (inspired by David and Brierley, 1985).[87] In a series of papers the authors go on to show that legal ori-

[82] See Stein (2003) in this handbook for a survey of the conglomerate literature.

[83] See Comment and Schwert (1995) for the early 1990s and Bebchuk, Coates, and Subramanian (2002) for 1996–2000.

[84] See Danielson and Karpoff (1998) for a detailed analysis of takeover defences in the U.S. Grundfest (1993) observed: "The takeover wars are over. Management won. [. . .] As a result, corporate America is now governed by directors who are largely impervious to capital market electoral challenges."

[85] The introduction of the anti-takeover laws has also shifted perceptions on state corporate law competition. This competition is not depicted as a "race to the bottom" anymore as in Cary (1974) or Bebchuk (1992). Instead Romano (1993) has argued in her influential book, entitled "the Genius of American Law", that competition between states in the production of corporate law leads to better laws. She goes as far as recommending the extension of such competition to securities regulation (Romano, 1998). On the other hand, Bebchuk and Ferrell (1999, 2001) have argued that it is hard to justify the race to pass anti-takeover laws as a race to the top. Supporting their view, Kamar (1998) has pointed out that network effects can create regulatory monopolies and that limited state competition may therefore be consistent with the existence of inferior standards that are hard to remove. He goes on to argue that the break up of the monopoly of the SEC over securities regulation could lead to convergence to the standards of the dominant producer of corporate law, Delaware.

[86] In the U.K. institutional investors have larger holdings and regulation allows them to jointly force companies to dismantle their pre-bid defences. For example, in the mid-1970s Lloyds Bank wanted to cap votes at 500 votes per shareholders, which would have left the largest twenty shareholders commanding 16% of the voting rights with 0.01% each. Institutional investors threatened to boycott Lloyd's issues and the plan was dropped (Black and Coffee, 1994). In 2001 institutional investors "encouraged" British Telecom to rescind a 15% ownership and voting power ceiling, a powerful pre-bid defence dating back to BT's privatisation.

[87] The La Porta et al. (1997, 1998) indices do not cover securities regulation and have been widely criticised, both conceptually and because the numbers are wrong for certain countries. Of course the direct correlation between "legal origin" and other variables is not affected by such criticism. Pistor (2000) broadens and improves the basic index design for a cross-section of transition countries. She shows that improvements in the

gin correlates with the size of stock markets,[88] ownership concentration, the level of dividend payments[89], corporate valuation and other measures of the financial system across a large cross-section of countries (La Porta et al., 1997, 1999, 2000a, 2002).[90] Other authors have applied the legal origin view to issues like cross-border mergers and the home bias.[91] Stulz and Williamson (2003) add language and religion (culture) as possible explanatory variables.

In the same vein the regulatory constraints in the U.S. that hamper intervention by large shareholders, previously criticised for giving too much discretion to management (e.g. by Roe, 1990, 1991, 1994; Black, 1990; and Grundfest, 1990), have been painted in a positive light as providing valuable protections to minority shareholders against expropriation or self-dealing by large shareholders, reversing the causality of the argument (see La Porta et al., 2000b; and Bebchuk, 1999, 2000).[92] In a recent reply, Roe (2002) argues that this argument is misconceived because it is based on a misunderstanding of corporate law. Law imposes very few limits on managerial discretion and agency costs, particularly in the United States, suggesting that the correlation between classifications of corporate law and ownership concentration is spurious or captures the influence of missing variables, for example the degree of product market competition. More damagingly, recent historical evidence shows that investor protection in the United Kingdom was not very strong before WWII (Cheffins, 2002), but ownership had already dispersed very quickly (Franks, Mayer, and Rossi, 2003).

Recently, some commentators have gone as far as predicting a world-wide convergence of corporate governance practice to the U.S. model (see e.g. Hansmann and Kraakman, 2001).[93] In a variant of this view, world-wide competition to attract corporate headquarters and investment is seen like the corporate law competition between

index levels were larger in countries that implemented voucher privatisations (opted for ownership dispersion), concluding that corporate finance drives changes in the index levels, not legal origin.

[88] Rajan and Zingales (2003) show that the correlation of legal origin and the size of stock markets did not hold at the beginning of the century.

[89] On corporate governance and payout policies see Allen and Michaely (2003).

[90] La Porta et al. (2000b) provide a summary of this view.

[91] The "legal origin" view's prediction that bidders from common law countries increase the value of civil law targets, because the post-bid entity has (value-enhancing) common law level investor protection is supported by recent studies of cross-border mergers (Bris and Cabolis, 2002; Rossi and Volpin, 2004). At the same time, recent acquisitions by U.S. (common law) firms were generally poor, producing very large losses in value for larger acquirors in particular (Moeller, Schlingemann, and Stulz, 2004, 2005). Dahlquist et al. (2003) relate investor protection to the size of free float in different countries and the "home bias".

[92] This reversal of causality is particularly important in the context of emerging markets because it provides and alternative "ex-post" rationalisation of the voucher privatisation experiment in the Czech Republic.

[93] Hansmann and Kraakman (2001) call the U.S. model the "standard shareholder oriented model". In the shareholder model "ultimate control over the corporation should be in the hands of the shareholder class; [..] managers [..] should be charged with the obligation to manage the corporation in the interests of its shareholders; [..] other corporate constituencies, such as creditors, employees, suppliers, and customers should have their interests protected by contractual and regulatory means rather than through participation in corporate governance; [..] non-controlling shareholders should receive strong protection from exploitation at the hands of controlling shareholders; [..] the principalmeasure of the interests of the public corporation's shareholder-

U.S. states portrayed by Romano (1993). Such competition is predicted to eventually bring about a single standard resembling the current law in Delaware or, at least, securities regulation standards as set by the U.S. SEC (see Coffee, 1999).[94]

Although few advocates of the Anglo-American model look back at the 1980s and the perceived strengths of the Japanese and German models at the time, there have been some attempts to reconcile these contradictions. Thus, some commentators have argued that poison pill amendments and other anti-takeover devices are actually an improvement because they eliminate partial bids "of a coercive character" (Hansmann and Kraakman, 2001). Others have also argued that the market for corporate control in the U.S. is more active than elsewhere, suggesting that U.S. anti-takeover rules are less effective than anti-takeover measures elsewhere (La Porta, Lopez-de-Silanes, and Shleifer, 1999). Finally, Holmstrom and Kaplan (2001) have argued that the hostile takeovers and leveraged buyouts of the 1980s are no longer needed as U.S. governance "has reinvented itself, and the rest of the world seems to be following the same path".[95]

Following the scandals dissatisfaction with U.S. corporate governance is on the rise again. There is little doubt that the Enron collapse, the largest corporate bankruptcy in U.S. history to date, was caused by corporate governance problems. Yet Enron had all the characteristics of an exemplary "Anglo-American" corporation. As stock prices are falling executive remuneration (compensation) at U.S. corporations looks increasingly out of line with corporate reality. At the same time the global corporate governance reform movement is pressing ahead, but not necessarily by imitating the U.S. model.[96] The most visible manifestations are corporate governance codes that have been adopted in most markets, except the U.S.[97]

6.2. *Views expressed in corporate governance principles and codes*

Following the publication of the Cadbury Report and Recommandations (1992) in the U.K., there has been a proliferation of proposals by various committees and interest groups on corporate governance principles and codes.[98] These policy documents have

ers is the market value of their shares in their firm." They contrast this "standard model" with the "manager oriented model", the "labour oriented model", the "state-oriented model" and the "stakeholder model".

[94] In Europe, The Netherlands now seems to be taking on Delaware's role. Andenas, Hopt, and Wymeersch (2003) survey the legal mobility of companies within the European Union.

[95] Holmstrom and Kaplan (2001) emphasise that the lucrative stock option plans of the 90s have replaced the disciplinary role of hostile takeovers and debt (see compensation section). They also stress the role of activist boards and investors (*op. cit*., p. 140).

[96] Indeed, on takeover regulation many countries are explicitly rejecting the U.S. model adopting mandatory bid rules and not the Delaware rules. At the same time pension funds are lobbying corporations to take into account the interests of multiple constituencies, under the banner of "corporate social responsibility".

[97] There are indications that, as a result of the Enron collapse, the U.S. too will join in this global development originating from other shores.

[98] The Cadbury Report and Recommandations (1992) is the benchmark for corporate governance codes. Cadbury also set the agenda on issues and provided an example of "soft regulation" the business community in

been issued by institutional investors and their advisors, companies, stock exchanges, securities markets regulators, international organisations and lawmakers.[99] We briefly take stock of these views here and contrast them with the general economic principles discussed in the models section (Section 5) as well as the available empirical evidence (Section 7).[100]

Codes provide recommendations on a variety of issues such as executive compensation, the role of auditors, the role of non-shareholder constituencies and their relation with the company, disclosure, shareholder voting and capital structure, the role of large

other countries was quick to endorse and emulate, for example the "comply or explain" principle of enforcement via moral suasion and implicit contracts. However, Cadbury did not invent the governance wheel. The subject was already receiving attention in Commonwealth countries like Hong Kong (1989) and Australia (1991).

Internationally, the OECD (1999) "Principles of Corporate Governance" have been the main catalyst for the development of further codes and a driver of law reform (see www.oecd.org). The OECD Principles were a direct response to the Asia/Russia/Brazil crisis (see Section 3.5).

In the U.K. Cadbury was followed by Greenbury (1995), Hampel (1998) and the "Combined Code". Other Commonwealth countries followed suit: Canada (Toronto Stock Exchange Committee on Corporate Governance, 1994), South Africa (King Committee, 1994), Thailand (SET, 1998), India (Confederation of Indian Industry, 1998), Singapore (Stock Exchange of Singapore, 1998), Malaysia (High Level Finance Committee on Corporate Governance, 1999) and the Commonwealth Association (1999).

In Continental Europe, corporate governance principles, recommendations and "codes of best practice" are also numerous. France has seen two Viénot Report (1995, updated in 1999), the Netherlands the Peters Report (1997), Spain the Olivencia Report (1998) and Belgium the Cardon Report (1998). Greece, Italy and Portugal followed in 1999, Finland and Germany in 2000, Denmark in 2001 and Austria in 2002. The European Association of Securities Dealers was first to issue European Principles and Recommendations (2000), followed by Eurochareholders (2000). From the investor side, there have been statements from France (AFG-ASFFI, 1998), Ireland (IAIM, 1992), Germany (DSW, 2000), the U.K. (PIRC, 1993, 1996, 1999; Hermes, 1999). These documents can be found at www.ecgi.org/codes.

In Asia, guidelines have been written for Japan (1998) and Korea (1999), in addition to the Commonwealth countries already mentioned. In Latin America, Brazil (1999), Mexico (1999) and Peru (2002) have their own guidelines. Undoubtedly, other countries are sure to follow.

In the U.S., there is no "Code" as such but corporations have been issuing corporate governance statements [e.g. General Motors' guidelines (1994), the National Association of Corporate Directors (NACD, 1996) and the Business Roundtable (1997)]. Pension funds also issue their own corporate governance principles, policies, positions and voting guidelines (TIAA-CREF, 1997; AFL-CIO, 1997; CalPERS, 1998 (see www.ecgi.org/codes); Confederation of Indian Industry, 1998, revised 1999). The American Bar Association published a "Directors Guidebook" (1994). The American Law Institute (1994) adopted and promulgated its "Principles of Corporate Governance" in 1992. Although not binding in nature, these principles are widely cited in U.S. case law.

[99] The codes have triggered an avalanche of corporate governance statements from companies often leading to the creation of new jobs, job titles ("Head of Corporate Governance"), competence centres and task-forces within companies. From the investors' side, countries and companies are starting to be ranked and rated according to corporate governance benchmarks. The proposals tabled at shareholder meetings are scrutinised and compared "best practice".

[100] Not all policy documents mentioned here are included in the list of references. An extensive list, full text copies and international comparisons (in particular Gregory, 2000, 2001a, 2001b, 2002) can be found on the codes pages of the European Corporate Governance Institute (www.ecgi.org).

shareholders and anti-takeover devices. But a quick reading of these codes quickly reveals their dominant focus on boards and board-related issues.[101] Topics covered by codes include: board membership criteria, separation of the role of chairman of the board and CEO, board size, the frequency of board meetings, the proportion of inside versus outside (and independent) directors, the appointment of former executives as directors, age and other term limits, evaluation of board performance, the existence, number and structure of board committees, meeting length and agenda, and assignment and rotation of members.[102] Interestingly, many of the most prominent concerns articulated in codes are not echoed or supported in current empirical research, as we will discuss in Section 7. The striking schism between firmly held beliefs of business people and academic research calls for an explanation. For instance, why do independent directors feature so prominently in codes but appear to add so little in event studies and regressions? Equally, why do institutional investors attach so much importance to the separation of the roles of chairman of the board and CEO, while the empirical evidence suggests that this separation hardly matters?

6.3. Other views

Some commentators of comparative corporate governance systems attempt to go beyond a simple comparison of one system to another. Thus, although Black (1990, 1998) criticises U.S. corporate governance rules for excessively raising the costs of large shareholder intervention, he is also critical of other countries' corporate governance standards. He argues that all countries fall short of what he would like U.S. governance to look like (Black, 2000a).[103] Taking a radically different and far more optimistic perspective Easterbrook (1997) has argued that no global standards of corporate governance are needed because "international differences in corporate governance are attributable more to differences in markets than to differences in law" (see also Easterbrook and Fischel, 1991). Since markets are unlikely to converge, neither will the law. Although some fine-tuning might be required locally, market forces will automatically create the regulatory underpinnings national systems need.

7. Empirical evidence and practice

The empirical literature on corporate governance is so extensive that it is a daunting task to provide a comprehensive survey in a single article. Fortunately a number of

[101] Gregory (2001a) compares 33 codes from 13 member states of the European Union and two pan-European codes to the OECD Principles. All the international and 28 national codes provide a board job-description and all the codes cover at least one board related issue. In contrast, only about 15 national codes cover anti-takeover devices. A similar picture emerges from comparisons of codes from outside the EU (Gregory, 2000, 2001b).

[102] Again, see Gregory (2000, 2001a, 2001b) for an extensive listing and comparisons.

[103] See Avilov et al. (1999), Black, Kraakman, and Hay (1996) and Black (2000b) in the context of emerging markets.

surveys of specific issues have appeared recently.[104] We shall to a large extent rely on these surveys and only cover the salient points in this section. In the introduction we have defined five different approaches to resolving collective action problems among dispersed shareholders: (i) hostile takeovers, (ii) large investors, (iii) boards of directors, (iv) CEO incentive schemes and (v) fiduciary duties & shareholder suits. Each of these approaches has been examined extensively and recent surveys have appeared on takeovers (Burkart, 1999),[105] the role of boards (Romano, 1996; and Hermalin and Weisbach, 2003), shareholder activism (Black, 1998; Gillan and Starks, 1998; Karpoff, 2001; and Romano, 2001), CEO compensation (Core, Guay, and Larcker, 2003; Bebchuk, Fried, and Walker, 2002; Gugler, 2001; Perry and Zenner, 2000; Loewenstein, 2000; Abowd and Kaplan, 1999; Murphy, 1999) and large shareholders (Short, 1994; Gugler, 2001[106] and Holderness, 2003). Not even these surveys cover everything. In particular research on the role of large investors is not fully surveyed—partly because research in this area has been rapidly evolving in recent years. The literature on fiduciary duties and shareholder suits is very limited.

7.1. Takeovers

Hostile takeovers are a powerful governance mechanism because they offer the possibility of bypassing the management to take permanent control of the company, by concentrating voting and cash-flow rights.[107] Corporate governance codes endorse hostile takeovers and the voting guidelines issued by investor groups come out very strongly against anti-takeover devices and for the mandatory disclosure of price sensitive information and toeholds.[108] Paradoxically disclosure and insider trading laws may actually make hostile takeovers harder, as Grossman and Hart (1983) have noted. Indeed, the market for corporate control should work better in regulatory environments with low shareholder protection and lax disclosure standards, so bidder incentives are not eroded by the free-riding problem. On the other hand, low shareholder protection can also give rise to excessive takeover activity by empire builders. Anti-takeover protections reduce the threat of hostile takeovers but both theory and empirical evidence suggest that they

[104] An earlier general survey taking an agency perspective is Shleifer and Vishny (1997a).

[105] Andrade, Mitchell, and Stafford (2001) survey the stylised facts on takeovers and mergers in the U.S. 1973–1998.

[106] Gugler (2001) surveys the English-language literature and draws on national experts to survey the local language literatures in Austria, Belgium, Germany, France, Italy, Japan, The Netherlands, Spain and Turkey.

[107] In the U.S. control changes often require board approval. In countries like the U.K. the bidder bypasses the management *and* the board; the change of control decision is the sovereign right of the target shareholders.

[108] For example, the OECD (1999) Principle I.E states that the "markets for corporate control should be allowed to function in an efficient and transparent manner". The Euro-Shareholders Guidelines (2000) state that "anti-takeover defences or other measures which restrict the influence of shareholders should be avoided" (Recommendation 3) and that "companies should immediately disclose information which can influence the share price, as well as information about those shareholders who pass (upwards or downwards) 5% thresholds" (Recommendation 5).

also strengthen the bargaining position of the target for the benefit of target shareholders. Finally, it is important to keep in mind that hostile takeovers are difficult to finance even in the most liquid capital markets. Despite their alleged importance, hostile takeovers are isolated instances and their study has been largely confined to the U.S. and the U.K.

7.1.1. Incidence of hostile takeovers

Takeovers are well publicized, but in sheer numbers they are relatively rare events. Even at the peak of the U.S. takeover wave in the 1980s, takeover rates (the number of bids as a percentage of the number of listed companies) rarely exceeded 1.5% and declined steeply afterwards (Comment and Schwert, 1995).[109] Hostile takeovers, the events that are of interest here, are even more elusive. Under standard definitions, even at their pre-1990 peak hostile bids never represented more than 30% of all U.S. deals (Schwert, 2000).[110] Between 1990 and 1998 only 4% of all U.S. deals were hostile at some stage and hostile bidders acquired 2.6% of the targets (Andrade, Mitchell, and Stafford, 2001).[111] The paucity of hostile deals is also evident outside the U.S.; however, there is an unusually high amount of hostile activity in Europe in 1999 (Table 2).

If hostile takeovers are a disciplining device for management they should predominantly affect poorly performing firms. This prediction is not borne out by the available empirical evidence. Successful U.S. takeover targets are smaller than other companies, but otherwise they do not differ significantly from their peers (Comment and Schwert, 1995).[112] The targets of hostile bids are likely to be larger than other targets.[113] Indicators of poor target management contribute little or are not significant (Schwert, 2000).[114] The available evidence for the U.K. also fails to show that the targets of

[109] The causes of such cycles in takeover activity are many, and their relative importance is an open issue. The 1980s U.S. takeover boom has been attributed to, *inter alia*, the 1986 Tax Reform Act and to the 1978 Bankruptcy Act; see Kaplan (1994b) for a discussion of the latter point.

[110] Other characteristics of U.S. hostile deals are that they are more likely to involve cash offers and multiple bidders. Also, hostile bids are less likely to succeed than uncontested bids (Schwert, 2000).

[111] For 1973–1979 8.4% of all deals were hostile at some stage, between 1980–1989 14.3%; hostile acquisitions were 4.1% and 7.1% respectively (Andrade, Mitchell, and Stafford, 2001). The full merger sample covers 4,300 completed deals on the CRSP tapes, covering all U.S. firms on the NYSE, AMEX and Nasdaq between 1973–1998.

[112] Comment and Schwert (1995) estimate the probability of a successful takeover as a function of anti-takeover devices, abnormal returns, sales growth, the ration of net-liquid assets to total assets, debt/equity ratios, maket/book ratios, price/earnings ratios and total assets (size) for 1977–1991. They report that the results for hostile takeovers do not differ significantly (p. 34). We discuss the anti-takeover device evidence below.

[113] This is consistent with the view that bids in the U.S. are classified as hostile when the target boards have a lot of bargaining power. The boards of larger companies are more likely to reject a bid, at least initially, to obtain a higher premium.

[114] Schwert (2000) covers the period 1975–1996 and considers four definitions of "hostile bid". He concludes that "the variables [..] that might reflect poor management, market to book ratios and return on assets, contribute little. The variables [..] that probably reflect the bargaining power of the target firm, such as firm size and the secular dummy variables, contribute most explanatory power" (p. 2624).

Table 2
Number of takeovers by region

	Australia	Canada	U.S.	EU15			Other
				Total	U.K.	ex-U.K.	
Number of announced uncontested takeovers**							
1989	81	184	1,188	550	316	234	114
1990	69	193	834	597	290	307	188
1991	107	269	790	817	252	565	363
1992	46	194	746	824	181	643	296
1993	100	215	789	803	196	607	456
1994	124	224	1,015	816	221	595	614
1995	162	296	1,106	806	219	587	753
1996	142	277	1,115	676	195	481	745
1997	107	258	1,150	574	201	373	726
1998	103	231	1,203	653	234	419	893
1999	100	289	1,236	801	271	530	1,180
Number announced contested takeovers***							
1989	3	6	45	36	32	4	10
1990	2		12	24	22	2	5
1991	8	1	7	34	31	3	2
1992	10	2	7	20	15	5	4
1993	10	1	11	15	11	4	5
1994	8	11	33	11	8	3	4
1995	18	19	59	22	14	8	7
1996	22	8	45	20	13	7	11
1997	12	17	27	23	11	12	5
1998	12	14	19	14	12	2	5
1999	15	6	19	42	21	21	6

Source: Thomson Financial Services Data (TFSD) and own calculations.
**Under the TFSD definition a tender offer that was recommended by board of the target company to its shareholders.
***Under the TFSD definition a tender offer that was initially rejected by the board of the target company.

successful hostile bids had poorer pre-bid performance than other targets (Franks and Mayer, 1996).[115]

Hostile takeover activity in the U.S. sharply declined after 1989. Most observers agree that managers effectively lobbied for protection from the market for corporate control.

[115] Franks and Mayer (1996) cover the period 1980 to 1986 and consider the pre-bid evolution of share prices (abnormal returns), dividend payouts, cash-flows and Tobin's Q. They find a 14 point difference in abnormal returns between successful hostile bids and accepted bids that is not statistically significant, a significant difference in Tobin's Q but no difference in dividend payouts or cash-flows. On Tobin's Q they observe that all values are larger than one, suggesting poor relative rather than absolute performance. Finally, companies with control changes have higher pre-bid stock returns that companies without control changes, the opposite of what the poor management hypothesis predicts.

The tightening of insider trading laws in the second half of the 1980s, a series of landmark cases in Delaware in 1985 and a new wave of anti-takeover laws made it virtually impossible to take over U.S. corporations without target board consent (see 7.1.4 below). As a result, few hostile takeover attempts were made and less than 25% of the bidders succeeded in taking control of the target (Bebchuk, Coates, and Subramanian, 2002). Another explanation attributes the decline in takeover activity to the demise of the junk bond market, the business cycle and the credit crunch associated with the Savings and Loans crisis (Comment and Schwert, 1995). Takeover activity has recently emerged in continental Europe in a number of spectacular cases where there were none before. Although there is no conclusive evidence in support it is possible that this change has brought about more managerial discipline. It is also a sign of the waning protection of national champions by European governments.

7.1.2. Correction of inefficiencies

If hostile takeovers correct managerial failure and enhance efficiency the value of the bidder and the target under joint control (V_{AB}) should be larger than the value of the bidder (V_A) and the target (V_B) separately, or $\Delta V \equiv [V_{AB} - V_A - V_B] > 0$. Generally, the change in value (ΔV) is taken to be the difference between the stand-alone pre-bid and the combined post-bid values in event studies. Other measures are based on changes in accounting data, such as cash flows or plant level productivity. Event studies find sizeable average premia ($\sim$ 24%) going to target shareholders in all U.S. acquisitions (Andrade, Mitchell, and Stafford, 2001) and higher premia for hostile takeovers (Schwert, 2000; Franks and Mayer, 1996)[116]. In all U.S. acquisitions the gain for bidder shareholders[117] and the overall gain are indistinguishable from zero (Andrade, Mitchell, and Stafford, 2001).[118] Although suggestive, the event study evidence cannot conclusively determine whether these premia arose from the correction of an inefficiency, or from synergies between bidders and targets,[119] or whether they simply constitute transfers away from bidding shareholders or other constituencies [see Burkart (1999) for an extensive discussion of this issue].[120]

[116] Schwert (2000) reports that the total premia under the *Wall Street Journal* and TFSD definitions of "hostile deal" are 11.5% and 6.7% higher than for all deals, in line with the previous findings of Franks and Harris (1989) who report total premia of 42% for hostile and 28% for uncontested and unrevised bids in the U.S. Franks and Mayer (1996) report premia of 30% for successful hostile and 18% for accepted bids in the U.K.

[117] Most U.S. bidders are not individuals, or tightly controlled bidding vehicles, but widely held companies under management control (Shleifer and Vishny, 1988).

[118] The result holds for all subperiods 1973–1998 for cumulative abnormal returns from twenty days before the bid to the close. During the announcement period the overall gains are slightly positive (1.8%), especially for large targets (3.0%) and no-stock transactions (3.6%).

[119] See Bradley (1980), and for evidence that this was the case in the 1980s Bradley, Desai and Kim (1983, 1988).

[120] Positive takeover premia could also result from the correction of market inefficiencies caused by short-term myopia or undervalued targets. The most influential surveys of the evidence of the 1980s rejected these

7.1.3. Redistribution

How can one disentangle redistributive gains from overall efficiency improvements? A number of studies have identified and sometimes quantified the amount of redistribution away from other corporate constituencies resulting from a takeover. The constituencies in the target firm that may be on the losing side include bondholders (Higgins and Schall, 1975; Kim and McConnell, 1977; Asquith and Kim, 1982; Warga and Welch, 1993), employees (Shleifer and Summers, 1988; Williamson, 1988; Schnitzer, 1995) and corporate pension plans (Pontiff, Shleifer, and Weisbach, 1990; Petersen, 1992). But there may also be outside losers like the bidding shareholders and unprotected debtholders as well as the tax authorities.

An alternative strategy attempts to pinpoint the sources of efficiency gains through clinical studies, but no general pattern has emerged from a wealth of facts (Kaplan, 2000). The source of gain for target shareholders, when overall gains are small or non-existent, has not been identified yet with precision.

7.1.4. Takeover defences[121]

As we have seen there are theoretical arguments for and against takeover defences. They reduce the disciplining role of hostile takeovers by reducing the average number of bids but they can also help the board extract higher premia from bidders. A large empirical literature has tried to estimate the (relative) size of these effects in the U.S. Before turning to this evidence, we review the availability, mechanics and incidence of different defence mechanisms.

Numerous pre-bid and post-bid defences are at the disposal of target companies in most jurisdictions. Pre-bid defences include capital structure, classified boards, supermajority requirements, cross-shareholdings, enhanced voting rights, voting right restrictions, subjection of share transfers to board approval and change of control clauses in major contracts.[122] The most potent pre-bid defences require shareholder approval. However, some important defences which can be introduced without shareholder approval include control clauses and cross-shareholdings in Europe, poison pills in the U.S.[123] and, until recently, block acquisitions larger than 10% in Korea (Black, Kraakman, and Tarassova, 2000; Chung and Kim, 1999). The incidence of anti-takeover provisions is well documented in the U.S. (Danielson and Karpoff, 1998;

explanations on the grounds that there is evidence that stock markets are efficient and that the stock price of targets that defeat a hostile bid often returns to close to the pre-bid level (Jensen and Ruback, 1983; Jarrell, Brickley, and Netter, 1988).

[121] For a recent, critical survey of takeover defences see Coates (2000).

[122] The list of possible post-bid defences is much longer and includes litigation, white knights, greenmail and the pac-man defence.

[123] European Counsel M&A Handbook 2000, pp. 26–43. See Weston, Siu, and Johnson (2001) for a detailed explanation of U.S. anti-takeover measures.

Rosenbaum, 2000) but less systematically in Europe and Asia.[124] In the U.S., firms protected by poison pills have relatively high institutional ownership, fewer blockholders and low managerial ownership, consistent with the view that institutional ownership presents a threat in a hostile takeover situation and that blockholders can prevent the adoption of poison pills (Danielson and Karpoff, 1998).

The evidence on the consequences of takeover defence adoption is mixed. Mikkelson and Partch (1997) show that CEOs are more likely to be replaced when hostile takeover activity is high, which is consistent with disciplining and entrenchment, i.e. when CEOs are able to protect themselves better they are less likely to be replaced. The wealth effects of pre-bid defence adoption has been measured in numerous event studies that generally find small negative abnormal returns. On balance, the results support the view that managerial entrenchment dominates the enhanced bargaining effect. However, contradictory evidence comes from Comment and Schwert (1995) who find that anti-takeover measures have increased bid premia, supporting the view that the enhanced bargaining effect dominates. Here the board literature provides an intriguing piece of evidence. Shareholders of target firms with independent boards (see board section) receive premia that are 23% higher than for targets with more captive boards (Cotter, Shivdasani, and Zenner, 1997), even when controlling for the presence of anti-takeover devices. This suggests that independent boards are more ready to use anti-takeover devices to the advantage of target shareholders than other boards.

The latest panel data evidence suggests that anti-takeover provisions in the United States have had a negative impact on firm value (Gompers, Ishii, and Metrick, 2003). The same study finds that from 1990 to 1998 investors who would have taken long positions in companies with "strong shareholder protections" (as measured by an index they construct) and short positions in companies with "weak shareholder protections" would have earned abnormal returns of 8.5% per year.[125] As striking as these numbers are, however, the authors acknowledge that it is not possible to interpret this finding as measuring the market value of "good governance". The difficulty is that such abnormal returns can represent at best unanticipated benefits from good governance and may reflect changes in the business environment not directly related to governance.

[124] Danielson and Karpoff (1998) provide a detailed analysis of the adoption of anti-takeover measure in a sample that roughly corresponds to the S&P 500 during 1984–1989. Some form of anti-takeover measure covers most of their sample firms and the median firm is protected by six measures. In Europe the most potent defence against a hostile takeover is a blockholder holding more than 50% of the voting rights; in continental Europe most companies with small (or no) blocks have statutory pre-bid defences similar to U.S. companies, for example voting right and transfer restrictions or special shares with the sole right to nominate directors for election to the board (Becht and Mayer, 2001); see large investor section.

[125] Using data on 24 different "corporate governance provisions" from the IRRC (2000a) (the data we report in Table 3) the authors compare the returns on two portfolios and relate the provisions to Tobin's Q.

Table 3
Corporate takeover defences in the U.S.

	Fall 1999	Fall 1997	Mid-1995	Mid-1993	Mid-1990
Number of Companies	*1900*	*1922*	*1500*	*1483*	*1487*
	%	%	%	%	%
External Control Provisions					
Blank Check Preferred Stock	89.1	87.6	85.0	n/a	n/a
Poison pill	56.0	51.9	53.3	53.6	51.0
Consider Non-financial effects of merger	7.3	6.6	7.2	7.5	6.5
Internal Control Provisions					
Advance Notice Requirement	61.4	49.2	43.8	n/a	n/a
Classified Board	58.7	58.4	59.7	58.1	57.2
Limit right to call special meeting	36.7	33.6	31.1	28.6	23.9
Limit action by written consent	34.6	32.2	31.1	28.1	23.7
Fair price	24.8	26.4	32.5	33.2	31.9
Supermajority vote to approve merger	15.3	14.8	17.8	18.1	16.9
Dual class stock	11.5	10.7	8.3	8.2	7.5
Eliminate cumulative voting	8.8	8.4	10.4	10.1	8.8
Unequal voting rights	1.6	1.6	2.0	2.1	2.3
Miscellaneous Provisions					
Golden parachutes	64.9	55.8	53.3	n/a	n/a
Confidential Voting	10.2	9.2	11.7	9.4	3.2
Cumulative Voting	10.2	11.4	14.4	15.7	17.7
Antigreenmail	4.1	4.6	6.0	6.3	5.6

State Anti-Takeover Laws

	Mid-1999	
	Number	% of states
States with Anti-Takeover Laws	42	82.4
Featuring		
Control Share Acquisition Laws	27	52.9
Fair Price Laws	27	52.9
2–5 Year Freeze-Out Laws	33	64.7
Cash-Out Laws	3	5.9
Profit Recapture	2	3.9
Severance/Pay Labor Contract Provisions	5	9.8
Greenmail Restrictions	6	11.8
Compensation Restrictions	2	3.9
Poison Pill Endorsement	25	49.0
Directors' Duties	31	60.8
States with No Takeover Provisions $(8 + D.C.)$	9	17.6

Source: Rosenbaum (2000) and IRRC (2000a).
Note: classification taken from Danielson and Karpoff (1998).

7.1.5. One-share-one-vote

Deviations from one-share-one-vote are often associated with the issuance of dual class stock and have been the source of considerable controversy.[126] Shares with different voting rights often trade at different prices and the resulting premia (discounts) have been related to takeover models (see theory section) and interpreted as a measure of the value of corporate control and "private benefits" (Levy, 1983; Rydqvist, 1992; Zingales, 1995; Nicodano, 1998).

Theory predicts that dual class premia vary with the relative size of dual class issues, the inequality of voting power, the value of the assets under control, the probability of a takeover (which itself depends on the regulatory environment), and the likelihood of a small shareholder being pivotal.[127] In addition, relative prices are affected by differences in taxation, index inclusion, dividend rights and/or stock market liquidity.

Empirical estimates of voting premia range from 5.4% to 82% and, taken at face value, suggest that the value of corporate control is large in Italy and relatively small in Korea, Sweden and the U.S.[128] In practice the studies at best imperfectly control for all the factors affecting the price differential, making it an unreliable measure of "the value of corporate control". Time-series evidence also suggests that dual class premia should be interpreted with caution. While premia have been rising from 20% in mid-1998 to 54% in December 1999 in Germany (Hoffmann-Burchardi, 2000), in Finland they have dropped from 100% in the 1980s to less than 5% today. Similarly in Sweden premia have declined from 12% in the late 1980s to less than 1% today[129], and in Denmark from 30% to 2% (Bechmann and Raaballe, 2003). In Norway the differential was actually negative in 1990–1993, but has risen to 6.4% in 1997 (Odegaard, 2002). It is, of course, possible that changes in the value of control explain these changes in premia but further research is required before one can conclude with any confidence that this is the case.

[126] See Seligman (1986) for a comprehensive history of the one-share-one-vote controversy in the U.S. In early corporations statutory voting right restrictions were the norm.

[127] Takeover regulation can prevent block transfers, require the bidder to offer the same price to all voting stockholders or force the inclusion of non-voting stockholders. Company statutes can have a similar effect, for example fair-price amendments in the U.S. Nenova (2003) attempts to control for these factors across countries using quantitative measures of the legal environment, takeover regulation, takeover defences and the cost of holding a control block in a cross-section regression, treating the control variables as exogenous.

[128] Canada 8–13%, Jog and Riding (1986), Robinson, Rumsey, and White (1996), Smith and Amoako-Adu (1995); France, mean 1986–1996 51.4%, Muus (1998); Germany, mean 1988–1997 26.3%, in 2000 50%, Hoffmann-Burchardi (1999, 2000); Israel, 45.5%, Levy (1983); Italy 82%, Zingales (1994); Korea 10%, Chung and Kim (1999); Norway, −3.2–6.4%, Odegaard (2002), Sweden 12%, Rydqvist (1996); Switzerland 18%, Kunz and Angel (1996); U.K. 13.3%, Megginson (1990); U.S., 5.4% Lease, McConnell, and Mikkelson (1983), mean 1984–1990 10.5%, median 3% Zingales (1995); see also DeAngelo and DeAngelo (1985) for the U.S. Lease, McConnell, and Mikkelson (1984) analyse the value of control in closely held corporations with dual class shares.

[129] Personal communication from Kristian Rydqvist.

7.1.6. Hostile stakes and block sales

Takeover bids for widely held companies are, of course, not the only way corporate control can be contested and sold. In blockholder systems, hostility can take the form of "hostile stakes" (Jenkinson and Ljungqvist, 2001) and control is completely or partially transferred through block sales (Holderness and Sheehan, 1988 for the U.S.; Nicodano and Sembenelli, 2004 for Italy; Böhmer, 2000 for Germany; Dyck and Zingales, 2004 for 412 control transactions in 39 countries).[130] Control premia vary between −4% and 65% (Dyck and Zingales, 2004).[131]

7.1.7. Conclusion and unresolved issues

Hostile takeovers are associated with large premia for target shareholders, but so far the empirical literature has not fully identified the source of the premia. It is difficult to disentangle the opposing entrenchment and bargaining effects associated with hostile takeover defences. The net effect of the adoption of takeover defences on target stock market value is slightly negative, suggesting that the entrenchment effect is somewhat larger than the bargaining effect.[132] Recent evidence from the board literature suggests that independent boards implement defences to increase the bargaining position of target shareholders while captured boards tend to implement defences that increase entrenchment (Cotter, Shivdasani, and Zenner, 1997).

Despite the widespread interest in hostile takeovers, the available empirical evidence is surprisingly sketchy. Although hostile takeovers are no longer confined to the U.S. and the U.K., there appears to be no recent study of hostile takeovers in other countries.

7.2. Large investors

Shareholder rights can differ significantly across OECD countries and even across firms within the same country. These institutional differences make it difficult to compare the actions and effects of large shareholders across countries or firms.

Most of the time large shareholder action is channelled through the board of directors. Large shareholders are in principle able to appoint board members representing their interests. When they have majority control of the board they can hire (or fire)

[130] Like dual-class premia, block premia can be interpreted as an indirect measure of "private benefits". However, block premia have the advantage that they are based on actual control transactions, not the marginal value of a vote in a potential transaction.

[131] In countries with a mandatory bid rule control transfers must be partial. A control block cannot be sold without making an offer to the minority shareholders. In such countries only block sales below the mandatory bid threshold are considered. This imposes serious limits on the comparability of the results across countries.

[132] This is corroborated by comparisons of announcement effects of anti-takeover amendments with a larger bargaining component relative to devices where entrenchment is likely to be prominent., e.g. Jarrell and Poulsen (1987).

management. Large shareholders can also exercise power by blocking ratification of unfavourable decisions, or possibly by initiating decisions.

In practice corporate law, corporate charters and securities regulations impose limits on these powers, which vary significantly across countries. Even a basic right like corporate voting and appointments to the board varies considerably across governance systems and corporate charters. For example, some countries' corporate law prescribes discrete control thresholds that give a blocking minority veto power over major decisions.[133] In Germany employees appoint 50% of the board members in large corporations (Prigge, 1998). In the U.K. the listing requirements of the London Stock Exchange require large shareholders to keep an arm's length relationship with companies, limiting the right of blockholders to appoint directors to the board.[134] Under the Dutch "structural regime" the corporate boards of larger companies must appoint themselves and their successors, with a consequent negative impact on corporate valuations (De Jong et al., 2005). In some Anglo-Dutch corporations special classes of shares have the sole right to nominate directors for election to the boards or to veto their removal (Becht and Mayer, 2001).

Initiation rights also vary considerably across jurisdictions. Thus, to remove a director, shareholders might have to show "cause", wait for three years, vote separately by share-class, pass a supermajority resolution or simply pass an ordinary resolution by majority vote.[135] In the U.S. shareholders cannot initiate fundamental transactions like mergers, and boards are broadly shielded from direct shareholder influence (Hansmann et al., 2004). In contrast, shareholder proposals can force mergers or charter amendments if they receive a majority in the U.K., Japan or France.[136] Ratification rights, on the other hand, are strikingly similar in most jurisdictions. The law prescribes a list of decisions that require shareholder approval, which can be extended in the charter.

Most empirical work on large investors has focused on simple hypotheses which are not always grounded in rigorous theoretical analysis. Much of the early work on large shareholders has been concerned with the implications of the trend towards shareholder dispersion and the effects of the decline of shareholder influence. We begin this section by tracing the available evidence on ownership and control patterns across countries and through time. We then address the empirical evidence on the causes and effects of

[133] For example corporate law in the Netherlands, Germany and Austria prescribes supermajorities for major decisions. Often the threshold can be increased via the statutes, but not decreased.

[134] A 30% + blockholder cannot appoint more than 5 out of 12 directors (Wymeersch, 2003). In the U.K. the distribution of blockholdings in listed companies tapers off abruptly at 30% (Goergen and Renneboog, 2001).

[135] Initiation rights differ across the U.S., depending on the state and, within any one state, the company bylaws (Clark, 1986, p. 105). Initiation rights are always strong in the U.K., where directors can be removed at any time by an ordinary resolution brought by a 20% + blockholder or coalition and a majority vote (Section 303 of the Companies Act 1985). The same is true in Belgium, where Article 518 of the company law explicitly states that the board cannot resist such a shareholder resolution. Obviously removal rights are closely related to the anti-takeover devices we discussed previously.

[136] In some unlisted companies shareholders exert direct control of the company through voting, for example in Germany and France (Hansmann and Kraakman, 2001).

ownership dispersion. In particular, we shall address the following questions: Does the presence of large investors or "relationship investing" improve corporate performance? Do large shareholders abuse their voting power? Do alternative forms of shareholder intervention (activism) improve company performance? Is there an empirical link between share blocks and stock market liquidity?

7.2.1. Ownership dispersion and voting control

As we pointed out in the theory section, with the exception of the U.S. some form of concentration of ownership and/or voting control is the most common corporate governance arrangement in OECD and developing countries.[137] The full impact and scope of this observation has only emerged very recently after a long period of confusion originally caused by Berle and Means (1932) with their assertions and empirical methodology.

The hypothesis that risk diversification leads to growing shareholder dispersion was first tested in 1924 by Warshow (1924). His study records an astonishing 250% increase in the number of shareholders between 1900 and 1923.[138] The test of the consequences for voting control followed. Means (1930) proposed that the new owners of the "modern corporation" no longer appointed the majority of directors on the board and, therefore, no longer controlled it. For 44% of the largest 200 U.S. corporations in 1929 no large investors were found, leading to the conclusion that "control is maintained in large measure separate from ownership" (Means, 1931b; Berle and Means, 1932).[139] This hypothesis has become received wisdom for corporations in the U.S. (Larner, 1966, 1970[140]; Herman, 1981; La Porta, Lopez-de-Silanes, and Shleifer, 1999), but also for the U.K. (Florence, 1947, 1953, 1961; Cubbin and Leech, 1983; Leech and Leahy, 1991; La Porta, Lopez-de-Silanes, and Shleifer, 1999), although other studies found that blockholders had never disappeared entirely in the U.S. (Temporary National Economic Committee, 1940[141]; Eisenberg, 1976;

[137] For supporting evidence see La Porta, Lopez-de-Silanes, and Shleifer (1999), Claessens, Djankov, and Lang (2000) and Faccio and Lang (2002) and voting block statistics based on modern disclosure standards (ECGN, 1997; Barca and Becht, 2001).

[138] Warshow (1924) could not determine the exact number of shareholders because they were masked by custodians (nominee accounts, banks) or, in modern parlance, "street names". There are no comparative early studies for other countries because his method relied on the existence of registered shares and in many countries corporations have always issued bearer shares. Warshow's study was updated by Means (1930) and additional evidence is reported in Berle and Means (1932). See Temporary National Economic Committee (1940, p. 198) for a survey of these and other classic studies using the Warshow method.

[139] A corporation was classified as management controlled if it had no known shareholder holding at least 5% of voting stock. Cases falling between 5 and 20% were classified as jointly management and minority controlled and "½ a company" was assigned to each category. Berle and Means (1932) used the same definition.

[140] Larner (1966) reduced the "management control" threshold to 10% and found that the fraction of management controlled firms had increased from 44% to 84.5%. Eisenberg (1976) argues that Larner's study was biased towards finding "management control".

[141] The Temporary National Economic Committee (1940) relied on the SEC to collect this data for the 200 non-financial corporations in 1937.

Demsetz and Lehn, 1985; Holderness and Sheehan, 1988) and the U.K.[142] Later research confirms that blocks are indeed rare in the U.S. (Edwards and Hubbard, 2000; Becht, 2001), but this finding has recently been challenged by Holderness (2006) who uses hand collected data to show that U.S. ownership is more concentrated than is widely believed. In the U.K. there is no doubt that a coalition of the largest 1-5 blockholders—usually institutional investors—can wield a substantial amount of voting power in most listed companies (Goergen and Renneboog, 2001).[143]

Means's method (see footnote 138) for measuring shareholder concentration has been criticised and extended by numerous authors, for example by Gordon (1945)[144], Florence (1947)[145] and Eisenberg (1976). One particular source of measurement error is due to disclosure rules[146]. Depending on how disclosed holdings are treated one can obtain significantly different measures of concentration. Thus, La Porta, Lopez-de-Silanes, and Shleifer (1999) and Claessens, Djankov, and Lang (2000)—using the Means method—find very little ownership concentration in Japan. However, adding the ten largest holders on record in Japan in 1997 gives a concentration ratio, defined as the percentage of shares held by these shareholders, of 48.5% (51.1% in 1975; Hoshi and Kashyap, 2001, p. 252). Inevitably, much research has been undertaken on the U.S. and the U.K. because the information about shareholdings in these countries is relatively easy to obtain. In contrast, in countries where corporations issue bearer shares infor-

[142] Florence (1961) reported that the median holding of the largest 20 holders in large U.K. companies fell from 35% in 1936 to 22% in 1951, a finding that was widely cited by Marris (1964) and other British managerial economists. However, Chandler (1976) argues that personal capitalism lasted longer than these numbers suggest and that British firms only adopted managerial capitalism in the 1970s. Consistent with Chandler's view is Hannah's (1974, 1976) observation that it was possible for bidders to bypass family controlled boards only as late as the 1950s. See Cheffins (2002) for a survey.

[143] Goergen and Renneboog (2003) explore the determinants of post-IPO diffusion rates in the U.K. and Germany.

[144] Gordon (1945) argued that we should "speak [..] of the separation of ownership and active leadership. Ordinarily the problem is stated in terms of the divorce between ownership and "control". This last word is badly overused, and it needs to be precisely defined [..]. Our procedure [..] will be to study the ownership of officers and directors and then to ascertain the extent to which non-management stockholdings are sufficiently concentrated to permit through ownership the wielding of considerable power and influence (control?) over management by an individual, group or another corporation" Gordon (1945, p. 24, footnote 20).

[145] Florence (1947) proposed a measure of "oligarchic" minority control based on the full distribution of the largest 20 blocks and actual board representation.

[146] Statistics based on shareholder lists underestimate concentration unless the cash-flow and voting rights that are ultimately held by the same person or entity are consolidated. At the first level, it has been common practice to add the holdings using surnames, addresses and other obvious linkages; see for example Leech and Leahy (1991, p. 1421). First level blocks held through intermediate companies are consolidated by tracing control (or ownership) chains and adding those that are ultimately controlled by the same entity. Means (1930) applied a discrete variant of this method and classified a closely held corporation controlled by a widely held corporation as widely held.

mation about shareholdings is generally not available.[147] Fortunately for researchers, modern securities regulation has begun to overcome this problem, at least in Europe.[148]

From a theoretical point of view static measures of concentration are not always satisfactory. What matters is not whether ownership and/or voting power are more or less concentrated on a permanent basis but the ability of shareholders to intervene and exercise control over management when required (see Manne, 1965; and Bolton and von Thadden, 1998b). If there is a well functioning market for corporate control (takeovers or proxy fights) managerial discretion is limited even when companies are widely held. On the other hand, when anti-takeover rules and amendments are in place shareholder intervention is severely limited, whether a large investor is present or not. In the Netherlands, relatively few corporations are widely held, yet the ability of shareholders to intervene is very limited.[149] Dynamic measures of concentration based on power indices can address some of these issues[150] but they have been considered in only a few studies (Leech, 1987b, 1987c;[151] Holderness and Sheehan, 1988; and Nicodano and Sembenelli, 2004).[152]

7.2.2. Ownership, voting control and corporate performance

We distinguish four generations of empirical studies that have tested the proposition that there is a link between ownership dispersion, voting control and corporate performance (value).

The first generation has tested the hypothesis that free-riding among dispersed shareholders leads to inferior company performance. Starting with Monsen, Chiu, and Cooley (1968) and Kamerschen (1968) numerous authors have regressed performance measures like profit rates and returns on assets on a Means-Larner type or Gordon type corporate control dummy.[153] In most regressions the dummy was not significant and the authors have rejected the hypothesis that greater dispersion results in lower performance (see the surveys by Short, 1994 and Gugler, 2001).

[147] Obviously, when companies issue bearer shares there is no shareholder list.

[148] In the U.S. voting blocks are disclosed under Section 13 of the 1934 Act that was introduced with the Williams Act in the 1960s. The standard provides for the disclosure of ultimate voting power of individual investors or groups, irrespective of the "distance" to the company, the control device used or the amount of cash-flow rights owned. A similar standard exists in the European Union (Directive 88/627/EEC). It is also spreading to Eastern Europe via the Union's accession process.

[149] Under the structural regime corporate boards operate like the board of the Catholic Church and its chairman: the bishops appoint the Pope and the Pope the bishops; Means (1930) illustration of what he meant by management control.

[150] They do not take into account statutory anti-takeover devices.

[151] Leech (1987a) proposed a set of power indices that are related to the size and distribution of blocks for a given probability of winning a board election and applied it to Berle and Means original data (Leech, 1987b), the TNEC data (Leech, 1987c) and 470 U.K. listed companies between 1983–1985 (Leech and Leahy, 1991).

[152] The exception is the "value of corporate votes" literature that uses Shapley values and other power indices to measure the value of corporate control, for example Zingales (1995).

[153] See footnotes 134 and 138 above.

The method was also applied in other countries, finding the owner-controlled firms significantly outperform manager-controlled firms in the U.K. (Radice, 1971; Steer and Cable, 1978; Cosh and Hughes, 1989; Leech and Leahy, 1991),[154] profitability is higher with family control in France (Jacquemin and de Ghellinck, 1980).[155]

Demsetz and Lehn (1985) explain that ownership concentration is endogenous. Some firms require large shareholder control while others don't. They argue that without accounting for this endogeneity it is to be expected that a regression of firm performance on a control dummy in a cross-section of heterogeneous firms should produce no statistically significant relation if the observed ownership-performance combinations are efficient.

Following Stulz (1988) a second generation of studies focuses on inside ownership by managers and considers the effects of takeover threats. The hypothesis is a hump-shaped relationship between concentrated ownership and market capitalization.[156] Outside ownership merely shifts the locus. Morck, Shleifer, and Vishny (1988) find some evidence of such a relationship. Similarly, McConnell and Servaes (1990) find a maximum at 40–50% insider ownership (controlling for ownership by institutional investors and blockholders). Short and Keasey (1999) find similar results for the U.K.[157]

The third generation continues to test the Stulz hypothesis but vastly improves the econometrics, showing reverse causation.[158] Using instrumental variable and panel techniques the studies find corporate performance causing managerial ownership (Kole, 1995; Cho, 1998), both determined by similar variables (Himmelberg, Hubbard, and Palia, 1999), or no relationship between ownership and performance (Demsetz and Villalonga, 2001). The impact of corporate performance on managerial ownership is not significant. An alternative approach looks for instruments in institutions where ownership concentration is not endogenous, for example in co-operatives with many members. However, these studies are likely to suffer from other biases, in particular sample selection (by definition) and missing variables.[159]

[154] Holl (1975) found no significant difference between owner and manager controlled firms.

[155] See Gugler (2001) for further details.

[156] Corporate value first increases as more concentrated insider ownership aligns incentives, but eventually decreases as the probability of hostile takeovers declines.

[157] They find a maximum at 15.6% insider ownership and a minimum at 41.9%.

[158] Typical econometric shortcomings of 1st and 2nd generation ownership-performance studies are reverse causality (endogeneity), sample selection, missing variables and measurement in variables. For example, Anderson and Lee (1997) show that many 2nd generation studies used data from unreliable commercial sources and correcting for these measurement errors can flip the results. See Börsch-Supan and Köke (2002) for a survey of econometric issues.

[159] Gorton and Schmid (1999) study Austrian cooperative banks where equity is only exchangeable with the bank itself and one member has one vote, hence the separation of ownership and control is proportional to the number of members. They find that the log ratio of the average wages paid by banks, relative to the reservation wage is positively related to the (log) of the number of co-operative members, controlling for other bank characteristics, period and regional effects. They conclude that agency costs, as measured by efficiency wages, are increasing in the degree of separation between ownership and control.

The fourth generation returns to the first generation specification and econometrics, but adds two missing variables, the legal system and voting rights held in excess of cash-flow rights.[160] They find no effects for European countries (Faccio and Lang, 2002) and a negative effect of large investors in Asia (Claessens, Djankov, and Lang, 2000).[161] La Porta, Lopez-de-Silanes, and Shleifer (1999) run a Q-regression for 27 countries but neither the cash-flow rights of controlling blockholders nor the legal system have a significant effect on corporate valuation.[162] It seems inevitable that a fifth generation study will emerge that addresses the econometric problems of the fourth generation.

7.2.3. Share blocks and stock market liquidity

The empirical link between secondary market liquidity and shareholder dispersion is well documented. Starting with Demsetz's (1968) classic study, measures of liquidity such as trading volume and bid-ask spreads have been shown to depend on the number of shareholders, even when controlling for other factors (Demsetz, 1968; Tinic, 1972; Benston and Hagerman, 1974). Equally, increases in the number of shareholders, for example after stock splits (Mukherji, Kim, and Walker, 1997) or decreases in the minimum trading unit (Amihud, Mendelson, and Uno, 1999) lead to higher secondary market liquidity. The inverse relationship also holds. An increase in ownership concentration, or a decrease in the "free float", depresses liquidity (Becht, 1999 for Belgium and Germany; Sarin, Shastri, and Shastri, 1999 for the U.S.).

The positive effect of stock market liquidity is also well documented. More liquid stocks command a price premium and offer a concomitantly lower risk adjusted return, reducing the cost of capital for the company (Stoll and Whaley, 1983; Amihud and Mendelson, 1986). Hence, companies have a measurable incentive to increase the number of shareholders, providing further evidence on the existence of a monitoring-liquidity tradeoff.

To our knowledge the role of liquidity in spurring monitoring has not been explored empirically. Instead the literature has focused on asymmetric information problems and informed investors as a source of illiquidity. Empirically, higher insider ownership reduces liquidity because it increases the probability of trading with an insider (Sarin, Shastri, and Shastri, 1999; Heflin and Shaw, 2000).

[160] However, the hypothesis is reversed. The authors do not expect to find that firms without a block perform worse than firms with a block, but expropriation of minority shareholders by the blockholders.

[161] The studies regress "excess-value" (the natural logarithm of the ratio of a firm's actual and its imputed value, as defined by Berger and Ofek, 1995) on Means-Larner control dummies and other control variables.

[162] La Porta, Lopez-de-Silanes, and Shleifer (1999) perform a number of bivariate comparisons of Means-Larner control groups for a larger set of variables.

7.2.4. Banks[163]

Traditionally the empirical corporate governance literature has taken a narrow view of delegated monitoring by banks and sought to measure bank involvement through the intensity of bank-industry links such as equity holdings, cross-holdings and/or (blank) proxies, board representation and interlocking directorates.[164]

Within this narrow view there is an empirical consensus that bank-industry ties in the U.S. were strong at the beginning of the century but became weak through anti-trust regulation and the Glass-Steagall act[165], were never strong in the U.K. but always strong in Germany[166] and Japan (Hoshi and Kashyap, 2001). A popular explanation for these patterns has been the different regulatory history in these countries (Roe, 1994).[167]

The empirical literature has documented that equity holdings by banks are not very common[168], but the presence of bankers on boards and their involvement in interlocking directorates is common.[169] Based on these empirical measures the literature has compared the performance of companies under "bank influence" to other companies, with mixed results.[170] Also, the influence of banks has been identified as an important

[163] For a more general review of banks and financial intermediation see Gorton and Winton (2003).

[164] This approach has a long tradition, for example Jeidels (1905) for Germany and the Pujo Committee (1913) for the U.S.

[165] See, for example, Carosso (1970, 1973, 1985), Chernow (1990), Tallman (1991), Tabarrok (1998), Calomiris (2000), Ramirez and DeLong (2001). The relative performance of J.P. Morgan controlled and other corporations has been investigated by DeLong (1991) and Ramirez (1995). Kroszner and Rajan (1997) investigate the impact of commercial banks on corporate performance before Glass-Steagall Kroszner and Rajan (1994) and Ramirez (1995) the impact of the Act itself.

[166] Edwards and Fischer (1994), Edwards and Ogilvie (1996) and Guinnane (2002) argue that bank influence and involvement in Germany is, and has been, very limited.

[167] The regulatory explanation of (low) bank involvement in industry is convincing for the U.S., but less so for other countries. In the U.K. no restrictions apply and banks have always kept an arm's length relationship to industry. In Japan the Allied occupation forces sought to impose Glass-Steagall type restrictions, yet the *keiretsu* found other ways of maintaining strong ties.

[168] In Germany banks hold many but not the largest blocks (Becht and Böhmer, 2003). However, they exert considerable voting power through blank proxies for absent blockholders (Baums and Fraune, 1995). There is also indirect evidence that banks' holdings of equity in non-financial firms were small at the end of the 19th century (Fohlin, 1997).

[169] Interlocking directorates started to become common in Germany towards the end of the 19th century (Fohlin, 1999b). At the beginning of the 1990s only 12.8% of companies were not connected to another by some personal link and 71% had a supervisory board interlock (Pfannschmidt, 1993; see Prigge, 1998, p. 959 for further references). Most of the links were created by representatives of banks and insurance companies (Pfannschmidt, 1993). The same was true for about half of the companies in Japan, also when the bank has extended a loan to the company (Kroszner and Strahan, 2001). In the U.S. 31.6% of the Forbes 500 companies in 1992 had a banker on board, but only 5.8% of the main bank lenders had board seats. Lenders are discouraged from appointing directors because of concerns about conflicts of interest and liability during financial distress (Kroszner and Strahan, 2001). Banks also drive board seat accumulation and overlap in Switzerland (Loderer and Peyer, 2002).

[170] For surveys of this evidence see Prigge (1998:1020) for Germany, Gugler (2001) and Section 7.2 for a review of the econometric problems. In addition to the usual endogeneity problems blocks held by banks can

driver of economic growth and for overcoming economic backwardness (Tilly, 1989; Gerschenkron, 1962; Schumpeter, 1934, 1939)[171], a view that has been challenged recently.[172]

Relationship banking[173] is a broader concept that emphasises the special nature of the business relationship between banks and industrial clients. Relationship banking, broadly defined is "the connection between a bank and customer that goes beyond the execution of simple, anonymous, financial transactions" (Ongena and Smith, 1998)[174]. The ability of banks to collect information about customers and their role in renegotiating loans gives them a role in corporate governance even if they hold no equity and have no board links.

The empirical literature documents that banking relations last from 7 to 30 years on average[175], depending on the country and sample.[176] Relationships last longer when they are exclusive (Ongena and Smith, 2000), depending on interest rates and the range of services provided by the bank to the firm (DeGryse and Van Cayseele, 2000). Most firms have multiple banking relationships.[177]

Event study evidence suggests that changes in banking relationships have an impact on stock prices. The announcement of a bank loan agreement (new or renewal) is associated with positive abnormal returns, while private placements or public issues have no or a negative effect (James, 1987), a finding that has been consistently confirmed for renewals (Lummer and McConnell, 1989; Best and Zhang, 1993;

arise from debt-to-equity conversion. The classic study for Germany is Cable (1985), the most recent study Gorton and Schmid (2000).

171 Banks collected capital, lent it to able entrepreneurs, advised and monitored them, helping their companies along "from the cradle to the grave" (Jeidels, 1905).

172 Within the traditional view Fohlin (1999a) shows that the contribution of Italian and German banks to mobilising capital was limited. Da Rin and Hellmann (2002) argue that banks helped to overcome coordination failures and played the role of "catalysts" in industrial development.

173 For a recent survey with emphasis on the empirical literature see Ongena and Smith (1998), with emphasis on the theoretical literature Boot (2000).

174 "Relationship banking" might involve board and equity links, but not necessarily. The labels "*Hausbank* system" for Germany and "Main Bank System" for Japan (Allen and Gale, 2000) are often associated with exclusive debt links cemented by equity control rights, but exclusive bank-firm relationships are also found in countries where banks hold little or no industrial equity, for example the U.S.

175 At the beginning of the 1990s the average relationship in Italy lasted 14 years (Angelini, Di Salvo, and Ferri, 1998), 22 in Germany (Elsas and Krahnen, 1998), 30 years in Japan (Horiuchi, Packer, and Fukuda, 1988), 15–21 years in Norway (Ongena and Smith, 1998), but only 7.8 years in Belgium (DeGryse and Van Cayseele, 1998) and 7 years in the U.S. (Cole, 1998). In a German sample that is more comparable to the U.S. samples the mean duration is only 12 years (Harhoff and Korting, 1998); see Ongena and Smith (2000), Table 2 for further references.

176 The cross-country and cross-study comparison must be treated with some caution because the studies suffer from the usual econometric problems that are typical for duration analysis to different degrees: right and left-censoring, stock sampling and other sampling biases.

177 For large firms, the median number of bank relationships is 13.9–16.4 in Italy, 6–8 in Germany, 7.7 in Japan and 5.2 in the U.S.; see Ongena and Smith (2000), Table 3 for further details and references.

Billett, Flannery, and Garfinkel, 1995).[178] The stock price reaction to loan commitments is also positive, in particular with usage fees (Shockley and Thakor, 1997). Acquisitions financed by bank loans are associated with positive bidder announcement returns, in particular when information asymmetries are important (Bharadwaj and Shivdasani, 2003). Equally, Kang, Shivdasani, and Yamada (2000) show that Japanese acquirers linked to banks make more valuable acquisitions than acquirers with more autonomous management.

7.3. Minority shareholder action

7.3.1. Proxy fights

Corporate voting and proxy fights received considerable attention in the early theoretical literature, drawing on the analogy between political and corporate voting (Manne, 1965). In the U.S. today, proxy fights are potentially very important because they allow dissident shareholders to remove corporate boards protected by a poison pill (see Section 5.1). Proxy fights are however not very common; occurring on average 17 times a year in the period 1979–1994, with 37 contests in 1989, at the peak of the hostile takeover boom (Mulherin and Poulsen, 1998, p. 287).[179] This timing is no coincidence; 43% of these proxy fights were accompanied by a hostile takeover bid (Mulherin and Poulsen, 1998, p. 289).[180] Proxy fights are usually brought by minority shareholders with substantial holdings (median stake 9.1%).[181] In other countries with dispersed shareholdings (see Section 7.2.1), such as the U.K., proxy fights are very rare.[182] The latest evidence suggests that proxy fights provide a degree of managerial disciplining and enhance shareholder value. Gains in shareholder wealth are associated with contest related acquisitions and restructuring under new management (Mulherin and Poulsen, 1998).[183]

[178] The evidence is mixed for new loans; see Ongena and Smith (2000), Table 1.

[179] Mulherin and Poulsen (1998) is the most complete study of proxy contests in the United States to date. Previous studies for smaller samples and/or shorter time periods include Dodd and Warner (1983), Pound (1988), DeAngelo and DeAngelo (1989), Borstadt and Zwirlein (1992) and Ikenberry and Lakonishok (1993). An interesting case study is Van-Nuys (1993).

[180] In the full sample 23% of the firms involved in contest were acquired.

[181] Furthermore, most proxy contests (68%) aim to appoint the majority of directors, just more than half are successful (52%), and most result in management turnover (61%); Mulherin and Poulsen (1998:289).

[182] There are notable exceptions, for example the small shareholder action at Rio Tinto PLC (in the United Kingdom) and Rio Tinto Ltd (in Australia) in May 2000 (http://www.rio-tinto-shareholders.com/).

[183] Mulherin and Poulsen (1998) sought to resolve the inconclusive findings of previous research. In agreement with theory, event studies had shown that proxy fights occur at underperforming firms and that they increase shareholder wealth when the contest is announced and over the full contest period. However, some studies found that targets did not underperform prior to the contests, and that shareholder wealth declines after the announcement, in particular after the contest has been resolved—and relatively more when the challenger is successful in placing directors on the board of the target (Ikenberry and Lakonishok, 1993).

7.3.2. Shareholder activism

After the decline in hostile takeovers in the U.S. at the beginning of the 1990s, shareholder activism has been identified as a promising new avenue for overcoming the problems of dispersed holdings and a lack of major shareholders (Black, 1992).[184] Typical forms of activism are shareholder proposals, "focus lists" of poor performers, letter writing and other types of private negotiations. Typical activist issues are calls for board reforms (see board section), the adoption of confidential voting and limits on excessive executive compensation (see compensation section). There is anecdotal evidence that activism is also on the rise in other countries, focusing on similar issues.[185]

In the United States, the filing of ordinary shareholder proposals[186] is much easier than a full proxy solicitation but these proposals are not binding for the board or management, making such proposals the preferred tool of U.S. activists. In Europe most countries allow shareholders to file proposals that are put to a vote at shareholder meetings (Baums, 1998; Deutsche Schutzvereinigung für Wertpapierbesitz, 2000).

The empirical literature on shareholder activism in the U.S. is surprisingly large and there are no less than four literature surveys (Black, 1998; Gillan and Starks, 1998; Karpoff, 2001; and Romano, 2001). They concur that shareholder activism, irrespective of form or aim, has a negligible impact on corporate performance. However, authors disagree on the cause and interpretation of this result.

Black (1998) concludes that institutional investors spend "a trivial amount of money" on overt activism and that their ability to conduct proxy fights and appoint directors is hindered by regulation[187] and other factors.[188] In contrast, Romano (2001) argues that shareholder activism in the U.S. has a limited impact because it focuses mainly on issues that are known to matter very little for company performance and value. Fund managers and/or trustees engage in this type of activism because they derive private benefits from it, such as promoting a political career.

The two explanations are, in fact, linked. Pension funds are subject to the same agency problems as corporations and pension fund regulation is concerned with minimising

[184] As we reported in the facts section, this development is closely related to the size of pension funds in the U.S., the largest in the OECD.

[185] Shareholder activism is the logical next step from the adoption of corporate governance codes and principles, pressing companies to implement the recommendations put forward in these documents (see http://www.ecgi.org for a listing and full-text copies of corporate governance codes).

[186] In the U.S. shareholder proposals are filed under Rule 14a-8 of the SEC's proxy rule. They are precatory in nature, i.e. even if a majority of the shares outstanding vote in favour of the proposal the board is not oblidged to implement the resolution.

[187] Initially (Black, 1992) argued that shareholder activism could overcome (regulation induced) shareholder passivity in the U.S.

[188] In the U.K. there are fewer regulatory barriers than in the U.S., but there are other reasons why institutional investors are reluctant to exercise voice, for example "imperfect information, limited institutional capabilities, substantial coordination costs, the misaligned incentives of money managers, a preference for liquidity, and uncertain benefits of intervention" (Black and Coffee, 1994).

investment and management risk for beneficiaries. Institutional activism pushes the corporate governance problem to a higher level, with even higher dispersion this time of policy holders (often with no voting right or "one-holder-one-vote" rules), no market for pension fund control and boards with poorly paid and/or trained trustees.[189] In the U.S., trustees of 401(k) plans are appointed by the corporation, raising conflict of interest issues laid bare in the recent collapse of Enron.[190]

7.3.3. Shareholder suits

Shareholder suits can complement corporate voting and potentially provide a substitute for other governance mechanisms. Once again the institutional details differ across countries.[191] In the U.S. shareholder litigation can take the form of derivative suits, where at least one shareholder brings the suit on behalf of the corporation, and direct litigation, which can be individual or class-action.[192] The incidence of shareholder suits in the U.S. is low. Between 1960–1987 a random sample of NYSE firms received a suit once every 42 years and including the OTC market, 29% of the sample firms attracted about half of the suits (Romano, 1991).[193] In Europe enforcing basic shareholder rights usually falls upon public prosecutors but direct shareholder litigation is also possible on some matters.

Three main hypotheses have been tested: who benefits more from shareholder suits, shareholders or lawyers; is there any evidence that managers are disciplined by shareholder litigation; and does shareholder litigation boost or replace other forms of monitoring?

The most comprehensive empirical study for the U.S. covers the period 1960–1987 (Romano, 1991)[194]. She finds that shareholders do not gain much from litigation, but their lawyers do. Most suits settle out of court, only half of them entail a recovery for

[189] See Myners (2001) for a recent policy report on pension fund management and governance in the U.K. His survey of U.K. pension fund trustees revealed that they received one day of training prior to taking up their job. Leech (2003) analyses the incentives for activism in the United Kingdom. Stapledon (1996) compares institutional shareholder involvement in Australia and the United Kingdom.

[190] Conflicts of interest and outright looting of pension fund assets were at the bottom of the collapse of the Maxwell media empire in the U.K. in 1992; Bower (1995) and Greenslade (1992).

[191] In most countries shareholders can appeal to the courts to uphold their basic rights, for example their voting and cash-flow rights. However, the extent and incidence of shareholder litigation differs substantially. Here we only deal with suits brought against managers or directors.

[192] The details of procedure and financial incentive differ for the two types of action (Clark, 1986). For derivative suits the recovery usually goes to the corporation, but it must reimburse a plaintiff's legal expenses, reducing the problem of shareholders at large free-riding on the shareholders bringing the suit. In practice lawyers have an incentive to seek out shareholders and offer to bear the cost if the suit is unsuccessful and take a large fee if it is successful. This provides lawyers with an incentive to settle for a low recovery fee and a high lawyer's fee (Klein and Coffee, 2000, p. 196).

[193] For more recent descriptive statistics on class action see Bajaj, Mazumdar, and Sarin (2001).

[194] Unfortunately the study has not been updated (Romano, personal communication).

shareholders and when they do the amount recovered per share is small.[195] In contrast, in 90% of the settled suits the lawyers are awarded a fee. There are some structural settlements but they are mostly cosmetic. The market is indifferent to the filing of a derivative suit but exhibits a negative abnormal return of −3.2% for class action.[196] There is little evidence that managers are disciplined by litigation. Executive turnover in sued firms is slightly higher, but managers almost never face financial losses.[197] Suits both help and hinder other types of monitoring. For example, blockholders are likely to get sued[198] but they also use the threat of a suit to force change or reinforce their voting power. There seems to be no comparable empirical evidence for other countries.

7.4. Boards[199]

7.4.1. Institutional differences

In practice the structure, composition and exact role of boards varies greatly between individual corporations (charters) and governance systems. The same is true for the rules governing the appointment and removal of a board member and their duties.[200] In formal terms, boards can have one or two tiers. One-tier boards are usually composed of executive directors and non-executive directors. In theory the executives manage and the non-executives monitor, but in practice one-tier boards are often close to management.[201] In a two tier board system there is a separate management board that is overseen by a supervisory board. Supervisory board members are barred from performing management functions.[202] Informally, both types of board can be more or less "captured" by management or dominated by blockholders.[203] To avoid the problem of capture by such interests, corporate governance recommendations emphasise the role of

[195] The recovery in derivative suits is only half as large as in direct (class) action.

[196] This could be related to the fact that the recovery in derivative suits is only half as large as in direct (class) action and that the class action recovery goes to shareholders, not the company itself. Indeed, the latter might be selling shareholders, i.e. no longer hold any shares in the company (Romano, 1991, p. 67).

[197] Compensation packages are unchanged and settlement fees are met by special insurance policies taken out by the company.

[198] As we pointed out elsewhere this is consistent with the view that shareholder suits limit self-dealing, but also with the view that they generally discourage block holding (Black, 1990).

[199] Recent surveys on the role of boards include Romano (1996), Bhagat and Black (1999) and Hermalin and Weisbach (2003).

[200] Despite these differences, the OECD Principles (1999) contain a long list of board responsibilities and prescribes basic elements of board structure and working required to fulfil its objectives.

[201] For example, it is (or used to be) common that the chairman of the board and the chief executive officer are the same person and in some countries they must be by law.

[202] Most countries have either one or the other system, but in France companies can choose.

[203] For example, it is common that the supervisory board is staffed with former members of the executive board, friends of the CEO or the blockholder.

"independent directors", non-executive directors who have no links with the company other than their directorship and no links with management or blockholders.[204]

The role of the board in approving corporate decisions also varies. In one system a decision that can be ratified by the board requires shareholder approval in another. Major decisions, like mergers and acquisitions, almost always require shareholder approval. In most systems the shareholders appoint and remove the board, but the rules vary substantially (see the large investor section). The board appoints the managers. In some countries boards have a formal duty vis-à-vis the employees of the company or, as in Germany, employees have the right to appoint directors. In the U.S. statutes that require boards to take into account the interests of non-shareholder constituencies are commonly portrayed as "anti-takeover rules" (Romano, 1993).[205]

7.4.2. Board independence

There are few formal models of boards (see theory section) and the empirical work has focused on loose hypotheses based on policy or practical insights and recommendations. The bulk of this work has investigated whether board composition and/or independence are related to corporate performance and typically rejects the existence of such a relationship.

To measure the degree of board independence, several criteria have been proposed.[206] Is the chief executive officer the chairman of the board? What is the proportion of independent directors on the board? Are there any board committees and how are they staffed? Coded into variables, the answers are related to performance measures like abnormal returns, Tobin's Q and/or the usual accounting measures with simple regression analysis. The evidence from the U.S. suggests that board composition and corporate performance are "not related" (Hermalin and Weisbach, 2003), the relationship is "uncertain" (Bhagat and Black, 1999) or is "at best ambiguous" (Romano, 1996).

7.4.3. Board composition

Most of these studies are subject to the econometric criticisms we highlighted in the large investor section. In the model of Hermalin and Weisbach (1998) board composition is endogenous and what we observe in a cross-section might be efficient. Hence, we would not expect to see a significant relationship between board structure and general performance. Does board composition affect performance or do the needs of companies

[204] Not surprisingly the exact definition of "independent" also varies a great deal and is the subject of constant debate. See the ECGN codes page (www.ecgn.org) for full text copies of such recommendations and definitions.

[205] See Hansmann et al. (2004) for a comprehensive discussion of the role of boards in a comparative perspective.

[206] Motivated by casual observation some studies have also investigated whether board size is related to performance.

affect their board composition? The empirical analysis of boards is also in need of third generation studies.

Warther's (1998) model predicts that boards only play a role in crisis situations and there is some evidence that this is true for independent boards. In the takeover context bidder shareholders protected by outsider dominated boards suffer less from overbidding (get smaller negative abnormal returns) than when boards are management-dominated (Byrd and Hickman, 1992). Also, outside boards are more likely to remove CEOs as a result of poor company performance (Byrd and Hickman, 1992).

7.4.4. Working of boards

Recommendations of "best practice" (e.g. European Association of Securities Dealers, 2000) advance the practical hypothesis that the working as well as the composition of boards matters for performance. This proposition has been tested indirectly since it is virtually impossible to devise a quantitative measure of the way a board is run on the inside. [207] Hence a practitioner's interpretation of the results of this empirical literature might be that the studies have simply failed to measure the dimension of boards that matters most for corporate performance—their functioning.

7.4.5. International evidence

The international evidence on the role of boards in corporate governance and their impact on corporate performance is sketchy or the relevant studies are not easily accessible. A notable exception is the U.K. where a number of studies have broadly confirmed the findings for the U.S. (Franks, Mayer, and Rennebook, 2001).

7.5. Executive compensation and careers[208]

7.5.1. Background and descriptive statistics

Executive compensation in the U.S. has risen continuously since 1970 (see Murphy, 1999) and in 2000 reached an all-time high, with the bulk of the increase stemming from option plans.[209] Compensation consultants estimate that for a comparable U.S.

[207] Vafeas (1999) finds a positive relationship between the frequency of board meetings and corporate performance, but obviously this too is a very crude measure of the effectiveness of the working of the board. In a study that has been very influential in the management literature, Lorsch and MacIver (1989) use the survey method to provide direct evidence on the working of boards. Adams (2003) uses board remuneration as a proxy for board effort, but does not control for endogeneity.

[208] For recent surveys see Bebchuk, Fried, and Walker (2002), Gugler (2001, p. 42), Perry and Zenner (2000), Loewenstein (2000), Abowd and Kaplan (1999) and Murphy (1999). Core, Guay, and Larcker (2003) survey the specialized literature on equity based compensation and incentives.

[209] Total compensation for the average U.S. CEOs increased from $1,770,000 in 1993 to $3,747,000 in 1997 (in 1992 CPI-deflated dollars). The value of options in this package rose from $615,000 to $1,914,000 and bonuses from $332,000 to $623,000; Perry and Zenner (2001, p. 461), Table 1.

CEO the basic compensation package alone is higher than total package in Germany, Spain, Sweden and Switzerland, and not much lower than in France or Japan.[210] In contrast, the total compensation of other management is similar across OECD countries and higher in Italy than in the U.S. (Abowd and Kaplan, 1999). The differential remains large when data are adjusted for company size.[211]

Executive contracts are supposed to provide explicit and implicit incentives that align the interests of managers with those of shareholders, as discussed in the theory section. The bulk of the empirical literature has focused on sensitivity of pay[212] (explicit incentives) and the dismissal of executives (implicit incentive) to corporate performance.[213] High levels of pay were justified with the extraordinary gains in wealth shareholders reaped through most of the 90s and incentive pay was characterised as one of the drivers behind the high market valuation of U.S. corporations (Holmstrom and Kaplan, 2001). Recently, while stock prices plummeted and executive pay did not, attention has shifted to asymmetries in the pay-performance relationship and the potential for self-dealing by CEOs.

7.5.2. Pay-performance sensitivity

In the early 1990s the consensus view in the literature was that the sensitivity of pay to performance in the U.S. was too low (see Baker and Hall, 2004; Jensen and Murphy, 1990).[214] Executives did not receive enough cash after good corporate performance and did not incur sufficient losses, through dismissal, after poor performance. The same conclusions were reached for other countries, most notably Japan (see Kaplan, 1994a). In the U.S. the sensitivity of executive pay to performance reached levels 2 to 10 times higher than in 1980 by 1994 (see Hall and Liebman, 1998). The dollar change in executive wealth normalised by the dollar change in firm value appears small and falls by a factor of ten with firm size, but the change in the value of the CEO's equity stake is

[210] The value of an executive compensation package is typically measured by the "after-tax value of salaries, short-term bonuses, deferred retirement bonuses, stockholdings, stock bonuses, stock options, dividend units, phantom shares, pension benefits, savings plan contributions, long term performance plans, and any other special items (such as a loan to the executive made at a below market rate)" (Antle and Smith, 1985). As we shall see, the most important and controversial item are stock options, an unprecedented rise in their use throughout the 90s and the terms on which they are granted.

[211] Cheffins (2003) explores whether there will be global convergence to U.S. pay levels and practices: how can U.S. pay levels remain so much higher than anywhere else, and why has this gap only opened up in the last decade and not earlier.

[212] See Rosen (1992) for an early survey of this literature.

[213] The accounting literature also emphasizes the technical problem of estimating the monetary value of top executive compensation packages. See Antle and Smith (1985), based on early work by Burgess (1963) and Lewellen (1968).

[214] The point was also emphasized in an early survey by Jensen and Zimmerman (1985).

large and increases with firm size.[215] The probability of dismissal remained unchanged between 1970 and 1995 (Murphy, 1999).[216]

The sensitivity of equity-based compensation with respect to firm value is about 53 times higher than that of the salary and bonus components (Hall and Liebman, 1998). However, even for median performance the annualised percentage increase in mean wealth for CEOs has been 11.5% for the period between 1982 and 94 (Hall and Liebman, 1998) and the size of CEO losses relative to the average appreciation of their stock holdings has been modest.

In other countries too, the use of equity-based compensation and pay-performance sensitivity has risen, but nowhere close to the U.S. level. In the U.K. the percentage of companies with an option plan has risen from 10% in 1979 to over 90% in 1985 (Main, 1999). However, the level of shareholdings and pay-performance sensitivity are about six times lower than in the U.S. (Conyon and Murphy, 2000).

7.5.3. Are compensation packages well-designed?

Agency theory predicts that incentive pay should be tied to performance relative to comparable firms, not absolute performance. And indeed, early studies found that changes in CEO cash compensation were negatively related to industry and market performance, but positively related to firm performance (Gibbons and Murphy, 1990)[217]. In contrast, equity-based compensation is hardly ever corrected for industry or market stock index movements, leading to a solid rejection of the relative performance evaluation (RPE) hypothesis in all recent surveys (Core, Guay, and Larcker, 2003, pp. 38–39; Bebchuk, Fried, and Walker, 2002; Abowd and Kaplan, 1999; Murphy, 1999).[218]

Agency theory can be used to determine the optimal exercise price of options when they are granted. The optimal price is a function of numerous factors and not the same for different firms. In practice most options are granted at the money (i.e. with an exercise price equal to the company's stock price on the day), a clear contradiction of the predictions of theory (Bebchuk, Fried, and Walker, 2002, p. 818).

[215] Baker and Hall (2004) document the firms size effect and discuss the merits of each measure. During 1974–1986 the median CEO gained or lost $3.25 for $1000 gained or lost by shareholders, adjusted for the risk of dismissal; but money equivalent of this threat was only $0.30 (Jensen and Murphy, 1990). In 1997 and 1998 the gain or loss was $10–11 per $1000 (unadjusted) (Perry and Zenner, 2000; Hall and Liebman, 2000). For an executive holding stock and options worth $20,000,000, a 10% change in stock prices implies a $2,000,000 change in wealth.

[216] Among S&P 500 firms average CEO turnover rates for low performers were 15% on and 11% from the 25th performance percentile upwards (Murphy, 1999).

[217] See Murphy (1999, p. 2535) for additional references.

[218] Several explanations of this puzzle have been put forward including accounting problems, tax considerations, the difficulty in obtaining performance data from rivals, worries about collusion between companies, the ability of managers to get back to absolute performance plans with appropriate financial instruments, but not a single one is very satisfactory.

Theory also predicts that incentive schemes and the adoption of such schemes should result in net increases in shareholder wealth. The latest evidence (based on "abnormal Q" regressions) rejects this prediction. An increase in CEO option holdings leads to a decrease in Tobin's Q, suggesting that CEOs hold too many options but not enough stock (Habib and Ljungqvist, 2002). However, event study evidence generally supports the theory (Morgan and Poulsen, 2001; DeFusco, Johnson, and Zorn, 1990; Brickley, Bhagat, and Lease, 1985; Larcker, 1983).[219]

Agency theory further predicts that incentive pay and blockholder monitoring or takeover threats are substitutes. Firms subject to blockholder monitoring or with family representatives on the board are less likely to implement stock option plans (Mehran, 1995; Kole, 1997) because more discipline substitutes for more sensitivity of pay. In contrast, without blockholder monitoring, CEOs are not paid as the theory predicts (Bertrand and Mullainathan, 2001, 2000). Boards protected by state anti-takeover laws (Bertrand and Mullainathan, 1998) or anti-takeover amendments (Borokhovich, Brunarski, and Parrino, 1997) (see takeover section) provide more incentive pay to compensate for less discipline from hostile takeovers, while in the U.K. takeover threats are higher while incentive pay and the level of pay are lower than in the U.S. (Conyon and Murphy, 2000). However, there are inconsistencies. Companies in industries with more disciplining takeovers should pay less, while in fact they pay more (Agrawal and Walkling, 1994; Agrawal and Knoeber, 1998). Although these results are suggestive, self-dealing is a plausible rival explanation—boards that are monitored less give more pay to their CEO cronies.[220]

7.5.4. Are managers paying themselves too much?

Few direct tests of the rival "self-serving manager" explanation of U.S. pay practices are available, but some studies attempt to get at the issue indirectly. Thus, there is evidence that management manipulates the timing of stock option grants (Yermack, 1997) and times the flow of good and bad news prior to the option grant (Aboody and Kasznik, 2000). This can be interpreted as evidence of self-dealing (Shleifer and Vishny, 1997a).

Another way of determining whether there has been self-dealing is to see whether CEO stock option plans (or bonus packages) have been approved by a shareholder vote. Even though in 2000 almost 99% of the plans proposed at major U.S. corporations received shareholder approval, the average percentage of votes cast against stock-option plans has increased from under 4% in 1988 to about 18% in 1995–1999 (IRRC, 2000b), 20.2% in 1999 and 23.3% in 2001 (IRRC, 2002). In some cases dilution levels are 70%

[219] Note that DeFusco, Johnson, and Zorn (1990) found a negative reaction in bond prices, interpreting the adoption of stock option plans as means for transferring wealth from bondholders to stockholders. An influential early study is Masson (1971).

[220] Bebchuk, Fried, and Walker (2002) express general skepticism about the substitution effect between incentive pay and disciplining through takeovers. They argue that boards can pay themselves and the CEO large amounts of money without reducing the value of the company enough to justify a takeover.

or more, especially in the technology sector, often associated with "evergreen" features (IRRC, 2002). There is rising concern about exemptions for "broadly based plans"[221], potential dilution of voting rights[222], broker voting[223], option repricing, payments in restricted stock, loans for share purchases, "evergreen plans"[224] and discount options (Thomas and Martin, 2000). In addition, activists are now worried that "at the same time that stock prices are falling, CEO pay continues to rise" (American Federation of Labor and Industrial Organizations, 2001).[225] These results are not strong direct evidence support for the self-serving manager hypothesis, but they can be re-interpreted as yet another failure of shareholder monitoring in the U.S.

In parallel with the takeover literature, yet another approach for distinguishing between self-serving and efficient behaviour brings in board composition and the power of the CEO vis-à-vis the board. Outside and independent directors on the board or on remuneration committees are thought to be (more) resistant to awarding self-serving compensation packages. In contrast, CEOs who are also the chairman of the board ("duality") are thought to lean more towards self-dealing. In the U.S., most corporations have a compensation committee comprising outside directors.[226] As a direct result of the Cadbury (1992) and Greenbury (1995) reports, U.K. issuers have remuneration committees[227] and in 1994 already they were 91% staffed with outside directors. Similarly, during 1991–1994 the proportion of U.K. boards with "duality" fell from 52% to 36% (Conyon and Peck, 1998). Both developments are also gaining ground in continental Europe.[228] So far, empirical studies have failed to detect that institutions and reforms have any impact on pay structure. In the U.S. committees staffed with directors close to management do not grant unusually generous compensation packages (Daily et al., 1998). In the U.K. in 1991–1994, the proportion of non-executive directors serving on boards

[221] Stock option plans that do not need shareholder approval if they benefit more than a certain proportion of non-officer employees.

[222] The IRRC (2001) estimates that the average potential dilution of the voting power of the currently outstanding shares from stock option plans was 13.1% for the S&P 500 and 14.6% for the S&P 1500 in 2000, higher than in previous years.

[223] Under NYSE rules brokers can vote shares without instructions from the beneficial owners. A recent study estimates that routine proposals that benefit from broker votes receive 14.2% more "yes" votes than other routine proposals of the same kind, making broker votes marginal for 5.2% of routine proposals (Bethel and Gillan, 2002).

[224] Evergreen plans reserve a small percentage of stock for award each year. Once approved the awards are made without shareholder approval. "Quasi-evergreen plans" have a limited lifetime, regular plans run indefinitely (Thomas and Martin, 2000, p. 62).

[225] The AFL-CIO has recently opened a Website campaigning against "runaway pay" in the U.S. see (http://www.paywatch.org).

[226] If not, under U.S. tax law compensation is not tax deductible for executives mentioned in the proxy statement (Murphy, 1999).

[227] See Conyon and Mallin (1997).

[228] See http://www.cgcodes.org for reports on the implementation of the pertinent governance recommendations in continental Europe.

and duality had no effect on compensation structure (Conyon and Peck, 1998).[229] CEOs monitored by a board with interlocking directors get more pay (Hallock, 1997).[230]

There is evidence that the extensive use of compensation experts and peer review increases pay in excess of what is warranted from a pure agency perspective. For example, CEOs with pay packages that lie below the median of their peers see their pay increase more quickly, *ceteris paribus* (Bizjak, Lemmon, and Neveen, 2000).

7.5.5. Implicit incentives

Implicit incentives typically take the form of executive dismissal or post-retirement board services. Post-retirement appointment to a board can be a powerful implicit incentive or, once again, a sign of self-dealing. In the U.S., CEO careers continue after retirement with 75% holding at least one directorship after two years. Almost half (49.5%) stay on their own board after retirement, in 18% of the cases as chairman (Brickley, Linck, and Coles, 1999).[231]

Most explicit and implicit incentives are written into CEO contracts that, under U.S. Federal Law, must be disclosed but had not been collected until recently (Minow, 2000). Preliminary analysis reveals that contracts range from "short and to the point" (Minow, 2000) to guaranteed benefits and perks of epic proportions.[232] Implicit benefits include severance pay for dismissal without "cause"[233] or in case of changes in control (acquisition of 15, 20 or 51% of the voting shares).[234] We expect that more analytic studies based on this data will shed more light on these issues.

7.5.6. Conclusion

To conclude, it has become difficult to maintain the view, based on data from the bull market of the early 90s, that U.S. pay practices provide explicit and implicit incentives for aligning the interests of managers with those of shareholders. Instead, the rival view that U.S. managers have the ability, the opportunity and the power to set their own pay at the expense of shareholders (Bebchuk, Fried, and Walker, 2002), increasingly prevails. We know relatively less about pay practices in other countries, but attempts to

[229] We are not aware of a direct test that exploits the time series variation of the U.K. reforms.

[230] Fich and White (2005) investigate the determinants of interlocks.

[231] Many corporate governance codes oppose the appointment of CEOs to their own boards after retirement.

[232] See http://www.thecorporatelibrary.com/ceos/. One of the more lavish contracts included a $10 million signing bonus, $2 million stock options at $10 a share below market, a "guaranteed bonus" of at least half a million dollars a year, a Mercedes for the executive and his wife, a corporate jet for commuting and first class air for the family once a month, including the executive's mother (Minow, 2000).

[233] The definition of cause is often stringent, for example "felony, fraud, embezzlement, gross negligence, or moral turpitude" (Minow, 2000).

[234] The latter, once again, weakens the potential monitoring role of blockholders in the U.S.

implement U.S. practices are controversial, as the long-standing debate in the U.K.[235] and recent rows in France[236] show. The institutional investor community is drawing its own conclusions and has tabled global guidelines on executive pay[237], while corporate America is under pressure to report earnings net of the cost of stock options.

7.6. Multiple constituencies

In addition to shareholders there are four major other constituencies: creditors (and other non-equity investors), employees, suppliers and clients. In parallel to the theory section we focus on the role and impact of the debtholder and employee constituencies in a comparative corporate governance perspective.

7.6.1. Debtholders

Many aspects of the role of debtholders in corporate governance are addressed in the empirical financial contracting literature.[238] These studies investigate the evolution impact and choice of general capital structures, or the effect of changes in leverage on stock prices, particularly in the context of corporate control transactions (see takeovers section).

The main theoretical rationale for sharing control between managers, shareholder *and* debtholders is their different role in restructuring and, in particular during financial distress (see theory section).

Is debt a commitment device for liquidation after poor performance? As usual, the role of debtholders differs appreciably between countries. For example, in the U.S. insolvency law is "softer" than in the U.K.[239], and judges are more lenient (Franks and Sussman, 2005a). Furthermore, regulation in the U.S. is subject to political intervention and lobbying, which further weakens the usefulness of debt as a commitment device (Berglöf and Rosenthal, 1999; Franks and Sussman, 2005a; Kroszner, 1999).[240] Basic statistics lend support to this view. In the U.S. the rate of

[235] Recently coalitions of U.K. institutional investors have been successful at curbing pay packages, even in the case of perceived excess among their own kind; Andrew Bolger, *Prudential bows to revolt over executive pay*, FT.com; May 08, 2002.

[236] Pierre Tran and David Teather, *Vivendi shareholders turn on Messier*, The Guardian; April 25, 2002.

[237] The proposed standard prescribes, *inter alia*, individual disclosure for individual executives, reporting of stock options as a cost to the company, shareholder voting on pay policy, appointment of an independent pay committee and limits on potential channels of self-dealing (e.g. loans to executives); International Corporate Governance Network (2002).

[238] For a comprehensive earlier survey see Harris and Raviv (1992).

[239] Under Chapter 11 of the 1978 Bankruptcy Code the debtor is allowed to stay in control and try to raise new cash. In the U.K. floating charge holders take control through the appointment of an Administrative Receiver who acts in their interest and replaces the board (Franks and Sussman, 2005b; Davies, Prentice, and Gower, 1997).

[240] Theory predicts that ex-ante commitment from dispersed debt is stronger than concentrated debt, yet systems that give creditors strong liquidation rights often do so through an agent, making it easier to renegotiate (e.g. the U.K. and Germany).

deviation from absolute priority rules is 77–78%[241] but it is close to zero in the U.K. (Franks and Sussman, 2005b).[242]

Recent work on venture capital financing lends more direct support to the importance of debtholder involvement by analysing the actual contracts signed between firms and the providers of finance.[243] Consistent with the theory they find that the financial constituencies[244] have control and liquidation rights that are contingent on performance and that control shifts between constituencies, again depending on performance (Kaplan and Strömberg, 2003).

7.6.2. Employees

The literature on employee involvement has focused on two questions: does employee involvement come at the expense of shareholders (reduce shareholder wealth), and if contracts are incomplete, is employee involvement efficient? There is little empirical evidence in support of the first question and, to our knowledge, no empirical evidence that would allow us to formulate an answer to the second question.

The incidence of employee involvement is often thought to be limited to Germany's mandatory codetermination and two-tier boards. In fact, employee involvement is also mandatory in Austria and the Netherlands[245] (two-tier boards), Denmark, Sweden, Luxembourg and France[246] (one-tier board). Companies operating in two or more member states of the European Union must have a "European Works Council".[247] Voluntary codetermination can be found in Finland and Switzerland (Wymeersch, 1998). In contrast, employees in Japan are not formally represented on the board (Hoshi, 1998), although Japanese corporations are run, supposedly, in the employees' and not the shareholders' interest (Allen and Gale, 2000). Compared to the wealth of opinions on

[241] For example Franks and Torous (1989).

[242] Note that these basic statistics are methodologically problematic. The U.S. studies suffer from sample bias, looking primarily at large companies with publicly traded debt and conditional on the outcome of the bankruptcy procedure. Hence, the results could be distorted towards more or less actual commitment in the U.S. at large. The statistics of Franks and Sussman (2005b) do not suffer from this problem because they were sponsored by a government-working group on the reform of insolvency law.

[243] Sahlman (1990), Black and Gilson (1998), Kaplan and Strömberg (2003).

[244] In theory a venture capitalist (universal bank) holding debt and equity represents two constituencies.

[245] In the Netherlands the board members of large *structuur* regime corporations have a duty to act "in the interest of the company" and shareholders do not appoint them. Formally the incumbent board members appoint new board members. In practice they are chosen jointly by capital and labour because the shareholders and the employees can challenge appointment in a specialised Court (Wymeersch, 1998, p. 1146).

[246] The French system provides for weak representation and has been called "a mockery" (Wymeersch, 1998, p. 1149).

[247] Council established under the European Works Council Directive (94/45/EC) to ensure that all company employees are "properly informed and consulted when decisions which affect them are taken in a Member State other than that in which they are employed." The Directive applies to companies and groups with at least 1,000 employees in the European Economic Area (the EU15, Norway, Iceland and Liechtenstein) as a whole and at least 150 in each of two or more Member States.

employee involvement, the empirical literature is small, even for countries were such institutions are known to exist, such as Germany.

German codetermination provides for mandatory representation of employees on the supervisory board of corporations[248] with three levels of intensity: full parity for coal, iron and steel companies (since 1951)[249], quasi-parity for other companies with more than 2000 employees (since 1976)[250] and 1/3 parity for those with 500–2000 employees (since 1994).[251] Media companies are exempt.

Does the degree of codetermination adversely affect shareholder wealth or company performance? If codetermination reduces shareholder wealth, shareholders will resent codetermination and they will try to bypass[252] or shift board rights to the general assembly. There is some evidence of the former but none for the latter. In 1976 most supervisory boards of corporations subject to the quasi-parity regime did not have to be consulted on important management decisions[253] (Gerum, Steinmann, and Fees, 1988), a clear violation of the recommendations in most corporate governance codes (see Section 6.2).[254]

If there are losses in shareholder wealth from codetermination, how large are they? Econometric studies of codetermination compare company or sector performance "before and after" the 1951, 1952, 1972 and 1976 reforms or their enforcement by the courts. These studies find no or small effects of codetermination (Svejnar, 1981, 1982; Benelli, Loderer, and Lys, 1987; Baums and Frick, 1999) and/or their samples and methodology are controversial (Gurdon and Rai, 1990; FitzRoy and Kraft, 1993).[255] A recent study relies on the cross-section variation of codetermination intensity, controlling for different types of equity control and company size. It finds codetermination reducing market-to-book-value and return on equity (Gorton and Schmid, 2004). Codetermination intensity and its incidence correlate with other factors that are known to matter for stock price and accounting measures of performance, in particular sector and company size, and it is doubtful that one can ever fully control for these factors.

[248] See Hopt et al. (1998) and Prigge (1998) for an overview; in what follows we only discuss corporations (*AGs*). The German-language literature is vast; see Streeck and Kluge (1999) or Frick, Kluge, and Streeck (1999) for recent examples.

[249] Shareholders and workers each appoint 50% of the board members. The chairman is nominated by the board and must be ratified by the general meeting and both sides of the board by majority vote.

[250] The chairman is chosen by the shareholder representatives and has a casting vote.

[251] Between 1952–1994 this regime applied to all corporations, and still does for corporations registered before 1994.

[252] For example by delegating sensitive tasks to shareholder dominated committees or allowing the shareholder appointed Chairman to add items to the agenda at will.

[253] The catalogue of decisions is long and includes mergers and acquisitions, patents and major contracts.

[254] In coal, iron and steel companies, where codetermination is most intense, more management decisions required formal approval from the supervisory board, an apparent contradiction to the general finding. However, one can argue that worker influence is so intense in these companies that the capital side of the supervisory board is too weak to apply a *de facto* opt-out of codetermination.

[255] Frick, Kluge, and Streeck (1999), Gerum and Wagner (1998).

8. Recent developments

Since we wrote our earlier survey (Becht, Bolton, and Röell, 2003) there have been several important developments in corporate governance both on the regulatory front and in academic research. First and foremost, in response to the corporate scandals that were unfolding while we were writing our survey, "the most sweeping securities law reforms since the New Deal"[256] have been implemented in the U.S. with the passage of the Sarbanes-Oxley act in July 2002, and also the reforms brought about subsequently by New York attorney general Eliot Spitzer in his settlement with the Wall Street investment banking industry. In Europe ongoing reform efforts were accelerated by the extraterritorial reach of the U.S. reforms and Europe's own corporate governance scandals.

Second, on the scholarly research front, the same corporate scandals have renewed interest in three major issues in corporate governance: (i) conflicts of interest among auditors, financial analysts and in investment banking more generally, (ii) executive compensation and earnings manipulation and, (iii) the role of the board of directors.

Third, despite these research efforts the gaps between scholarly research and the fast moving world of corporate governance we identified in our original survey has probably widened. Practitioners and policy makers were fast off the mark in implementing reform and it will take several years for academia to digest the flurry of reform activity and other developments we have observed since the scandals broke. In particular we still know very little about: (i) the comparative merits of mandatory rules preferred by U.S. reformers versus the more market oriented reforms pursued in Europe (through voluntary codes and "comply or explain"); (ii) the growing importance of corporate governance ratings and indices; (iii) the role of hedge funds and private equity firms in European corporate governance and restructuring[257]; (iv) the advantages and disadvantages of different board election systems; (v) the mechanisms that allow an economy like China, with vaguely defined property rights and minimal shareholder protections to raise external capital and grow at astonishing rates.

We shall briefly review here the major developments in these areas, the debates they have given rise to, and also mention what in our view are the most significant advances in scholarly research in the past two years. Inevitably, given the enormous literature the corporate scandals and subsequent reforms have spawned, our brief discussion in this section could not be comprehensive. The changes that have taken place in the past three or four years have been so momentous that only a historian standing back from these events will be able to piece the whole picture together.

[256] As characterized by David Skeel (2005).

[257] A coalition of minority investors led by a London based hedge fund recently forced the resignation of the CEO of the Deutsche Börse AG. During the heyday of *Deutschland AG* corporate governance such a development would have been unthinkable.

8.1. Regulatory responses to corporate scandals

8.1.1. The Sarbanes-Oxley act

The Sarbanes-Oxley act (SOX) is a direct response to key governance failings at Enron and WorldCom. It targets primarily the kinds of abuses in earnings manipulation and financial reporting uncovered by the Enron and WorldCom failures. Its main aim is to restore confidence in company financial statements by dramatically increasing penalties for misreporting earnings performance and reducing conflicts of interest for two main groups of monitors of firms, auditors and analysts. In addition, SOX provides stronger protections for whistle-blowers.

The provision in SOX that has perhaps drawn the most attention is the stiff criminal penalties CEOs and CFOs face if they are now found to knowingly or willingly falsify financial statements. Post SOX, CEOs and CFOs must personally certify public accounts and if they are later found to have falsely reported earnings they may face steep jail sentences. What is more, to the subsequent great irritation of the management community, SOX requires CEOs to also assess and attest to internal controls (for small companies, the costs involved can be a significant deterrent to going public). To limit CEO's incentives to manipulate earnings, SOX also now requires CEOs to reimburse any contingent payments they received based on past overstated earnings. What is more, companies are now forbidden from extending loans to CEOs (repayable in company shares), thus banning a dubious practice that had taken extreme proportions in the case of Worldcom[258].

To further strengthen financial reporting SOX reduces the conflicts of interest in auditing that have arisen with the rapid growth in consulting activities by the major auditing firms. It has been argued that an important reason why Arthur Andersen has been so lax in monitoring Enron' s accounts is that by probing the firm's accounting practices too deeply it risked losing its most valuable consulting client. The SOX legislation targets this basic conflict with several new regulations. First, the auditor of a firm is strictly limited in its consulting activities for that firm. Second, the auditing firm is now selected by an audit committee entirely composed of independent directors instead of by the CFO. Third, the entire accounting profession is now regulated by a new body, the public chartered accountants oversight board, charged with monitoring the accounting firms. Fourth, to further reduce the risk of collusion between the auditor and firm, the lead accounting partner must rotate every five years[259]. Finally, SOX also requires

[258] Bernie Ebbers received loans from Worldcom worth a staggering total of $400 million.

[259] The reforms stopped short of implementing more radical proposals requiring rotation of the entire auditing firm after a fixed period of time, as in Italy (every 9 years). Interestingly, it is the implementation of this rule that prompted Parmalat to do all its accounting manipulation off-shore, where it was allowed to continue to retain its old auditor. Had Italy required this rotation of auditors for all activities, including off-shore ones chances are that the Parmalat scandal would never have happened.

greater disclosure of off-balance sheet transactions to reduce the risk of Enron-style accounting manipulation.

Another interesting provision that is aimed at reducing the risk of financial fraud is the greater protections given by SOX to whistle-blowers. Should they lose their jobs for exposing financial wrongdoing then SOX guarantees whistle-blowers' reinstatement, as well as back pay and legal fees. Unfortunately however, SOX requires that whistle-blowers file a complaint with the Occupational Safety and Health Administration (OSHA) a division of the Labor Department, which has little financial or accounting expertise and so far has dismissed most cases as frivolous complaints. Inevitably the OSHA's extreme conservatism has quickly undermined the effectiveness of this important reform[260].

There are many interesting aspects of this new securities law that merit a deeper discussion than we can provide here: the political battles surrounding the passage of the law; what its effects have been; whether it is an adequate response to the types of abuses that have been uncovered by the corporate scandals; and whether its benefits in terms of strengthening the quality of financial reporting outweigh the greater compliance and auditing costs. Several recent contributions provide such an in depth analysis, among which Ribstein (2003), Gordon (2003), Romano (2005a) and Skeel (2005).

8.1.2. Other U.S. reforms

Congress was not the only U.S. institution to pursue corporate governance reform. The New York Stock Exchange revised its listing rules and imposed de facto mandatory rules. It now requires, for example, that listed companies must have a majority of independent directors, with a tightened definition of independence. It also requires companies to have a nominating/corporate governance committee and a compensation committee composed entirely of independent directors.

The SEC also swung into motion and attempted to reform the proxy voting process, making shareholder voting more effective, in particular board elections.[261] This proposal met with considerable resistance from the corporate sector and has been defeated. The SEC's proposed reforms on board elections have also re-ignited a peripheral debate among U.S. legal scholars on the old question of the respective positions of federal regulations and state law (in particular the role of Delaware corporate law) in regulating corporate governance.[262] We discuss the core economic issues in this debate at greater length in Section 3.

[260] See Deborah Solomon and Kara Scannell, "SEC Is Urged to Enforce 'Whistle-Blower' Provision", *The Wall Street Journal*, 15 November, 2004.

[261] See Bebchuk (2003a, 2003b).

[262] Roe (2005) argues that the "federal response" (by Congress, the NYSE and the SEC) shows that there is no regulatory competition between U.S. states: Delaware has a monopoly and when Delaware law gets out of bounds, the Federal authorities step in Romano (2005b) argues that the U.S. corporate scandals cannot be attributed to shortcomings of Delaware law.

8.1.3. Eliot Spitzer and conflicts of interest on Wall Street

The corporate scandals of 2001 also led to investigations by Eliot Spitzer at the major Wall Street investment banks into possible conflicts of interest among "sell-side" analysts. It was alleged that these conflicts may have induced some leading analysts (most notoriously Henry Blodget and Jack Grubman) to produce misleading research and rosy earnings forecasts, and thereby participate in a vast peddling scheme of new equity deals underwritten by their firms. Spitzer quickly uncovered striking evidence of widespread tainted investment advice designed to support the placement of lucrative IPO's and mergers of client firms. At the same time a number of academic studies have appeared that report related evidence of, (i) investment bank-affiliated analysts providing excessively optimistic recommendations (see in particular Hong and Kubik, 2003), (ii) analysts' compensation being tied to profits generated at the underwriting arm of their firm (see, for example, Michaely and Womack, 1999) and, (iii) of small unsophisticated investors being influenced more by the recommendations of analysts that have clear potential conflicts of interest than the more seasoned institutional investors (see Malmendier and Shanthikumar, 2004).

Spitzer's investigations and law suits against the major Wall Street investment banks eventually gave rise to a major settlement in December 2002, whereby the investment banks agreed to set aside $450 million to finance independent research over a five year period and to pay fines amounting to $900 million. In addition, the settlement required stronger separation between in-house analysts and their bank's underwriting arm, as well as greater disclosure of their potential conflict of interest. Thus, for example, analysts are now prohibited from going on road shows to market new issues.

As striking as these reforms are, however, they stop way short of proposals that were hotly debated during the settlement negotiations, mainly, (i) the branding of in-house research as "sales literature" and, (ii) the establishment of completely independent research and advisory institutions to be financed collectively according to a pre-specified formula by the investment banking industry. Arguably, the reinforced "Chinese walls" that now separate analysts from their corporate finance colleagues can still be circumvented, so that the potential for a conflict of interest among sell-side analysts remains and could again give rise to rosy recommendations in the next IPO wave[263].

8.1.4. European reforms

In Europe the response to the corporate scandals has been more restrained and has relied more on self regulation, corporate governance codes and the "comply or explain" principle. Codes play a bigger role in Europe than in the rest of the world and their

[263] See Randall Smith, "Regulators set accord with Securities Firms, But Some Issues Persist", *Wall Street Journal*, 23 December 2003. See also, Bolton, Freixas, and Shapiro (2005) for an analysis of the merits and drawbacks of fully independent research and advisory institutions.

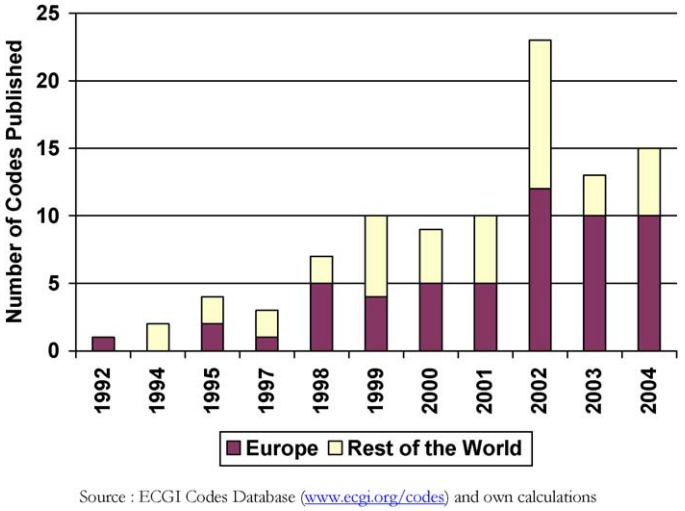


Figure 1. Number of corporate governance codes published by year.

adoption has increased substantially after the publication of the first set of OECD Principles (1999) and again after the collapse of Enron (see Figure 1). European corporate governance practice has also been affected to some extent by the extra-territorial reach of U.S. reforms and corporate efforts to harmonize standards with the United States, in particular for auditing.

8.2. Executive compensation and earnings manipulation

Following the string of corporate scandals of 2001–2002, many commentators did not fail to notice that the executives of Enron, WorldCom, and the other failed corporations had been richly compensated almost all the way up to the failure of their companies. While there had been concerns about excesses in executive compensation and the insufficient sensitivity of CEO pay to performance prior to the corporate scandals, these concerns were largely muffled by the extraordinary rise in stock prices over the 1990s. However, when the technology bubble burst, the lofty rewards CEOs had been able to secure no longer seemed justified given the companies' subsequent dismal performance[264]. How could CEOs be paid so much when their stock-performance was so poor?

[264] In the summer of 2002 *The Financial Times* published a survey of the 25 largest financially distressed firms since January 2001 and found that top executives in these firms walked away with a total of $3.3 billion in compensation. In particular, Kenneth Lay, the CEO of Enron received total compensation of $247 million, Jeffrey Skilling, the former CEO and President of Enron received $89 million and Gary Winnick, the CEO of Global Crossing received $512 million in total cumulative compensation (see *The Financial Times*, July 31, 2002).

If executive compensation and stock options could no longer be rationalized straightforwardly as incentive-efficient pay, what were the true determinants of CEO pay? This question has received a lot of attention from corporate governance scholars in recent years and a number of competing explanations have been proposed. One hypothesis put forward by Bebchuk and Fried (2004) is that CEO pay is mainly driven by CEO power to extract rents and by failures in corporate governance. They argue that the most highly compensated CEOs have essentially been able to set their own pay through captured boards and remuneration committees. However, to camouflage the extent of their rent extraction activities CEOs have cloaked their pay packages in the guise of incentive efficient pay.

An alternative line put forward by Bolton, Scheinkman, and Xiong (2003), Jensen, Murphy, and Wruck (2004) links the excesses of CEO pay to the technology bubble of the 1990 and the excess emphasis over this period on short-term stock performance. Bolton, Scheinkman, and Xiong (2003) expand the classical principal-agent framework of optimal incentive contracting to incorporate the possibility of stock price bubbles and characterize the optimal CEO compensation contract in this context. They find that when large differences of opinion among shareholders fuel a bubble, the optimal compensation contract induces a greater short-term CEO orientation and encourages actions that fuel speculation and short-term stock price performance at the expense of long-run firm fundamental value. This provides an explanation for why compensation committees and boards representing the interests of shareholders may have chosen to structure CEO pay in such a way that CEOs were able to profit early from a temporary speculative stock price surge.

Staying within the classical agency framework, Hermalin (2005) proposes yet another explanation. He points to the trend over the 1990s towards greater board independence, a higher proportion of externally recruited CEOs, a decrease in the average tenure of CEOs, and higher forced CEO turnover to suggest that these trends alone could explain why CEO pay has increased so much over this period. In a more competitive environment, with riskier and more demanding jobs, CEOs may simply have required better compensation.

Several other explanations have been proposed, too numerous to survey comprehensively in this short update. We shall only discuss briefly another important line of research linking executive compensation with accounting and stock-price manipulation. Besides the major accounting frauds uncovered in the Enron, WorldCom and more recently the AIG scandals, it has been widely documented that the technology bubble has been accompanied by a substantial growth in earnings restatements. Thus, Levitt (2002) points out that while there were only 6 restatements in 1992 and 5 in 1993, there were over 700 restatements over the period of 1997 to 2000. In addition, a number of recent empirical studies have uncovered a positive statistical relation between stock-based

compensation and earnings manipulation as measured by restatements,[265] discretionary accruals[266] and SEC accounting enforcement actions.[267]

More generally, stock-based compensation also appears to have led to other forms of corporate malfeasance beyond just earnings manipulation. Peng and Röell (2006) find that there is a direct statistical link between CEO option-based compensation and the incidence of securities class action lawsuit filings, over and above the indirect link through earnings manipulation. In an influential paper that has recently generated significant attention in the press and precipitated major regulatory investigations, Heron and Lie (2006) show that many companies' uncanny ability to time option awards at a time of fleetingly low stock prices can be attributed to the backdating of option awards: not only has the quantitative significance of the abnormal gains declined drastically after the SEC imposed much faster disclosure of grants, but the gains that still remain are concentrated among firms that miss filing deadlines.

8.3. Reforming the board of directors

The corporate scandals have also set off a raging debate on the role of the board of directors and its effectiveness in monitoring management. Many observers have pointed out that Enron had an exemplary board by the corporate governance standards of the day, with a larger than average number of independent directors and with greater incentive compensation for directors. Nevertheless, Enron's board clearly failed to protect Enron's shareholders.

At WorldCom the failures of the board were more obvious. Interestingly, in an effort to restore trust and to signal that the new company would have impeccable corporate governance standards, the bankruptcy court commissioned a study by Richard C. Breeden—former SEC chairman—to recommend new rules for the board of directors and the compensation and audit committees. As a result, part of the bankruptcy-reorganization agreement for WorldCom has been to require the new company to emerge from Chapter 11 (renamed MCI) to introduce a strengthened and more independent board as well as other corporate governance changes.

In his report, Breeden (2003) made several concrete proposals for reforming the board, which define a new benchmark for spotless corporate governance. Breeden recommends that all directors should be independent, that the chairman of the board should not be the CEO, that at least one new director be elected each year to the board, that shareholders be allowed to nominate their own candidates for election to the board (by allowing them to include their chosen candidates in the management's proxy statement), that the CEO be banned from sitting on other boards, that directors of MCI be banned from sitting on more than two other company boards, that board members be required

[265] Burns and Kedia (2005) and Richardson, Tuna, and Wu (2003).

[266] Bergstresser and Philippon (2006), Gao and Shrieves (2002), Cheng and Warfield (2005) and Peng and Röell (2006).

[267] Johnson, Ryan, and Tian (2006) and Erickson, Hanlon, and Maydew (2004).

to visit company facilities and meet with the CFO and General Counsel in the absence of the CEO, etc.

Needless to say, most publicly traded companies in the U.S. today are far from living up to this standard. Perhaps the Breeden standard is just excessive, especially if the company already has gained the trust of its shareholders. But, it is less clear whether one of Breeden's proposals initially advocated by the SEC, to allow shareholders to include their own candidates for election on the board in the management proxy statement, is excessive[268]. Some corporate governance scholars, in particular Bebchuk and Fried (2004), have strongly argued in favor of this reform. But the business community and other commentators generally perceive this to be a radical overly interventionist rule (see Symposium on Corporate Elections, Bebchuk, 2003b).

At the heart of this debate on board reform lies a fundamental unresolved economic question on the exact role of the board. Should the board of directors be seen as having only an (inevitably adversarial) monitoring role, or should directors also play an advisory role? And, even if the board's role is mainly one of oversight, will the board be able to effectively play this role if it has to rely on a CEO wary of the directors' response to disclose the relevant information about the company's operations? Beyond the role of the board there is also an unresolved question as to the exact role of the CEO. Is the CEO simply an agent for shareholders whose excesses need to be reigned in, or does he play a more important leadership role? If it is up to the CEO to determine and implement the overall strategy for the corporation then shouldn't one expect that even directors with the best intentions will defer to the CEOs judgment? All these questions have not received much attention prior to the corporate scandals and much more analysis and research is needed to be able to answer them conclusively and thus come to a determination of the appropriate policy towards boards.

8.4. Other major research themes

Besides the three issues we have touched on so far, several other themes have received a lot of attention since the publication of our survey. We briefly discuss the ones that have caught our attention in this section.

8.4.1. Corporate governance and serial acquisitions

During the 1980s and early 1990s bidder shareholders did not gain much from corporate acquisitions but, on average, bidders did not overpay either. New evidence, however, shows that during the last takeover wave this was no longer true (see in particular

[268] The SEC proposal was that instead of forcing shareholders, who want to propose a candidate for the board in opposition to the candidates nominated by management, to undertake a full-scale proxy fight, to facilitate the nomination through a two-step procedure. The first step being some event to be defined that forces the company to open the proxy to shareholder nominees, and the second step being a vote on candidates nominated by the shareholders.

Moeller, Schlingemann, and Stulz, 2005). Between 1998 and 2001 bidders incurred significant losses from acquisitions. This loss distribution is highly skewed, with only a few acquirers exhibiting very large abnormal returns in the days surrounding the announcement of the deal. The losses for acquirer shareholders were larger than the gains for target shareholders, so that on net corporate value was dissipated on a massive scale through the last merger wave. Many examples of poor acquisitions were driven by poor corporate governance at the acquiring firms. The anecdotal evidence on the major corporate governance scandals at least highlights how a corporate governance breakdown made it possible for management to engage in runaway acquisition programmes at WorldCom, Enron, Hollinger, Vivendi and Parmalat, among others.

8.4.2. Stock returns and corporate governance

As we have highlighted in our survey, the debate on how much value good governance can produce has been revived by the striking finding of Gompers, Ishii, and Metrick (2003) that from 1990 to 1998 investors long on companies with good governance and short on companies with bad governance (as measured by an index they construct) would have earned abnormal returns of 8.5% on average per year. Although the authors themselves cautioned about the interpretation of their findings many subsequent commentators were less careful and took their study to provide conclusive evidence of the link between good governance and high stock returns. As the recent study by Core, Guay, and Rusticus (2006) shows, however, the interpretation that good stock performance is driven by good governance that most commentators have adopted is problematic. In particular, they find that although governance appears indeed to be related to profit performance, there is no evidence from analysts' forecasts and earnings announcements that the stock market was in any way surprised by firms' performance. As they argue, one cannot, therefore, attribute the differences in stock returns to market surprises about earnings performance.

8.4.3. Corporate governance and ownership structure

Why is ownership of listed companies in the United Kingdom, the United States and Japan so much more dispersed than in other countries? We reviewed a broad range of hypotheses in our original survey and concluded that we could not distinguish them properly because the available ownership data was limited to recent cross-sections. Fortunately, data collection of long ownership time-series is starting to shed new light on this question. In two important recent studies Franks, Mayer, and Rossi (2003) put together an ownership time-series for the United Kingdom and establish that ownership in the United Kingdom has dispersed very quickly once a company has been taken public or following mergers and acquisitions. They find, in particular, that rapid dispersion occurred and substantial amounts of external finance were raised even in the early 19th century, at a time when corporate law gave very little protection to minority shareholders.

Other recent studies have revisited the link between ownership concentration and shareholder monitoring. Thus, Anderson and Reeb (2003) study the performance of family-controlled listed firms, which they point out represent a significant proportion of the largest listed companies even in the U.S. (18% of the S&P 500). They find that family firms consistently outperform their peers, as measured by both accounting yardsticks like return on assets and market-valuation measures such as Tobin's q. This above average performance can also be seen in the lower cost of debt financing for family-run firms (Anderson, Mansi, and Reeb, 2003). This evidence thus provides strong support for the view that ownership concentration improves governance and performance at least for family owned firms.

8.4.4. Shareholder activism and fund voting patterns

Since August 2004, a new SEC regulation requires U.S. mutual fund companies and registered investment management companies voting on behalf of investors to divulge how they have voted on proxy issues. The SEC data on fund voting patterns has recently become available, and it has been analyzed in two recent studies. Interestingly, Rothberg and Lilien (2005) have found that mutual funds almost always vote with management on operational issues and social or ethical issues, but they often vote against management on anti-takeover (34% vote against) and executive compensation (59%) issues. In addition, Stock pickers tend to vote against management less often than index funds, and in particular less often than big fund families, which abstain or vote against management 19% of the time. Davis and Kim (2005) focus more specifically on conflicts of interest arising from business ties between mutual funds and their corporate clients: many mutual fund companies derive substantial revenues from their involvement in corporate benefit plans. They find no sign that proxy voting depends on whether a firm is a client or not. However, in the aggregate, mutual fund families with heavy business ties are less likely to vote in favour of shareholder proposals opposed by management.

8.4.5. Corporate governance and the media

The watchdog role of the media is a very new area of inquiry that is starting to yield sketchy but tantalising insights.[269] Dyck and Zingales (2003a) point out that journalists, like analysts, are under pressure to accentuate the positive as a means of ensuring continued preferential access to company information sources. They measure media capture by the degree to which the presentation of material in company press releases—in particular, the emphasis on GAAP earnings versus unstandardized and possibly massaged "Street" earnings—is mirrored in press reports. They find that, in particular, non-WSJ coverage and that of less well researched firms (in terms of analyst following) is more

[269] Sherman (2002) describes the surprising blindness of the financial press to obvious red flags in Enron's publicly available financial reports in the period before the scandal broke.

likely to echo the company's "spin" in stressing "Street" earnings whenever the company's press release does so. There is an interesting cyclicality in spin. In the post-2000 downturn, even though company press releases emphasised Street earnings more, the press became more focused on GAAP; and Dyck and Zingales (2003b) also find that Harvard Business School case-writers rely more on independent sources during downturns. The authors attribute the cyclicality in spin to higher demand for news during stock market boom periods: if company news sources are in relatively fixed supply, they are able to exert more pressure on journalists during booms. This line of work is plausible but still somewhat speculative; we can expect it to be a growing area of research.

8.4.6. Corporate governance and taxes

Desai, Dyck, and Zingales (2004) point out that government tax enforcement can play a useful role in deterring false disclosure and theft by company insiders. They find that the increased vigor of tax enforcement under Putin reduced control premia in Russia, especially in the extractive industries (oil, gas and minerals) that were targeted most by the stricter enforcement policies. Paradoxically, announcements of increased tax enforcement had a positive stock price impact, especially for companies that seemed to be diverting shareholder value and avoiding taxes the most by selling oil at suspiciously low prices. Conversely, poor corporate governance is found to hinder the collection of corporate tax revenue in cross-country comparisons.

9. Conclusion

Our earlier survey concluded by attempting to take stock of the voluminous research output on corporate governance over the past two decades. There is no need to repeat the same exercise again here even if some of our assessments and conclusions might well be different in light of the important events that have unfolded over the past three years and in light of the new research we have discussed. What is certainly apparent from our brief review of the most recent developments is that research on corporate governance has continued with ever greater intensity. Remarkably, despite this voluminous outpouring of research there is still enormous interest in the field and in the issues. However, although much ground has been covered some of the long-standing deepest questions are still poorly understood, such as the role of the state in the economy, how corporate governance should be approached in emerging market countries, the link between politics, sociology and governance, and why there is such a diversity of governance arrangements around the world. In this respect, the ambitious new book by Gourevitch and Shinn (2007), which takes on some of these core issues, may well show the way to a promising new area of research.

References

Aboody, D., Kasznik, R. (2000). "CEO stock option awards and the timing of corporate voluntary disclosures". Journal of Accounting and Economics 29 (1), 73–100.

Abowd, J.M., Kaplan, D.S. (1999). "Executive compensation: six questions that need answering". Journal of Economic Perspectives 13 (4), 145–168.

Adams, R. (2001). "The dual role of boards as advisers and monitors". PhD dissertation, University of Chicago.

Adams, R.B. (2003). "What do boards do? Evidence from board committee and director compensation data". Federal Reserve Bank of New York, mimeo.

Admati, A.R., Pfleiderer, P., Zechner, J. (1994). "Large shareholder activism, risk sharing, and financial market equilibrium". Journal of Political Economy 102, 1097–1130.

Aghion, P., Hermalin, B. (1990). "Legal restrictions on private contracts can enhance efficiency". Journal of Law, Economics and Organization 6, 381–409.

Aghion, P., Tirole, J. (1997). "Formal and real authority in organizations". Journal of Political Economy 105, 1–29.

Aghion, P., Bolton, P. (1987). "Contracts as a barrier to entry". American Economic Review 77, 388–401.

Aghion, P., Bolton, P. (1992). "An incomplete contracts approach to financial contracting". Review of Economic Studies 59, 473–494.

Aghion, P., Bolton, P., Fries, S. (1999). "Optimal design of bank bailouts: the case of transition economies". Journal of Institutional and Theoretical Economics 155, 51–70.

Aghion, P., Bolton, P., Tirole, J. (2000). Exit Options in Corporate Finance: Liquidity versus Incentives. Harvard University, Cambridge, MA.

Agrawal, A., Knoeber, C.R. (1998). "Managerial compensation and the threat of takeover". Journal of Financial Economics 47 (2), 219–239.

Agrawal, A., Walkling, R.A. (1994). "Executive careers and compensation surrounding takeover bids". Journal of Finance 49 (3), 985–1014.

Allen, F., Michaely, R. (2003). "Payout policy". In: Constantinides, G., Harris, M., Stulz, R. (Eds.), Handbook of the Economics of Finance, Volume 1A Corporate Finance. North-Holland, Amsterdam, pp. 337–429.

Allen, F., Gale, D. (2000). Comparing Financial Systems. MIT Press, Cambridge and London.

American Federation of Labor and Industrial Organizations (2001). Runaway CEO Pay: What's Happening and What You Can Do About It. (http://www.aflcio.org/paywatch/.)

Amihud, Y., Mendelson, H. (1986). "Asset pricing and the bid-ask spread". Journal of Financial Economics 17, 223–249.

Amihud, Y., Mendelson, H., Uno, J. (1999). "Number of shareholders and stock prices: evidence from Japan". Journal of Finance 54, 1169–1184.

Andenas, M., Hopt, K.J., Wymeersch, E. (Eds.) (2003). Free Movement of Companies in EC Law. Oxford University Press, Oxford.

Anderson, R.C., Lee, D.S. (1997). "Ownership studies: the data source does matter". Journal of Financial and Quantitative Analysis 32, 311–330.

Anderson, R.C., Reeb, D.M. (2003). "Founding-family ownership and firm performance: evidence from the S&P 500". Journal of Finance 58, 1301–1328.

Anderson, R.C., Mansi, S.A., Reeb, D.M. (2003). "Founding family ownership and the agency cost of debt". Journal of Financial Economics 68, 263–285.

Andrade, G., Mitchell, M., Stafford, E. (2001). "New evidence and perspectives on mergers". Journal of Economic Perspectives 15, 103–120.

Angelini, P., Di Salvo, R., Ferri, G. (1998). "Availability and cost of credit for small businesses: customer relationships and credit cooperatives". Journal of Banking and Finance 22 (6–8), 925–954.

Antle, R., Smith, A. (1985). "Measuring executive compensation: methods and an application". Journal of Accounting Research 23 (1), 296–325.

Aoki, M. (1990). "Toward an economic model of the Japanese firm". Journal of Economic Literature 28, 1–27.

Aoki, M., Patrick, H., Sheard, P. (1994). "The Japanese main bank system: an introductory overview". In: Aoki, M., Patrick, H. (Eds.), The Japanese Main Bank System: Its Relevance for Developing and Transforming Economies. Oxford University Press, Oxford and New York.

Asquith, K.P., Kim, E.H. (1982). "The impact of merger bids on the participating firms' security holders". Journal of Finance 37, 1209–1228.

Avilov, G., Black, B., Carreau, D., Kozyr, O., Nestor, S., Reynolds, S. (1999). "General principles of company law for transition economies". Journal of Corporation Law 24, 190–293.

Ayres, I., Cramton, P. (1994). "Relational investing and agency theory". Cardozo Law Review 15, 1033–1066.

Ayres, I., Gertner, R. (1989). "Filling gaps in incomplete contracts: an economic theory of default rules". Yale Law Journal 99 (1), 87–130.

Bacon, J., Brown, J. (1975). Corporate Directorship Practices: Role Selection, and Legal Status of the Board. American Society of Corporate Secretaries, New York.

Bagnoli, M., Lipman, B.L. (1988). "Successful takeovers without exclusion". Review of Financial Studies 1, 89–110.

Bajaj, M., Mazumdar, S.C., Sarin, A. (2001). Securities Class Action Settlements: An Empirical Analysis. Haas School of Business, University of California-Berkeley, mimeo. (http://ssrn.com/abstract=258027.)

Baker, G.P., Hall, B.J. (2004). "CEO incentives and firm size". Journal of Labor Economics 22 (4), 767–798.

Barca, F., Becht, M. (Eds.) (2001). The Control of Corporate Europe, The Control of Corporate Europe. Oxford University Press, Oxford.

Baron, D.P. (1983). "Tender offers and management resistance". Journal of Finance 38, 331–343.

Bartlett, J. (1994). Venture Capital: Law, Business Strategies, and Investment Planning. John Wiley, New York.

Baums, T. (1998). "Shareholder representation and proxy voting in the European Union: a comparative study". In: Hopt, K.J., Kanda, H., Roe, M.J., Wymeersch, E., Prigge, S. (Eds.), Comparative Corporate Governance. The State of the Art and Emerging Research. Clarendon Press, Oxford, pp. 545–564.

Baums, T., Frick, B. (1999). "The market value of the codetermined firm". In: Blair, M.M., Roe, M.J. (Eds.), Employees and Corporate Governance. Brookings Institution Press, Washington, pp. 206–235.

Baums, T., Fraune, C. (1995). "Institutionelle Anleger und Publikumsgesellschaft: Eine Empirische Untersuchung". Die Aktiengesellschaft 3, 97–112.

Bebchuk, L.A. (1992). "Federalism and the corporation: the desirable limits on state competition in corporate law". Harvard Law Review 105, 1435.

Bebchuk, L.A. (1999). A Rent-Protection Theory of Corporate Ownership and Control. NBER, Cambridge, MA.

Bebchuk, L.A. (2000). "Using options to divide value in corporate bankruptcy". European Economic Review 44, 829–843.

Bebchuk, L.A. (2003a). "The case for shareholder access to the ballot". Business Lawyer 59, 43–66.

Bebchuk, L.A. (2003b). Symposium on Corporate Elections. Harvard Law and Economics Discussion Paper No. 448. (http://ssrn.com/abstract=471640.)

Bebchuk, L.A., Ferrell, A. (1999). "Federalism and corporate law: the race to protect managers from takeovers". Columbia Law Review 99, 1168–1199.

Bebchuk, L.A., Ferrell, A. (2001). "A new approach to takeover law and regulatory competition". Virginia Law Review 87, 1168.

Bebchuk, L.A., Fried, J.M. (2004). Pay without Performance. Harvard University Press, Cambridge, MA.

Bebchuk, L.A., Hart, O. (2001). "Takeover bids vs. proxy fights in contests for corporate control". NBER Working Paper Series #8633.

Bebchuk, L.A., Roe, M. (1999). "A theory of path dependence in corporate ownership and governance". Stanford Law Review 52, 127–170.

Bebchuk, L.A., Coates, J., Subramanian, G. (2002). "The powerful antitakeover force of staggered boards: theory, evidence & policy". Stanford Law Review 54, 887–951.

Bebchuk, L.A., Kraakman, R., Trianis, G. (2000). "Stock pyramids, cross-ownership, and dual class equity". In: Morck, R.K. (Ed.), Concentrated Corporate Ownership. University of Chicago Press, NBER Conference Report Series, Chicago and London.

Bebchuk, L.A., Fried, J.M., Walker, D.I. (2002). "Managerial power and rent extraction in the design of executive compensation". The University of Chicago Law Review 69, 751–846.
Bechmann, K., Raaballe, J. (2003). "A regulation of bids for dual class shares. Implication: two shares—one price". European Journal of Law and Economics 15 (1), 17–46.
Becht, M. (1999). "European corporate governance: trading off liquidity against control". European Economic Review 43, 1071–1083.
Becht, M. (2001). "Beneficial ownership in the U.S." In: Barca, F., Becht, M. (Eds.), The Control of Corporate Europe. Oxford University Press, Oxford, pp. 285–299.
Becht, M., Mayer, C. (2001). "Introduction". In: Becht, M. (Ed.), The Control of Corporate Europe. Oxford University Press, Oxford, pp. 1–45.
Becht, M., Böhmer, E. (2003). "Voting control in German corporations". International Review of Law and Economics 23, 1–29.
Becht, M., Bolton, P., Röell, A. (2003). "Corporate governance and control". In: Constantinides, G., Harris, M., Stulz, R. (Eds.), Handbook of Economics and Finance, Part I. North Holland, Amsterdam, pp. 1–109.
Benelli, G., Loderer, C., Lys, T. (1987). "Labor participation in corporate policy-making decisions: West Germany's experience with codetermination". Journal of Business 60, 553–575.
Benston, G.J., Hagerman, R.L. (1974). "Determinants of bid-asked spreads in the over-the-counter market". Journal of Financial Economics 1, 353–364.
Berger, P.G., Ofek, E. (1995). "Diversification's effect on firm value". Journal of Financial Economics 37, 39–65.
Berglöf, E. (1990). "Corporate control and capital structure—essays on property rights and financial contracts". A Dissertation for the Doctor's Degree in Business Administration. Stockholm School of Economics, Stockholm.
Berglöf, E. (1994). "A control theory of venture capital finance". Journal of Law, Economics and Organization 10, 247–267.
Berglöf, E., von Thadden, E.-L. (1994). "Short-term versus long-term interests: capital structure with multiple investors". Quarterly Journal of Economics 109, 1055–1084.
Berglöf, E., Rosenthal, H. (1999). The Political Economy of American Bankruptcy: The Evidence from Roll Call Voting, 1800–1978. Princeton University, mimeo.
Bergstresser, D., Philippon, T. (2006). "CEO incentives and earnings management". Journal of Financial Economics 80 (3), 511–529.
Bergstrom, C., Hogfeldt, P., Molin, J. (1997). "The optimality of the mandatory bid rule". Journal of Law, Economics and Organization 13, 433–451.
Berkovitch, E., Israel, R. (1996). "The design of internal control and capital structure". Review of Financial Studies 9, 209–240.
Berle, A.A. (1931). "Corporate powers as powers in trust". Harvard Law Review 44, 1049.
Berle, A.A. (1932). "For whom corporate managers are trustees: a note". Harvard Law Review 45, 1365.
Berle, A.A., Means, G.C. (1930). "Corporations and the public investor". The American Economic Review 20, 54–71.
Berle, A.A., Means, G. (1932). The Modern Corporation and Private Property. The Macmillan Company, New York.
Berle, A.A. (1954). The 20th Century Capitalist Revolution. Harcourt Brace, New York.
Bernheim, B.D., Whinston, M. (1985). "Common marketing agency as a device for facilitating collusion". RAND Journal of Economics 16, 269–281.
Bernheim, B.D., Whinston, M. (1986a). "Menu auctions, resource allocation, and economic influence". Quarterly Journal of Economics 101, 1–31.
Bernheim, B.D., Whinston, M. (1986b). "Common agency". Econometrica 54 (4), 923–942.
Bertrand, M., Mullainathan, S. (1998). "Executive compensation and incentives: the impact of takeover legislation". NBER Working Paper 6830.
Bertrand, M., Mullainathan, S. (2001). "Are CEOs rewarded for luck? The ones without principals are". Quarterly Journal of Economics 116, 901–932.

Bertrand, M., Mullainathan, S. (2000). "Agents with and without principals". American Economic Review 90, 203–208.

Best, R., Zhang, H. (1993). "Alternative information sources and the information content of bank loans". Journal of Finance 48 (4), 1507–1522.

Bethel, J.E., Gillan, S. (2002). "The impact of the institutional and regulatory environment on shareholder voting". Financial Management Journal 31 (4), 29–54.

Bhagat, S., Black, B.S. (1999). "The uncertain relationship between roard composition and firm performance". Business Lawyer 54, 921–963.

Bharadwaj, A., Shivdasani, A. (2003). "Valuation effects of bank financing in acquisitions". Journal of Financial Economics 67 (1), 113–148.

Bhide, A. (1993). "The hidden costs of stock market liquidity". Journal of Financial Economics 34, 31–51.

Biais, B., Perotti, E. (2002). "Machiavellian privatization". American Economic Review 92, 240–258.

Bianco, M., Casavola, P., Ferrando, A. (1997). "Pyramidal groups and external finance: an empirical investigation". Servizio Studi Banca d'Italia, Rome, Italy, Working Paper.

Billett, M.T., Flannery, M.J., Garfinkel, J.A. (1995). "The effect of lender identity on a borrowing firm's equity return". Journal of Finance 50 (2), 699–718.

Bizjak, J.M., Lemmon, M.L., Naveen, L. (2000). "Has the use of peer groups contributed to higher levels of executive compensation?" (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=252544.)

Black, B.S. (1990). "Shareholder passivity reexamined". Michigan Law Review 89, 520.

Black, B.S., Coffee, J.C. Jr. (1994). "Hail Britannia?: Institutional investor behavior under limited regulation". Michigan Law Review 92, 1997–2087.

Black, B.S., Gilson, R.J. (1998). "Venture capital and the structure of capital markets: banks versus stock markets". Journal of Financial Economics 47, 243–277.

Black, B. (1992). "Agents watching agents: the promise of institutional investor voice". UCLA Law Review 39, 811–892.

Black, B. (1998). "Shareholder activism and corporate governance in the U.S." In: Newman, P. (Ed.), The New Palgrave Dictionary of Economics and the Law. Macmillan Reference Limited, London and Basingstoke.

Black, B. (2000a). "The core institutions that support strong securities markets". Business Lawyer 55, 1565–1607.

Black, B. (2000b). "Strengthening Brazil's securities markets". Revista de Direito Mercantil, Economico e Financiero 120, 41–55. (Also available at: http://ssrn.com/abstract=247673.)

Black, B., Kraakman, R., Tarassova, A. (2000). "Russian privatization and corporate governance: what went wrong?" Stanford Law Review 52, 1731–1808.

Black, B., Kraakman, R., Hay, J. (1996). "Corporate law from scratch". In: Frydman, R., Gray, C.W., Rapaczynski, A. (Eds.), Corporate Governance in Central Europe and Russia, Volume 2. Insiders and the State. Central European University Press, distributed by Oxford University Press, New York, Budapest.

Bloch, F., Hege, U. (2000). Multiple Shareholders and Control Contests. Tilburg University, mimeo.

Böhmer, E. (2000). "Business groups, bank control, and large shareholders: an analysis of German takeovers". Journal of Financial Intermediation 9, 117–148.

Bolton, P. (1995). "Privatization, and the separation of ownership and control: lessons from Chinese enterprise reform". Economics of Transition 3 (1), 1–12.

Bolton, P., Freixas, X. (2000). "Equity, bonds, and bank debt: capital structure and financial market equilibrium under asymmetric information". Journal of Political Economy 108, 324–351.

Bolton, P., Scharfstein, D.S. (1996). "Optimal debt structure and the number of creditors". Journal of Political Economy 104, 1–25.

Bolton, P., von Thadden, E.-L. (1998a). "Liquidity and control: a dynamic theory of corporate ownership structure". Journal of Institutional and Theoretical Economics 154, 177–211.

Bolton, P., von Thadden, E.-L. (1998b). "Blocks, liquidity, and corporate control". Journal of Finance 53, 1–25.

Bolton, P., Xu, C. (2001). "Ownership and managerial competition: employee, customer, or outside ownership". STICERD Discussion Paper No. TE/01/412, London School of Economics.

Bolton, P., Freixas, X., Shapiro, J. (2005). "Conflicts of interest, information provision and competition in the financial services industry". Journal of Financial Economics, in press.
Bolton, P., Scheinkman, J.A., Xiong, W. (2003). "Executive compensation and short-termist behavior in speculative markets". NBER Working Paper Series, vol. w9722.
Boot, A.W.A. (2000). "Relationship banking: what do we know?" Journal of Financial Intermediation 9, 7–25.
Borokhovich, K.A., Brunarski, K.R., Parrino, R. (1997). "CEO contracting and antitakeover amendments". Journal of Finance 52, 1495–1517.
Börsch-Supan, A., Köke, J. (2002). "An applied econometricians' view of empirical corporate governance studies". German Economic Review 3 (3), 295–326.
Borstadt, L.F., Zwirlein, T.J. (1992). "The efficient monitoring role of proxy contests: an empirical analysis of post-contest control changes and firm performance". Financial Management 21 (3), 22–34.
Bower, T. (1995). Maxwell: The Final Verdict. HarperCollins, London.
Bradley, M. (1980). "Interfirm tender offers and the market for corporate control". Journal of Business 53, 345–376.
Bradley, M., Desai, A., Kim, E.H. (1983). "The rationale behind interfirm tender offers: information or synergy". Journal of Financial Economics 11, 183–206.
Bradley, M., Desai, A., Kim, E.H. (1988). "Synergistic gains from corporate acquisitions and their division between the stockholders of the target and acquiring firms". Journal of Financial Economics 21, 3–40.
Bratton, W.W., McCahery, J.A. (1999). "Comparative corporate governance and the theory of the firm: the case against global cross reference". Columbia Journal of Transnational Law 38, 213–297.
Breeden, R. (2003). "Restoring trust". (http://news.findlaw.com/hdocs/docs/worldcom/corpgov82603rpt.pdf.)
Brickley, J.A., Bhagat, S., Lease, R.C. (1985). "The impact of long-range managerial compensation plans on shareholder wealth". Journal of Accounting and Economics 7, 115–129.
Brickley, J.A., Linck, J.S., Coles, J.L. (1999). "What happens to CEOs after they retire? New evidence on career concerns, horizon problems, and CEO incentives". Journal of Financial Economics 52, 341–377.
Bris, A., Cabolis, C. (2002). "Adopting better corporate governance: evidence from cross-border mergers". ICF Working Paper No. 02-32, Yale University, New Haven.
Bulow, J., Huang, M., Klemperer, P. (1999). "Toeholds and takeovers". Journal of Political Economy 107, 427–454.
Burgess, L.R. (1963). Top Executive Pay Packages. Free Press, New York.
Burkart, M. (1995). "Initial shareholdings and overbidding in takeover contests". Journal of Finance 50 (5), 1491–1515.
Burkart, M. (1999). "The economics of takeover regulation". Discussion Paper, Stockholm School of Economics, Stockholm.
Burkart, M., Panunzi, F. (2006). "Agency conflicts, ownership concentration, and legal shareholder protection". Journal of Financial Intermediation 15 (1), 1–31.
Burkart, M., Gromb, D., Panunzi, F. (1997). "Large shareholders, monitoring, and the value of the firm". Quarterly Journal of Economics 112, 693–728.
Burkart, M., Gromb, D., Panunzi, F. (1998). "Why higher takeover premia protect minority shareholders". Journal of Political Economy 106, 172–204.
Burns, N., Kedia, S. (2005). "Do executive stock options generate incentives for earnings management? Evidence from accounting restatements". Journal of Financial Economics, in press.
Byrd, J.W., Hickman, K.A. (1992). "Do outside directors monitor managers? Evidence from tender offer bids". Journal of Financial Economics 32, 195–221.
Cable, J.R. (1985). "Capital market information and industrial performance: the role of West German banks". Economic Journal 95 (377), 118–132.
Cadbury Committee (1992). Report of the Committee on the Financial Aspects of Corporate Governance. Gee and Co., London.
Calomiris, C.W. (2000). U.S. Bank Deregulation in Historical Perspective. Cambridge University Press.
Calomiris, C.W., Kahn, C.M. (1991). "The role of demandable debt in structuring optimal banking arrangements". American Economic Review 81, 497–513.

Cardon Report (1998). Brussels Stock Exchange, Report of the Belgian Commission on Corporate Governance. Now included as Part I of the Dual Code of the Brussels Stock Exchange and the Belgian Banking & Finance Commission. Corporate Governance for Belgian Listed Companies, Belgium. (www.cbf.be/pe/pec/en_ec01.htm.)

Carleton, W.T., Nelson, J.M., Weisbach, M. (1998). "The influence of institutions on corporate governance through private negotiations: evidence from TIAA-CREF". The Journal of Finance 53, 1335–1362.

Carlin, W., Mayer, C. (2003). "Finance, investment and growth". Journal of Financial Economics 69 (1), 191–226.

Carosso, V. (1970). "Washington and Wall Street: the new deal and investment bankers, 1933–1940". Business History Review 44, 425–445.

Carosso, V.P. (1973). "The wall street money trust from Pujo through Medina". Business History Review 47, 421–437.

Carosso, V.P. (1985). "American private banks in international banking and industrial finance, 1870–1914". In: Atack, J. (Ed.), Business and Economic History, Second series, Volume 14. University of Illinois, Champaign.

Carver, T.N. (1925). The Present Economic Revolution in the U.S. Little, Brown and Company, Boston.

Cary, W.L. (1974). "Federalism and corporate law: reflections upon delaware". The Yale Law Journal 83, 663.

Chandler, A.D. (1976). "The development of modern management structure in the U.S. and the U.K." Hannah, L. (Ed.), Management Strategy and Business Development: An Historical and Comparative Study. Macmillan, London.

Chang, C. (1992). "Capital structure as an optimal contract between employees and investors". Journal of Finance 47, 1141–1158.

Cheffins, , B.R. (2002). "Putting Britain on the Roe Map: the emergence of the Berle-Means Corporation in the U.K." McCahery, J.A., Moerland, P., Raaijmakers, T., Renneboog, L. (Eds.), Corporate Governance Regimes: Convergence and Diversity. Oxford University Press, Oxford.

Cheffins, B. (2003). "Will executive pay globalize along American lines?" Corporate Governance: An International Review 11, 12–24.

Chemla, G. (2005). "Hold-up, stakeholders and takeover threats". Journal of Financial Intermediation 14, 376–397.

Cheng, Q., Warfield, T. (2005). "Equity incentives and earnings management". The Accounting Review 80, 441–476.

Chernow, R. (1990). The House of Morgan. Simon & Schuster, London.

Chidambaran, N.K., John, K. (2000). "Relationship investing: large shareholder monitoring with managerial cooperation". Working Paper # FIN-98-044, Stern School of Business, New York University, New York, NY.

Cho, M.-H. (1998). "Ownership structure, investment, and the corporate value: an empirical analysis". Journal of Financial Economics 47, 103–321.

Chung, K.H., Kim, J.-K. (1999). "Corporate ownership and the value of a vote in an emerging market". Journal of Corporate Finance: Contracting, Governance and Organization 5, 35–54.

Claessens, S., Djankov, S., Lang, L.H.P. (2000). "The separation of ownership and control in East Asian corporations". Journal of Financial Economics 58, 81–112.

Claessens, S., Djankov, S., Fan, J., Lang, L. (2002). "Expropriation of minority shareholders in East Asia". Journal of Finance 57, 2741–2771.

Clark, R.C. (1986). Corporate Law. Little, Brown, Boston.

Coates, J.C. IV (2000). "The contestability of corporate control: a critique of the scientific evidence on takeover defenses". Texas Law Review 79, 271.

Coffee, J.C. (1991). "Liquidity versus control: the institutional investor as corporate monitor". Columbia Law Review 91, 1277–1366.

Coffee, J.C. (1999). "The future as history: the prospects for global convergence in corporate governance and its implications". Northwestern University Law Review 93, 641–708.

Coffee, J.C. (2002). "Convergence and its critics: what are the preconditions to the separation of ownership and control?" In: McCahery, J.A., Moerland, P., Raaijmakers, T., Renneboog, L. (Eds.), Corporate Governance Regimes: Convergence and Diversity. Oxford University Press, Oxford.

Cole, R.A. (1998). "The importance of relationships to the availability of credit". Journal of Banking and Finance 22 (6–8), 959–977.

Comment, R., Schwert, G.W. (1995). "Poison or placebo? Evidence on the deterrence and wealth effects of modern antitakeover measures". Journal of Financial Economics 39, 3–43.

Commonwealth Association (1999). Corporate Governance Codes and Principles in the Commonwealth, Principles of Best Business Practice for the Commonwealth Provenance Commonwealth Association for Corporate Governance (CACG).

Confederation of Indian Industry (CII) (1998). Desirable Corporate Governance: A Code, CII, New Delhi, April 1998. (http://www.combinet.org/governance/finalver/listof.htm.)

Conyon, M.J., Murphy, K.J. (2000). "The Prince and the Pauper? CEO pay in the U.S. and U.K." Economic Journal 110, 640–671.

Conyon, M.J., Peck, S.I. (1998). "Recent developments in U.K. corporate governance". In: Buxton, T., Chapman, P., Temple, P. (Eds.), Britain's Economic Performance. Routledge, London and New York.

Conyon, M., Mallin, C. (1997). "A review of compliance with cadbury". Journal of General Management 22, 24–37.

Core, J.E., Guay, W.R., Larcker, D.F. (2003). "Executive equity compensation and incentives: a survey". Federal Reserve Bank of New York Economic Policy Review 8 (1), 27–50.

Core, J.E., Guay, W.R., Rusticus, T.O. (2006). "Does weak governance cause weak stock returns? An examination of firm operating performance and investors' expectations". Journal of Finance 61 (2).

Cornelli, F., Li, D.D. (2002). "Risk arbitrage in takeovers". Review of Financial Studies 15, 837–868.

Cosh, A.D., Hughes, A. (1989). "Ownership, management incentives and company performance: an empirical analysis for the U.K. 1968–80". Discussion Paper No. 11/89, University of Cambridge.

Cotter, J.F., Shivdasani, A., Zenner, M. (1997). "Do independent directors enhance target shareholder wealth during tender offers?" Journal of Financial Economics 43, 195–218.

Cubbin, J.S., Leech, D. (1983). "The effect of shareholding dispersion on the degree of control in British companies: theory and measurement". Economic Journal 93, 351–369.

Cushing, H.A. (1915). Voting Trusts; A Chapter in Recent Corporate History. Macmillan Co., New York.

Da Rin, M., Hellmann, T. (2002). "Banks as catalysts for industrialization". Journal of Financial Intermediation 11, 366–397.

Dahlquist, M., Pinkowitz, L., Stulz, R.M., Williamson, R. (2003). "Corporate governance and the home bias". Journal of Financial and Quantitative Analysis 38, 87–110.

Daily, C., Johnson, J.L., Ellstrand, A.E., Dalton, D.R. (1998). "Compensation committee composition as a determinant of CEO compensation". Academy of Management Journal 41, 209–220.

Danielson, M.G., Karpoff, J.M. (1998). "On the uses of corporate governance provisions". Journal of Corporate Finance: Contracting, Governance and Organization 4, 347–371.

David, R., Brierley, J.E.C. (1985). Major Legal Systems in the World Today: An Introduction to the Comparative Study of Law, 3rd edn. Stevens, London.

Davies, P.L., Prentice, D.D., Gower, L.C.B. (1997). Gower's Principles of Modern Company Law, 6th edn. Sweet & Maxwell, London.

Davis, G.F., Useem, M. (2002). "Top management, company directors, and corporate control". In: Pettigrew, A., Thomas, H., Whittington, R. (Eds.), Handbook of Strategy and Management. Sage Publications, London, pp. 233–259.

Davis, G.F. Kim, E.H. (2005). "Business ties and proxy voting by mutual funds". Journal of Financial Economics, in press.

DeAngelo, H., DeAngelo, L. (1985). "Managerial ownership of voting rights: a study of public corporations with dual classes of common stock". Journal of Financial Economics 14, 33–69.

DeAngelo, H., DeAngelo, L. (1989). "Proxy contests and the governance of publicly held corporations". Journal of Financial Economics 23 (1), 29–59.

Debreu, G. (1959). Theory of Value: An Axiomatic Analysis of Economic Equilibrium. Yale University Press, New Haven and London.

DeFusco, R.A., Johnson, R.R., Zorn, T.S. (1990). "The effect of executive stock option plans on stockholders and bondholders". Journal of Finance 45, 617–627.

DeGryse, H., Van Cayseele, P. (1998). "Informatie en de rol van financiele intermediatie". In: De Bondt, R., Veugelers, R. (Eds.), Informatie en Kennis in de Economie. Leuvense Universitaire Pers, pp. 463–477.

DeGryse, H., Van Cayseele, P. (2000). "Relationship lending within a bank-based system: evidence from European small business data". Journal of Financial Intermediation 9, 90–109.

De Jong, A., DeJong, D.V., Mertens, G., Wasley, C.E. (2005). "The role of self-regulation in corporate governance: evidence and implications from the Netherlands". Journal of Corporate Finance 11, 473–503.

DeLong, J.B. (1998). "Robber barons". In: Aslund, A. (Ed.), Perspectives on Russian Economic Development. Carnegie Endowment, Washington, D.C.

DeLong, J.B. (1991). "Did J.P. Morgan's men add value? An economist's perspective on financial capitalism". In: Temin, P. (Ed.), Inside the Business Enterprise: Historical Perspectives on the Use of Information. University of Chicago Press for National Bureau of Economic Research, Chicago and London, pp. 205–236.

Demsetz, H. (1968). "The cost of transacting". The Quarterly Journal of Economics 82, 33–53.

Demsetz, H. (1986). "Corporate control, insider trading, and rates of return". American Economic Review 76, 313–316.

Demsetz, H., Lehn, K. (1985). "The structure of corporate ownership: causes and consequences". Journal of Political Economy 93, 1155–1177.

Demsetz, H., Villalonga, B. (2001). "Ownership structure and corporate performance". Journal of Corporate Finance 7, 209–233.

Desai, M.A., Dyck, A., Zingales, L. (2004). Theft and taxes, ECGI—Finance Paper No. 63/2004 (December). (http://ssrn.com/abstract=629350.)

Deutsche Schutzvereinigung fuer Wertpapierbesitz (2000). DSW Europe Study. A Study comparing Minority Shareholders' Rights, Voting Rights and Proxy Voting in Europe. DSW and IRRC, Duesseldorf and Washington, D.C.

Dewatripont, M. (1993). "The 'leading shareholder' strategy, takeover contests and stock price dynamics". European Economic Review 37, 983–1004.

Dewatripont, M., Maskin, E. (1995). "Contractual contingencies and renegotiation". Rand Journal of Economics 26, 704–719.

Dewatripont, M., Tirole, J. (1994). "A theory of debt and equity: diversity of securities and manager-shareholder congruence". Quarterly Journal of Economics 109, 1027–1054.

Dey Committee (1994). Guidelines for Improved Corporate Governance. Toronto Stock Exchange Committee on Corporate Governance in Canada, Toronto Stock Exchange, Toronto, Canada.

Diamond, D.W. (1984). "Financial intermediation and delegated monitoring". Review of Economic Studies 51, 393–414.

Diamond, D.W. (1989). "Reputation acquisition in debt markets". Journal of Political Economy 97, 828–862.

Diamond, D.W. (1991). "Monitoring and reputation: the choice between bank loans and directly placed debt". Journal of Political Economy 99, 689–721.

Diamond, D.W. (1993). "Bank loan maturity and priority when borrowers can refinance". In: Mayer, C., Vives, X. (Eds.), Capital Markets and Financial Intermediation. Cambridge University Press, Cambridge, New York and Melbourne, pp. 46–68.

Diamond, D.W., Rajan, R.G. (2001). "Liquidity risk, liquidity creation, and financial fragility: a theory of banking". Journal of Political Economy 109, 287–327.

Dodd, M. (1932). "For whom are corporate managers trustees?" Harvard Law Review 45, 1145.

Dodd, P., Warner, J.-B. (1983). "On corporate governance: a study of proxy contests". Journal of Financial Economics 11 (1–4), 401–438.

Dunlavy, C.A. (1998). "Corporate governance in late 19th century Europe and the U.S.: the case of shareholder voting rights". In: Hopt, K.J., Kanda, H., Roe, M.J., Wymeersch, E., Prigge, S. (Eds.), Comparative

Corporate Governance. The State of the Art and Emerging Research. Oxford University Press, Oxford, pp. 5–40.
Dyck, A., Zingales, L. (2003a). "The media and asset prices". Unpublished.
Dyck, A., Zingales, L. (2003b). "The bubble and the media". In: Cornelius, P.K., Kogut, B. (Eds.), Corporate Governance and Capital Flows in a Global Economy. Oxford University Press, New York and Oxford, pp. 83–102. Chapter 4.
Dyck, A., Zingales, L. (2004). "Private benefits of control: an international comparison, forthcoming". Journal of Finance 59 (2), 537–600.
Easterbrook, F.H. (1997). "International corporate differences: market or law?" Journal of Applied Corporate Finance 9, 23–29.
Easterbrook, F.H., Fischel, D.R. (1981). "The proper role of target's management in responding to a tender offer". Harvard Law Review 94, 1161–1204.
Easterbrook, F.H., Fischel, D.R. (1991). The Economic Structure of Corporate Law. Harvard University Press, Cambridge, MA and London.
Edwards, F.R., Hubbard, R.G. (2000). "The growth of institutional stock ownership: a promise unfulfilled". Journal of Applied Corporate Finance 13, 92–104.
Edwards, J., Fischer, K. (1994). Banks, Finance and Investment in Germany. Cambridge University Press, Cambridge, New York and Melbourne.
Edwards, J., Ogilvie, S. (1996). "Universal banks and German industrialization: a reappraisal". Economic History Review 49, 427–446.
Eells, R.S.F. (1960). The Meaning of Modern Business: An Introduction to the Philosophy of Large Corporate Enterprise. Columbia University Press, New York.
Eisenberg, M.A. (1976). The Structure of the Corporation. Little, Brown and Company, Boston and Toronto.
Elsas, R.K., Krahnen, J.P. (1998). "Is relationship lending special? Evidence from credit file data in Germany". Journal of Banking and Finance 22, 1283–1316.
Erickson, M., Hanlon, M., Maydew, E. (2004). "Is there a link between executive equity holdings and accounting fraud?" Journal of Accounting Research, in press.
European Association of Securities Dealers (2000). EASD Corporate Governance Principles and Recommendations. EASD, Brussels.
European Corporate Governance Network (1997). The Separation of Ownership and Control: A Survey of 7 European Countries. Preliminary Report to the European Commission. European Corporate Governance Network, Brussels.
Euroshareholders (2000). Euroshareholders Corporate Governance Guidelines 2000. European Shareholders Group, Brussels.
Faccio, M., Lang, H.P. (2002). "The ultimate ownership of western European corporations". Journal of Financial Economics 65 (3), 365–395.
Faure-Grimaud, A., Gromb, D. (2004). "Public trading and private incentives". Review of Financial Studies 17 (4), 985–1014.
Fich, E.M., White, L.J. (2005). "Why do CEOs reciprocally sit on each other's boards?" Journal of Corporate Finance 11, 175–195.
Fischel, D., Bradley, M. (1986). "The role of liability rules and the derivative suit in corporate law: a theoretical and empirical analysis". Com. L. Rev. 71, 261.
Fishman, M.J. (1988). "A theory of preemptive takeover bidding". Rand Journal of Economics 19, 88–101.
FitzRoy, F.R., Kraft, K. (1993). "Economic effects of codetermination". Scandinavian Journal of Economics 95, 365–375.
Florence, P.S. (1947). "The statistical analysis of joint stock company control". Journal of the Royal Statistical Society 110, 1–26.
Florence, P.S. (1953). The Logic of British and American Industry; A Realistic Analysis of Economic Structure and Government. Routledge and K. Paul, London.
Florence, P.S. (1961). Ownership, Control and Success of Large Companies: An Analysis of English Industrial Structure and Policy, 1936–1951. Sweet & Maxwell, London.

Fohlin, C. (1997). "Bank securities holdings and industrial finance before World War I: Britain and Germany compared". Business and Economic History 26, 463–475.
Fohlin, C. (1999a). "Capital mobilisation and utilisation in latecomer economies: Germany and Italy compared". European Review of Economic History 3, 139–174.
Fohlin, C. (1999b). "The rise of interlocking directorates in imperial Germany". Economic History Review 52, 307–333.
Franks, J.R., Harris, R.S. (1989). "Shareholder wealth effects of corporate takeovers: the U.K. experience 1955–1985". Journal of Financial Economics 23 (2), 225–249.
Franks, J., Mayer, C. (1995). "Ownership and control". In: Siebert, H. (Ed.), Trends in Business Organization: Do Participation and Cooperation Increase Competitiveness? Mohr, Siebeck, Tubingen.
Franks, J., Mayer, C. (1996). "Hostile takeovers and the correction of managerial failure". Journal of Financial Economics 40, 163–181.
Franks, J., Mayer, C. (2001). "Ownership and control of German corporations". Review of Financial Studies 14 (4), 943–977.
Franks, J., Mayer, C., Rossi, G. (2003). "The origin and evolution of ownership and control". European Corporate Governance Institute Finance Working Paper N° 09/2003.
Franks, J.R., Sussman, O. (2005a). "Financial innovations and corporate insolvency". Journal of Financial Intermediation 14 (3), 283–317. Said Business School, Oxford University, Oxford, U.K. mimeo.
Franks, J.R., Sussman, O. (2005b). "Financial distress and bank restructuring of small to medium size U.K. companies". Review of Finance 9 (1), 65–96.
Franks, J.R., Torous, W.N. (1989). "An empirical investigation of U.S. firms in reorganization". Journal of Finance 44, 747–769.
Franks, J.R., Mayer, C., Renneboog, L. (2001). "Who disciplines management in poorly performing companies?" Journal of Financial Intermediation 10, 209–248.
Freeman, R.B., Lazear, E.P. (1995). "An economic analysis of works councils". In: Rogers, J., Streeck, W. (Eds.), Works Councils: Consultation, Representation, and Cooperation in Industrial Relations. National Bureau of Economic Research Comparative Labor Markets Series. University of Chicago Press, Chicago and London.
Frick, B., Kluge, N., Streeck, W. (Eds.) (1999). Die wirtschaftlichen Folgen der Mitbestimmung. Campus Verlag, Frankfurt/Main, New York.
Fukao, M. (1995). Financial Integration, Corporate Governance, and the Performance of Multinational Companies. Brookings Institution, Washington, D.C.
Galbraith, J.K. (1967). The New Industrial State. H. Hamilton, London.
Gao, P., Shrieves, R.E. (2002). "Earnings management and executive compensation: a case of overdose of option and underdose of salary?" University of Tennessee, Knoxville, TN, unpublished.
Gerschenkron, A. (1962). Economic Backwardness in Historical Perspective: A Book of Essays. Harvard University Press, Cambridge, MA.
Gerum, E., Steinmann, H., Fees, W. (1988). Der Mitbestimmte Aufsichtsrat: Eine Empirische Untersuchung. Poeschel, Stuttgart.
Gerum, E., Wagner, H. (1998). "Economics of labor co-determination in view of corporate governance". In: Hopt, K.J., Kanda, H., Roe, M.J., Wymeersch, E., Prigge, S. (Eds.), Comparative Corporate Governance: The State of the Art and Emerging Research. Oxford University Press, Clarendon Press, Oxford and New York, pp. 341–360.
Gibbons, R., Murphy, K.J. (1990). "Relative performance evaluation for chief executive officers". Industrial and Labor Relations Review 43, 30S–51S.
Gibbons, R., Murphy, K.J. (1992). "Optimal incentive contracts in the presence of career concerns: theory and evidence". Journal of Political Economy 100, 468–505.
Gillan, S., Starks, L. (1998). "A survey of shareholder activism: motivation and empirical evidence". Contemporary Finance Digest 2, 10–34.
Gilson, R. (1981). "A structural approach to corporations: the case against defensive tactics in tender offers". Stanford Law Review 33, 819–891.

Gilson, R., Kraakman, R. (1991). "Reinventing the outside director: an agenda for institutional investors". Stanford Law Review 43, 863–906.
Gilson, R.J. (2001). "Unocal fifteen years later (and what we can do about it)". Delaware Journal of Corporate Law 26, 491.
Gilson, R.J. (2002). "Lipton and Rowe's apologia for delaware: a short reply". Delaware Journal of Corporate Law 27, 37–52.
Gilson, R.J., Schwartz, A. (2001). "Sales and elections as methods for transferring corporate control". Theoretical Inquiries in Law 2 (2), 783.
Goergen, M., Renneboog, L. (2001). "Strong managers and passive institutional investors in the U.K." In: Barca, F., Becht, M. (Eds.), The Control of Corporate Europe. Oxford University Press, Oxford, pp. 259–284.
Goergen, M., Renneboog, L. (2003). "Why are the levels of control (so) different in German and UK companies? Evidence from initial public offerings". Journal of Law and Organization 19 (1), 141–175.
Goldberg, V.P. (1976). "Regulation and administered contracts". Bell Journal of Economics and Management Science 7 (426), 439–441.
Gomes, A., Novaes, W. (2000). "Sharing of control as a corporate governance mechanism". Discussion Paper, Wharton School of Business, Philadelphia, PA.
Gompers, P.A., Ishii, J., Metrick, A. (2003). "Corporate governance and equity prices". Quarterly Journal of Economics 118 (1), 107–155.
Gompers, P., Lerner, J. (1999). The Venture Capital Cycle. MIT Press, Cambridge.
Gordon, J.N. (2003). "Governance gailures of the enron board and the new information order of Sarbanes-Oxley". University of Connecticut Law Review 35 (symposium issue), 1125.
Gordon, R.A. (1945). Business Leadership in the Large Corporation. The Brookings Institution, Washington, D.C.
Gorton, G., Schmid, F. (1999). "Corporate governance, ownership dispersion and efficiency: empirical evidence from Austrian cooperative banking". Journal of Corporate Finance: Contracting, Governance and Organization 5, 119–140.
Gorton, G., Schmid, F.A. (2000). "Universal banking and the performance of German firms". Journal of Financial Economics 58, 29–80.
Gorton, G., Schmid, F.A. (2004). "Capital, labor, and the firm: a study of German codetermination". Journal of the European Economic Association 2, 863–905.
Gorton, G., Winton, A. (1998). "Banking in transition economies: does efficiency require instability?" Journal of Money, Credit, and Banking 30 (3), 621–650.
Gorton, G., Winton, A. (2003). "Banking". In: Constantinides, G., Harris, M., Stulz, R. (Eds.), Handbook of the Economics of Finance. North Holland, Amsterdam.
Gourevitch, P.A., Shinn, J. (2007). Political Power and Corporate Control: The New Global Politics or Corporate Governance. Princeton University Press, Princeton, NJ. (http://press.princeton.edu/titles/8086.html.)
Greenbury Committee (1995). Study Group on Directors' Remuneration, Final Report, Gee, London.
Greenslade, R. (1992). Maxwell's Fall. Simon & Schuster, London.
Gregory, H.J. (2000). International Comparison of Corporate Governance: Guidelines and Codes of Best Practice in Developing and Emerging Markets. Weil, Gotshal & Manges LLP, New York.
Gregory, H.J. (2001a). International Comparison of Corporate Governance: Guidelines and Codes of Best Practice in Developed Markets. Weil, Gotshal & Manges LLP, New York.
Gregory, H.J. (2001b). International Comparison of Board "Best Practices"—Investor Viewpoints. Weil, Gotshal & Manges LLP, New York.
Gregory, H.J. (2002). Comparative Study of Corporate Governance Codes relevant to the European Union and its Member States, Report to the European Commission. Weil, Gotshal & Manges LLP, New York.
Gromb, D. (1993). "Is one share—one vote optimal?" Discussion Paper No. 177, Financial Markets Group, London School of Economics, London.
Grossman, S.J., Hart, O.D. (1983). "An analysis of the principal-agent problem". Econometrica 51, 7–45.
Grossman, S.J., Hart, O.D. (1986). "The costs and benefits of ownership: a theory of vertical and lateral integration". Journal of Political Economy 94, 691–719.

Grossman, S.J., Hart, O.D. (1988). "One share—one vote and the market for corporate control". Journal of Financial Economics 20, 175–202.
Grossman, S., Hart, O. (1980). "Takeover bids, the free-rider problem and the theory of the corporation". Bell Journal of Economics 11, 42–64.
Grundfest, J.A. (1990). "Subordination of American capital". Journal of Financial Economics 27, 89–114.
Grundfest, J.A. (1993). "Just vote no: a minimalist strategy for dealing with barbarians inside the gates". Stanford Law Review 45, 857–937.
Gugler, K. (Ed.) (2001). Corporate Governance and Economic Performance. Oxford University Press, Oxford.
Guinnane, T.W. (2002). "Delegated monitors, large and small: Germany's banking system, 1800–1914". Journal of Economic Literature 40, 73–124.
Gurdon, M.A., Rai, A. (1990). "Codetermination and enterprise performance: empirical evidence from West Germany". Journal of Economics and Business 42, 289–302.
Gutiérrez, M. (2003). "An economic analysis of corporate directors' fiduciary duties". Rand Journal of Economics 34, 516–535.
Habib, M.A., Ljungqvist, A.P. (2002). Firm Value and Managerial Incentives. NYU and London Business School, mimeo.
Hall, B.J., Liebman, J.B. (1998). "Are CEOs really paid like bureaucrats?" Quarterly Journal of Economics 113, 653–691.
Hall, B.J., Liebman, J.B. (2000). "The taxation of executive compensation". In: Poterba, J. (Ed.), Tax Policy and the Economy. MIT Press, Cambridge, MA.
Hallock, K.F. (1997). "Reciprocally interlocking boards of directors and executive compensation". Journal of Financial and Quantitative Analysis 32 (3), 331–344.
Hallock, K.F., Murphy, K.J. (Eds.) (1999). The Economics of Executive Compensation. 2 vols. Elgar; distributed by American International Distribution Corporation Williston, Cheltenham, U.K. and Northampton, MA.
Hampel Committee (1998). Committee on Corporate Governance, Final Report. Gee, London.
Hannah, L. (1974). "Takeover bids in Britain before 1950: an exercise in business 'Pre-History' ". Business History 16, 65–77.
Hannah, L. (Ed.) (1976). Management Strategy and Business Development: An Historical and Comparative Study. Macmillan, London.
Hansmann, H. (1996). The Ownership of Enterprise. The Belknap Press of Harvard University Press, Cambridge, MA.
Hansmann, H., Kraakman, R. (2001). "The end of history for corporate law". Georgetown Law Journal 89, 439–468.
Hansmann, H., Hertig, G., Hopt, K.J., Kanda, H., Rock, E.B., Kraakman, R., Davies, P. (2004). The Anatomy of Corporate Law: A Comparative and Functional Approach. Oxford University Press, Oxford.
Harhoff, D., Korting, T. (1998). "Lending relationships in Germany—empirical evidence from survey data". Journal of Banking and Finance 22 (10,11), 1317–1353.
Harris, M., Raviv, A. (1988a). "Corporate control contests and capital structure". Journal of Financial Economics 20, 55–86.
Harris, M., Raviv, A. (1988b). "Corporate governance: voting rights and majority rules". Journal of Financial Economics 20, 203–235.
Harris, M., Raviv, A. (1992). "Financial contracting theory". In: Laffont, J.J. (Ed.), Advances in Economic Theory: Sixth World Congress. Volume 2. Econometric Society Monographs, no. 21. Cambridge University Press, Cambridge, New York and Melbourne, pp. 64–150.
Hart, O. (1988). "On SEC's one-share-one-vote decision". Wall Street Journal (July 14).
Hart, O. (1995). Firms Contracts, and Financial Structure. Oxford University Press, London.
Hart, O., Moore, J. (1990). "Property rights and the nature of the firm". Journal of Political Economy 98, 1119–1158.
Hart, O., Moore, J. (1995). "Debt and seniority: an analysis of the role of hard claims in constraining management". American Economic Review 85, 567–585.

Hart, O., Moore, J. (1996). "The governance of exchanges: members' cooperatives versus outside ownership". Oxford Review of Economic Policy 12 (4), 53–69.
Hart, O., Moore, J. (1998). "Cooperatives vs. outside ownership". NBER Working Paper No. w6421.
Hart, O., Shleifer, A., Vishny, R. (1997). "The proper scope of government: theory and an application to prisons". Quarterly Journal of Economics 112 (4), 1126–1161.
Hartzell, J.C., Starks, L.T. (2003). "Institutional investors and executive compensation". Journal of Finance 58, 2351–2374.
Healy, P.M., Palepu, K.G. (2001). "Information asymmetry, corporate disclosure, and the capital markets: a review of the empirical disclosure literature". Journal of Accounting and Economics 31 (1–3), 405–440.
Heflin, F., Shaw, K.W. (2000). "Blockholder ownership and market liquidity". Journal of Financial and Quantitative Analysis 35, 621–633.
Hellman, T. (1997). A Theory of Corporate Venture Capital. Stanford University Graduate School of Business, Stanford.
Hellwig, M.F. (2000a). "Financial intermediation with risk aversion". Review of Economic Studies 67, 719–742.
Hellwig, M.F. (2000b). "On the economics and politics of corporate finance and corporate control". In: Vives, X. (Ed.), Corporate Governance. Theoretical and Empirical Perspectives. Cambridge University Press, Cambridge, pp. 95–134.
Hennessy, C.A., Levy, A. (2002). A Unified Model of Distorted Investment: Theory and Evidence. Haas School of Business, U.C. Berkeley.
Hermalin, B. (2005). "Trends in corporate governance". Journal of Finance 60 (5), 2351–2384.
Hermalin, B.E., Weisbach, M.S. (1991). "The effects of board composition and direct incentives on firm performance". Financial Management 20, 101–112.
Hermalin, B.E., Weisbach, M.S. (1998). "Endogenously chosen boards of directors and their monitoring of the CEO". American Economic Review 88, 96–118.
Hermalin, B.E., Weisbach, M.S. (2003). "Boards of directors as an endogenously determined institution: a survey of the economic literature". Federal Reserve Bank of New York Economic Policy Review 9, 7–26.
Herman, E.S. (1981). Corporate Control, Corporate Power. Cambridge University Press, New York.
Heron, R.A., Lie, E. (2006). "Does backdating explain the stock price pattern around executive stock option grants?" Journal of Financial Economics, in press.
Hessen, R. (1983). "The modern corporation and private property: a reappraisal". Journal of Law and Economics 26, 273–289.
Higgins, R.C., Schall, L.D. (1975). "Corporate bankruptcy and conglomerate merger". Journal of Finance 30 (1), 93–113.
High Level Finance Committee on Corporate Governance (1999). Chapter 5, The Malaysian Code on Corporate Governance. High Level Finance Committee Report on Corporate Governance, Kuala Lumpur. (http://www.combinet.org/governance/finalver/listof.htm.)
Himmelberg, C.P., Hubbard, G.R., Palia, D. (1999). "Understanding the determinants of managerial ownership and the link between ownership and performance". Journal of Financial Economics 53, 353–384.
Hirschman, A.O. (1970). Exit, Voice, and Loyalty; Responses to Decline in Firms, Organizations, and States. Harvard University Press, Cambridge, MA.
Hirshleifer, D. (1995). "Mergers and acquisitions: strategic and informational issues". In: Jarrow, R.A., Maksimovic, V., Ziemba, W.T. (Eds.), Handbooks of Operations Research and Management Science. Elsevier, Amsterdam, pp. 839–885.
Hirshleifer, D., Thakor, A.V. (1994). "Managerial performance, boards of directors and takeover bidding". Journal of Corporate Finance: Contracting, Governance and Organization 1, 63–90.
Hirshleifer, D., Titman, S. (1990). "Share tendering strategies and the success of hostile takeover bids". Journal of Political Economy 98, 295–324.
Hoffmann-Burchardi, U. (1999). "Corporate governance rules and the value of control—a study of German dual-class shares", LSE Financial Markets Group Discussion Paper No. 315, London.

Hoffmann-Burchardi, U. (2000). "Unlocking Germany's corporate gates: the Vodafone-Mannesmann takeover and German corporate governance". Financial Markets Group Review, London School of Economics, London, UK.

Holderness, C.G., Sheehan, D.P. (1988). "The role of majority shareholders in publicly held corporations: an exploratory analysis". Journal of Financial Economics 20, 317–346.

Holderness, C. (2003). "A survey of blockholders and corporate control". Federal Reserve Bank of New York Economic Policy Review 9, 51–64.

Holderness, C.G. (2006). "A contrarian view of ownership concentration in the United States and around the world", AFA 2006 Boston Meetings. (Available at SSRN: http://ssrn.com/abstract=686175.)

Holl, P. (1975). "Effect of control type on the performance of the firm in the U.K." Journal of Industrial Economics 23, 257–271.

Holmstrom, B. (1979). "Moral hazard and observability". Bell Journal of Economics 10, 74–91.

Holmstrom, B. (1999). "Managerial incentive problems—a dynamic perspective". Review of Economic Studies 66 (1), 169–182.

Holmstrom, B., Kaplan, S.N. (2001). "Corporate governance and merger activity in the United States: making sense of the 1980s and 1990s". Journal of Economic Perspectives 15 (2), 121–144.

Holmstrom, B., Nalebuff, B. (1992). "To the raider goes the surplus: a re-examination of the free-rider problem". Journal of Economics and Management Strategy 1 (1), 37–62.

Holmstrom, B., Tirole, J. (1993). "Market liquidity and performance monitoring". Journal of Political Economy 101, 678–709.

Holmstrom, B., Ricart i Costa, J. (1986). "Managerial incentives and capital management". Quarterly Journal of Economics 101, 835–860.

Hong, H., Kubik, J.D. (2003). "Analyzing the analysts: career concerns and biased earnings forecasts". Journal of Finance 58, 313–351.

Hopt, K.J., Kanda, H., Roe, M.J., Wymeersch, E., Prigge, S. (Eds.) (1998). Comparative Corporate Governance: The State of the Art and Emerging Research. Clarendon Press, Oxford and New York.

Hopt, K.J., Wymeersch, E. (Eds.) (2003). Capital Markets and Company Law. Oxford University Press, Oxford.

Horiuchi, A., Packer, F., Fukuda, S. (1988). "What role has the 'Main Bank' played in Japan?" Journal of the Japanese and International Economy 2 (2), 159–180.

Hoshi, T. (1998). "Japanese corporate governance as a system". In: Hopt, K.J., Kanda, H., Roe, M.J., Wymeersch, E., Prigge, S. (Eds.), Comparative Corporate Governance. The State of the Art and Emerging Research. Clarendon Press, Oxford, pp. 847–875.

Hoshi, T., Kashyap, A. (2001). Corporate Financing and Governance in Japan. MIT Press, Cambridge, MA.

Hoshi, T., Kashyap, A., Scharfstein, D. (1990). "The role of banks in reducing the costs of financial distress in Japan". Journal of Financial Economics 27, 67–88.

Huddart, S. (1993). "The effect of a large shareholder on corporate value". Management Science 39 (11), 1407–1421.

Ikenberry, D., Lakonishok, J. (1993). "Corporate governance through the proxy contest: evidence and implications". Journal of Business 66 (3), 405–435.

International Corporate Governance Network (2002). Executive Remuneration—The Caucus Race? Consultative Document. London, ICGN. (www.icgn.org.)

IRRC (2000a). Takeover Defenses 2000. Investor Responsibility Research Center, Washington, D.C.

IRRC (2000b). Management Proposals on Executive Compensation Plans. IRRC Governance Service Background Report A. Investor Responsibility Research Center, Washington, D.C.

IRRC (2001). Losing Value. Investor Responsibility Research Center, Washington, D.C.

IRRC (2002). Stock Plan Dilution 2002: Overhang from Stock Plans at S&P Super 1500 Companies. Investor Responsibility Research Center, Washington, D.C.

Jacquemin, A., de Ghellinck, E. (1980). "Familial control, size and performance in the largest French firms". European Economic Review 13 (1), 81–91.

James, C. (1987). "Some evidence on the uniqueness of bank loans". Journal of Financial Economics 19 (2), 217–235.

Jarrell, G.A., Poulsen, A.B. (1987). “Shark repellents and stock prices: the effects of antitakeover amendments since 1980”. Journal of Financial Economics 19, 127–168.
Jarrell, G.A., Brickley, J.A., Netter, J.M. (1988). “The market for corporate control: the empirical evidence since 1980”. Journal of Economic Perspectives 2, 49–68.
Jeidels, O. (1905). Das Verhältnis der deutschen Grossbanken zur Industrie: mit besonderer Berücksichtigung der Eisenindustrie. Duncker & Humblot, Leipzig.
Jenkinson, T.J., Ljungqvist, A. (2001). “The role of hostile stakes in German corporate governance”. Journal of Corporate Finance 7 (4), 397–446.
Jensen, M.C. (1986). “Agency costs of free cash flow, corporate finance, and takeovers”. American Economic Review 76, 323–329.
Jensen, M.C. (1989). “The eclipse of the public corporation”. Harvard Business Review 67, 61–74.
Jensen, M.C. (2002). “Value maximization and the corporate objective function”. In: Andriof, J., Waddock, S., Rahman, S., Husted, B. (Eds.), Unfolding Stakeholder Thinking. Greenleaf Publishing, Sheffield, U.K.
Jensen, M.C., Meckling, W.H. (1976). “Theory of the firm: managerial behavior, agency costs and ownership structure”. Journal of Financial Economics 3, 305–360.
Jensen, M.C., Murphy, K.J. (1990). “Performance pay and top-management incentives”. Journal of Political Economy 98, 225–264.
Jensen, M.C., Ruback, R.S. (1983). “The market for corporate control: the scientific evidence”. Journal of Financial Economics 11, 5–50.
Jensen, M.C., Zimmerman, J.L. (1985). “Management compensation and the managerial labor market”. Journal of Accounting and Economics 7 (1–3), 3–9.
Jensen, M.C., Murphy, K.J., Wruck, E.G. (2004). “Remuneration: where we’ve been, how we got to here, what are the problems, and how to fix them”. Harvard NOM Workin Paper No. 04-28; ECGI—Finance Working Paper No. 44/2004 (Available at SSRN: http://ssrn.com/abstract=561305 or DOI: 10.2139/ssrn.561305.)
Jog, V., Riding, A. (1986). “Price effects of dual-class shares”. Financial Analysts’ Journal 42, 58–67.
John, K., Senbet, L. (1998). “Corporate governance and board effectiveness”. Journal of Banking and Finance 22 (4), 371–403.
John, K., Kedia, S. (2000). “Design of corporate governance: role of ownership structure, takeovers, bank debt and large shareholder monitoring”. Working Paper FIN-00-048, Leonard N. Stern School of Business, New York University, New York, NY.
Johnson, S.A., Ryan, H.E., Tian, Y.S. (2006). “Managerial incentives and corporate fraud: the sources of incentives matter”, EFA 2006 Zurich Meetings. (Available at SSRN: http://ssrn.com/abstract=395960.)
Johnson, S. (2000). “Corporate governance in the Asian financial crisis”. Journal of Financial Economics 58, 141–186.
Josephson, M. (1934). The Robber Barons; The Great American Capitalists, 1861–1901. Harcourt Brace and Company, New York.
Kahn, C., Winton, A. (1998). “Ownership structure, speculation, and shareholder intervention”. Journal of Finance 53, 99–129.
Kamar, E. (1998). “A regulatory competition theory of indeterminacy in corporate law”. Columbia Law Review 98, 1908–1959.
Kamerschen, D.R. (1968). “The influence of ownership and control on profit rates”. The American Economic Review 58, 432–447.
Kang, J.-K., Shivdasani, A., Yamada, T. (2000). “The effect of bank relations on investment decisions: an investigation of Japanese takeover bids”. Journal of Finance 55 (5), 2197–2218.
Kang, J.-K., Stulz, R.M. (2000). “Do banking shocks affect borrowing firm performance? An analysis of the Japanese experience”. Journal of Business 73, 1–23.
Kaplan, S.N., Strömberg, P. (2003). “Financial contracting theory meets the real world: an empirical analysis of venture capital contracts”. Review of Economic Studies 70 (2), 281–315.
Kaplan, S. (Ed.) (2000). Mergers and Productivity. The University of Chicago Press, Chicago.
Kaplan, S. (1994a). “Top executive rewards and firm performance: a comparison of Japan and the U.S.” Journal of Political Economy 102 (3), 510–546.

Kaplan, S. (1994b). "Federated's acquisition and bankruptcy: lessons and implications". Washington University Law Quarterly 72, 1103–1126.

Karpoff, J.M. (2001). "The impact of shareholder activism on target companies: a survey of empirical findings". University of Washington School of Business, University of Washington. (Available at SSRN: http://ssrn.com/abstract=885365.)

Kim, E.H., McConnell, J.J. (1977). "Corporate mergers and the co-insurance of corporate debt". Journal of Finance 32, 349–365.

King Committee (1994). The King Report on Corporate Governance. Institute of Directors of Southern Africa, Johannesburg, November 1994.

Klein, B., Crawford, R., Alchian, A.A. (1978). "Vertical integration, appropriable rents, and the competitive contracting process". Journal of Law and Economics 21 (2), 297–326.

Klein, W.A., Coffee, J.C. (2000). Business Organization and Finance. Legal and Economic Principles. Foundation Press, New York, NY.

Knoeber, C.R. (1986). "Golden parachutes, shark repellents, and hostile tender offers". American Economic Review 76, 155–167.

Kole, S.R. (1995). "Measuring managerial equity ownership: a comparison of sources of ownership data". Journal of Corporate Finance: Contracting, Governance and Organization 1, 413–435.

Kole, S.R. (1997). "The complexity of compensation contracts". Journal of Financial Economics 43, 79–104.

Kovenock, D. (1984). A Note on Takeover Bids. Purdue University.

Kraakman, R.H., Park, H., Shavell, S. (1994). "When are shareholder suits in shareholders' interests?" Georgetown Law Review, 1733–1775.

Krasa, S., Villamil, A.P. (1992). "Monitoring the monitor: an incentive structure for a financial intermediary". Journal of Economic Theory 57, 197–221.

Kreps, D.M. (1990). "Corporate culture and economic theory". In: Alt, J.E., Shepsle, K.A. (Eds.), Perspectives on Positive Political Economy. Cambridge University Press, Cambridge, pp. 90–143.

Kroszner, R. (1999). Is the Financial System Politically Independent? Perspectives on the Political Economy of Banking and Financial Regulation. George J. Stigler Center for the Study of the Economy and the State Working Papers Series, University of Chicago.

Kroszner, R.S., Strahan, P.E. (2001). "Obstacles to optimal policy: the interplay of politics and economics in shaping bank supervision and regulation reforms". In: Mishkin, F.S. (Ed.), Prudential Supervision: What Works and What Doesn't, NBER Conference Report Series. University of Chicago Press, Chicago and London.

Kroszner, R.S., Rajan, R.G. (1994). "Is the Glass-Steagall act justified? A study of the U.S. experience with universal banking before 1933". American Economic Review 84, 810–832.

Kroszner, R.S., Rajan, R.G. (1997). "Organization structure and credibility: evidence from commercial bank securities activities before the Glass-Steagall act". Journal of Monetary Economics 39, 475–516.

Kunz, R.M., Angel, J.J. (1996). "Factors affecting the value of the stock voting right: evidence from the swiss equity market". Financial Management 25, 7–20.

Kyle, A.S., Vila, J.-L. (1991). "Noise trading and takeovers". Rand Journal of Economics 22, 54–71.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A. (1998). "Law and finance". Journal of Political Economy 106, 1113–1155.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A. (1999). "Corporate ownership around the world". Journal of Finance 54, 471–517.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A. (2002). "Investor protection and corporate valuation". Journal of Finance 57, 1147–1170.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., Vishny, R.W. (1997). "Legal determinants of external finance". Journal of Finance 52, 1131–1150.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., Vishny, R.W. (2000a). "Agency problems and dividend policies around the world". Journal of Finance 55, 1–33.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., Vishny, R.W. (2000b). "Investor protection and corporate governance". Journal of Financial Economics 58, 3–27.

Larcker, D.F. (1983). "The association between performance plan adoption and corporate capital investment". Journal of Accounting and Economics 5 (1), 3–30.
Larner, R.J. (1966). "Ownership and control in the 200 largest non-financial corporations, 1929–1963". The American Ecnomic Review 16, 781–782.
Larner, R.J. (1970). Management Control and the Large Corporation. Dunellen, New York.
Lease, R.C., McConnell, J.J., Mikkelson, W.H. (1983). "The market value of control in publicly-traded corporations". Journal of Financial Economics 11, 439–471.
Lease, R.C., McConnell, J.J., Mikkelson, W.H. (1984). "The market value of differential voting rights in closely held corporations". Journal of Business 57, 443–467.
Leech, D. (1987a). "Ownership concentration and the theory of the firm: a simple-game-theoretic approach". Journal of Industrial Economics 35, 225–240.
Leech, D. (1987b). "Corporate ownership and control: a new look at the evidence of berle and means". Oxford Economic Papers, N.S. 39, 534–551.
Leech, D. (1987c). "Ownership concentration and control in large U.S. corporations in the 1930s: an analysis of the TNEC sample". Journal of Industrial Economics 35, 333–342.
Leech, D. (2003). "Incentives to corporate governance activism". In: Waterson, M. (Ed.), Competition, Monopoly and Corporate Governance, Essays in Honour of Keith Cowling, Edward Elgar, in press.
Leech, D., Leahy, J. (1991). "Ownership structure, control type classifications and the performance of large British companies". Economic Journal 101, 1418–1437.
Leland, H.E., Pyle, D.H. (1977). "Informational asymmetries, financial structure, and financial intermediation". Journal of Finance 32, 371–387.
Levin, J. (1995). Structuring Venture Capital, Private Equity, and Entrepreneurial Transactions. Little, Brown, Boston. Chapter 9.
Levitt, A. (2002). Take on the Street. Pantheon Books, New York.
Levy, H. (1983). "Economic evaluation of voting power of common stock". Journal of Finance 38, 79–93.
Lewellen, W.G. (1968). Executive Compensation in Large Industrial Corporations. Columbia University Press, New York.
Liefmann, R. (1909). Beteiligungs und Finanzierungsgesellschaften; eine Studie über den modernen Kapitalismus und das Effektenwesen (in Deutschland, den Vereinigten Staaten, England, Frankreich, Belgien und der Schweiz). G. Fischer, Jena.
Liefmann, R. (1920). Kartelle und Trusts: und die Weiterbildung der volkswirtschaftlichen Organisation. E. Moritz, Stuttgart.
Lippmann, W. (1914). Drift and Mastery; An Attempt to Diagnose the Current Unrest. H. Holt & co., New York.
Lipton, M., Rowe, P.K. (2002). "Pills, polls and professors: a reply to professor Gilson". Delaware Journal of Corporate Law 27 (1), 1–55.
Loderer, C., Peyer, U. (2002). "Board overlap, seat accumulation and share prices". European Financial Management 8, 165–192.
Loewenstein, M.J. (2000). "The conundrum of executive compensation". Wake Forest Law Review 35 (1), 1–30.
Lorsch, J., MacIver, E. (1989). Pawns or Potentates. Harvard Business School Press, Boston, MA.
Lummer, S.L., McConnell, J.J. (1989). "Further evidence on the bank lending process and the capital-market response to bank loan agreements". Journal of Financial Economics 25 (1), 99–122.
Macey, J.R., McChesney, F.S. (1985). "A theoretical analysis of corporate greenmail". Yale Law Journal 95, 13–61.
Macey, J.R. (1992). "An economic analysis of the various rationales for make shareholders the exclusive beneficiaries of corporate fiduciary duties". Stetson Law Review 21, 23–44.
Maher, M., Andersson, T. (2000). "Corporate governance: effects on firm performance and economic growth". Discussion Paper, OECD, Paris.
Main, B.G. (1999). "The rise and fall of executive share options in Britain". In: Carpenter, J., Yermack, D. (Eds.), Executive Compensation and Shareholder Value: Theory and Evidence. Kluwer Academic Press, Dordrecht, pp. 83–113.

Malmendier, U., Shanthikumar, D.M. (2004). "Are investors naive about incentives?" NBER Working Paper No. W10812. (http://ssrn.com/abstract=601114.)
Manne, H.G. (1964). "Some theoretical aspects of share voting". Columbia Law Review 64, 1427–1445.
Manne, H.G. (1965). "Mergers and the market for corporate control". Journal of Political Economy 73, 110–120.
Manning, B. (1958). "Book review: The American Stockholder, by J.A. Livingston". Yale Law Journal 67, 1477–1496.
Marris, R. (1964). The Economic Theory of Managerial Capitalism. Free Press of Glencoe, Glencoe, Illinois.
Mason, E.S. (Ed.) (1959). The Corporation in Modern Society. Harvard University Press, Cambridge, MA.
Masson, R.T. (1971). "Executive motivations, earnings, and consequent equity performance". Journal of Political Economy 79 (6), 1278–1292.
Maug, E. (1997). "Boards of directors and capital structure: alternative forms of corporate restructuring". Journal of Corporate Finance: Contracting, Governance and Organization 3, 113–139.
Maug, E. (1998). "Large shareholders as monitors: is there a trade-off between liquidity and control?" Journal of Finance 53, 65–98.
Mayer, C. (1988). "New issues in corporate finance". European Economic Review 32, 1167–1183.
McCauley, R.N., Zimmer, S.A. (1994). "Exchange rates and international differences in the cost of capital". In: Amihud, Y., Levich, R.M. (Eds.), Exchange Rates and Corporate Performance. New York University, New York, pp. 119–148.
McConnell, J.J., Servaes, H. (1990). "Additional evidence on equity ownership and corporate value". Journal of Financial Economics 27, 595–612.
Mead, E.S. (1912). Corporation Finance, 2nd edn. D. Appleton and Company, New York and London.
Mead, E.S. (1928). Corporation Finance, 6th edn. D. Appleton and Company, New York and London.
Mead, E.S. (1903). Trust Finance; A Study of the Genesis, Organization, and Management of Industrial Combinations. D. Appleton and Company, New York.
Means, G. (1930). "The diffusion of stock ownership in the U.S." Quarterly Journal of Economics 44, 561–600.
Means, G.C. (1931a). "The growth in the relative importance of the large corporation in American economic life". The American Economic Review 21, 10–42.
Means, G.C. (1931b). "The separation of ownership and control in American industry". Quarterly Journal of Economics 46, 68–100.
Megginson, W.L. (1990). "Restricted voting stock, acquisition premiums, and the market value of corporate control". Financial Review 25, 175–198.
Mehran, H. (1995). "Executive compensation structure, ownership, and firm performance". Journal of Financial Economics 38, 163–184.
Michaely, R., Womack, K.L. (1999). "Conflict of interest and the credibility of underwriter analyst recommendations". Review of Financial Studies 12, 653–686.
Mikkelson, W.H., Megan Partch, M. (1997). "The decline of takeovers and disciplinary managerial turnover". Journal of Financial Economics 44, 205–228.
Minow, N. (2000). CEO Contracts 1999: Introduction. (http://www.thecorporatelibrary.com.)
Mirrlees, J.A. (1976). "The optimal structure of incentives and authority within an organization". Bell Journal of Economics 7, 105–131.
Mirrlees, J.A. (1999). "The theory of moral hazard and unobservable behaviour: Part I". Review of Economic Studies 66, 3–21.
Mitchell, J. (1990). "Perfect equilibrium and intergenerational conflict in a model of cooperative enterprise growth". Journal of Economic Theory 51 (1), 48–76.
Mitchell, J. (2000). "Bad debts and the cleaning of banks' balance sheets: an application to economies in transition, revised". (http://www.ecare.ulb.ac.be/ecare/Janet/janet.htm.)
Moeller, S.B., Schlingemann, F.P., Stulz, R.M. (2004). "Firm size and the gains from acquisitions". Journal of Financial Economics 73, 201–228.
Moeller, S.B., Schlingemann, F.P., Stulz, R.M. (2005). "Wealth destruction on a massive scale? A study of acquiring-firm returns in the recent merger wave". Journal of Finance 60, 757–782.

Monks, A.G., Minow, N. (2001). Corporate Governance, 2nd edn. Blackwell Publishing.
Monsen, R.J., Chiu, J.S., Cooley, D.E. (1968). "The effect of separation of ownership and control on the performance of the large firm". Quarterly Journal of Economics 82, 435–451.
Morck, R., Shleifer, A., Vishny, R.W. (1988). "Management ownership and market valuation: an empirical analysis". Journal of Financial Economics 20, 293–315.
Morgan, A.G., Poulsen, A.B. (2001). "Linking pay to performance—compensation proposals in the S&P 500". Journal of Financial Economics 62 (3), 489–523.
Mukherji, W., Kim, Y.H., Walker, M.C. (1997). "The effect of stock splits on the ownership structure of firms". Journal of Corporate Finance: Contracting, Governance and Organization 3, 167–188.
Mulherin, H.-J., Poulsen, A.-B. (1998). "Proxy contests and corporate change: implications for shareholder wealth". Journal of Financial Economics 47 (3), 279–313.
Muller, H., Warneryd, K. (2001). "Inside versus outside ownership: a political theory of the firm". RAND Journal of Economics 32, 527–541.
Murphy, K. (1999). "Executive compensation". In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 3. North Holland, Amsterdam, pp. 2485–2563.
Muus, C.K. (1998). "Non-voting shares in France: an empirical analysis of the voting premium". Working Paper Series Finance & Accounting, Johann Wolfgang Goethe Universität, Frankfurt am Main.
Myers, S.C. (1977). "Determinants of corporate borrowing". Journal of Financial Economics 5, 147–175.
Myners, P. (2001). Institutional Investment in the U.K.: A Review. H.M. Treasury, London.
Narayanan, M.P. (1985). "Managerial incentives for short-term results". Journal of Finance 40 (5), 1469–1484.
Neher, D.V. (1999). "Staged financing: an agency perspective". Review of Economic Studies 66, 255–274.
Nenova, T. (2003). "The value of corporate votes and control benefits: a cross-country analysis". Journal of Financial Economics 68, 325–351.
Nicodano, G. (1998). "Corporate groups, dual-class shares and the value of voting rights". Journal of Banking and Finance 22, 1117–1137.
Nicodano, G., Sembenelli, A. (2004). "Private benefits, block transaction premiums and ownership structure". International Review of Financial Analysis 13 (2), 227–244.
Noe, T.H., Rebello, M.J. (1996). The Design of Corporate Boards: Composition, Compensation, Factions and Turnover. Georgia State University, mimeo.
Odegaard, B.A. (2002). Price Differences Between Equity Classes. Corporate Control, Foreign Ownership or Liquidity? Evidence from Norway. Norwegian School of Management, Oslo, mimeo.
Olivencia Report (1998). Comisión Especial para el Estudio de un Código Etico de los Consejos de Administración de las Sociedades, El Gobierno de las Sociedades Cotizadas, Spain. (www.ecgi.org.)
Ongena, S., Smith, D.C. (1998). "Quality and duration of banking relationships". In: Birks, D. (Ed.), Global Cash Management in Europe. MacMillan Press, pp. 225–235.
Ongena, S., Smith, D.C. (2000). "Bank relationships: a review". In: Zenios, S.A., Harker, P. (Eds.), Performance of Financial Institutions. Cambridge University Press, Cambridge, pp. 221–258.
Organisation for Economic Co-operation and Development (1999). OECD Principles of Corporate Governance, OECD, Paris.
Pagano, M., Röell, A. (1998). "The choice of stock ownership structure: agency costs, monitoring, and the decision to go public". Quarterly Journal of Economics 113, 187–225.
Pagano, M., Volpin, P.F. (2005a). "The political economy of corporate governance". American Economic Review 95 (4), 1005–1030.
Pagano, M., Volpin, P.F. (2005b). "Workers, managers, and corporate control". The Journal of Finance 60 (2), 841–868.
Pagano, M., Röell, A., Zechner, J. (2002). "The geography of equity listing: why do companies list abroad?" The Journal of Finance 57, 2651–2694.
Peng, L., Röell, A. (2006). "Executive pay, earnings manipulation and shareholder litigation". Review of Finance, in press.
Perotti, E.C., Spier, K.E. (1993). "Capital structure as a bargaining tool: the role of leverage in contract renegotiation". American Economic Review 83, 1131–1141.

Perry, T., Zenner, M. (2000). "CEO compensation in the 1990s: shareholder alignment or shareholder expropriation?" Wake Forest Law Review 35 (1), 123–152.
Perry, T., Zenner, M. (2001). "Pay for performance? Government regulation and the structure of compensation contracts". Journal of Financial Economics 62 (3), 453–488.
Peters Report (1997). Secretariat Committee on Corporate Governance, Corporate Governance in the Netherlands—Forty Recommendations. The Netherlands. (www.ecgi.org.)
Petersen, M.A. (1992). "Pension reversions and worker-stockholder wealth transfers". Quarterly Journal of Economics 107, 1033–1056.
Petersen, M.A., Rajan, R.G. (1994). "The benefits of lending relationships: evidence from small business data". Journal of Finance 49, 3–37.
Pettigrew, A., Thomas, H., Whittington, R. (Eds.) (2002). Handbook of Strategy and Management. Sage Publications, London.
Pfannschmidt, A. (1993). Personelle Verflechtungen über Aufsichtsräte: Mehrfachmandate in deutschen Unternehmen. Gabler, Wiesbaden.
Pistor, K. (2000). "Patterns of legal change: shareholder and creditor rights in transition economies". The European Business Organisation Law Review 1 (1), 59–108.
Pontiff, J., Shleifer, A., Weisbach, M.S. (1990). "Reversions of excess pension assets after takeovers". Rand Journal of Economics 21, 600–613.
Porter, M.E. (1992a). "Capital disadvantage: America's failing capital investment system". Harvard Business Review 65–82.
Porter, M.E. (1992b). "Capital choices: changing the way America invests in industry". Journal of Applied Corporate Finance 5 (2), 4–16.
Pound, J. (1988). "Proxy contests and the efficiency of shareholder oversight". Journal of Financial Economics 20 (1/2), 237–265.
Prigge, S. (1998). "A survey of German corporate governance". In: Hopt, K.J., Kanda, H., Roe, M.J., Wymeersch, E., Prigge, S. (Eds.), Comparative Corporate Governance. The State of the Art and Emerging Research. Clarendon Press, Oxford, pp. 943–1044.
Prowse, S.D. (1990). "Institutional investment patterns and corporate financial behavior in the U.S. and Japan". Journal of Financial Economics 27, 43–66.
Pujo Committee (1913). Report of the committee appointed pursuant to House resolutions 429 and 504 to investigate the concentration of control of money and credit. United States Congress House Committee on Banking and Currency, Government printing office.
Radice, H.K. (1971). "Control type, profitability and growth in large firms: an empirical study". Economic Journal 81, 547–562.
Raheja, C.G. (2005). "Determinants of board size and composition: a theory of corporate boards". Journal of Financial and Quantitative Analysis 40 (2), 283–306.
Rajan, R., Zingales, L. (2000). "The governance of the new enterprise". In: Vives, X. (Ed.), Corporate Governance. Theoretical and Empirical Perspectives. Cambridge University Press, Cambridge.
Rajan, R., Zingales, L. (2003). "The great reversals: the politics of financial development in the twentieth century". Journal of Financial Economics 69 (1), 5–50.
Ramirez, C.D. (1995). "Did J.P. Morgan's men add liquidity? Corporate investment, cash flow, and financial structure at the turn of the twentieth century". Journal of Finance 50, 661–678.
Ramirez, C.D., DeLong, J.B. (2001). "Understanding America's hesitant steps toward financial capitalism: politics, the depression, and the separation of commercial and investment banking". Public Choice 106, 93–116.
Ribstein, L.E. (2003). "International implications of Sarbanes-Oxley: raising the rent on U.S. law". Journal of Corporate Law Studies 3 (2).
Richardson, S., Tuna, I., Wu, M. (2003). "Capital market pressures and earnings management: the case of earnings restatements". University of Pennsylvania, unpublished.
Ripley, W.Z. (1927). Main Street and Wall Street. Little Brown and Company, Boston.
Roberts, J., Van den Steen, E. (2000). "Shareholder interests, human capital investments and corporate governance". Stanford University Graduate School of Business Working Paper 1631, Stanford.

Robinson, C., Rumsey, J., White, A. (1996). "Market efficiency in the valuation of corporate control: evidence from dual class equity". Revue Canadienne des Sciences de l'Administration/Canadian Journal of Administrative Sciences 13, 251–263.

Roe, M.J. (1990). "Political and legal restraints on ownership and control of public companies". Journal of Financial Economics 27, 7–41.

Roe, M.J. (1991). "A political theory of American corporate finance". Columbia Law Review 91, 10–67.

Roe, M.J. (1994). Strong Managers, Weak Owners: The Political Roots of American Corporate Finance. Princeton University Press, Princeton, N.J.

Roe, M.J. (1996). "From antitrust to corporation governance? The corporation and the law: 1959–1994". In: Kaysen, C. (Ed.), The American Corporation Today. Oxford University Press, New York, pp. 102–127.

Roe, M.J. (2002). "Corporate law's limits". The Journal of Legal Studies 31, 233–271.

Roe, M. (2005). "Regulatory competition in making corporate law in the United States—and its limits". Oxford Review of Economic Policy 21 (2), 232–242.

Roll, R. (1986). "The Hubris hypothesis of corporate takeovers". Journal of Business 59, 197–216.

Romano, R. (1991). "The shareholder suit: litigation without foundation?" Journal of Law, Economics and Organization 7, 55–87.

Romano, R. (1993). The Genius of American Corporate Law. AEI Press, Washington, D.C.

Romano, R. (1996). "Corporate law and corporate governance". Industrial and Corporate Change 5 (2), 277–339.

Romano, R. (1998). "Empowering investors: a market approach to securities regulation". Yale Law Journal 107, 2359–2430.

Romano, R. (2001). "Less is more: making institutional investor activism a valuable mechanism of corporate governance". Yale Journal on Regulation 18, 174–251.

Romano, R. (2005a). "The Sarbanes-Oxley act and the making of quack corporate governance". Yale Law Journal 114 (7), 1521–1611.

Romano, R. (2005b). "Is regulatory competition a problem or irrelevant for corporate governance?" Oxford Review of Economic 21 (2), 212–231.

Rosen, S. (1992). "Contracts and the Market for executives". In: Werin, L., Wijkander, H. (Eds.), Contract Economics. Blackwell, Cambridge, MA and Oxford, pp. 181–211.

Rosenbaum, V. (2000). Corporate Takeover Defenses 2000. Investor Responsibility Research Center, Washington, D.C.

Rosenstein, S., Wyatt, J.G. (1990). "Outside directors, board independence, and shareholder wealth". Journal of Financial Economics 26, 175–191.

Rossi, S., Volpin, P. (2004). "Cross-country determinants of mergers and acquisitions". Journal of Financial Economics 74 (2), 277–304.

Rostow, E.V. (1959). "To whom and for what ends are corporate managements responsible?" In: Mason, E.S. (Ed.), The Corporation in Modern Society. Harvard University Press, Cambridge, MA.

Rothberg, B.G., Lilien, S.B. (2005). Mutual funds and proxy voting: new evidence on corporate governance. Baruch College Working Paper, City University of New York, New York. (http://ssrn.com/abstract=669161.)

Rydqvist, K. (1992). "Dual-class shares: a review". Oxford Review of Economic Policy 8, 45–57.

Rydqvist, K. (1996). "Takeover bids and the relative prices of shares that differ in their voting rights". Journal of Banking and Finance 20, 1407–1425.

Sabel, C. (1996). "Ungoverned production: an American view of the novel universalism of Japanese production methods and their awkward fit with current forms of corporate governance". Worldspeaker, Electronic Magazine, Summer.

Sahlman, W.A. (1990). "The structure and governance of venture-capital organizations". Journal of Financial Economics 27, 473–521.

Sarin, A., Shastri, K.A., Shastri, K. (1999). "Ownership structure and stock market liquidity". Santa Clara University, California, mimeo.

Scharfstein, D. (1988). "The disciplinary role of takeovers". Review of Economic Studies 55, 185–199.

Schmidt, K. (1996). "The costs and benefits of privatization—an incomplete contracts approach". Journal of Law, Economics and Organization 12, 1–24.
Schnitzer, M. (1995). " 'Breach of trust' in takeovers and the optimal corporate charter". Journal of Industrial Economics 43, 229–259.
Schumpeter, J.A. (1939). Business Cycles: A Theoretical, Historical, and Statistical Analysis of the Capitalist Process. McGraw-Hill, New York, London.
Schumpeter, J.A. (1934). The Theory of Economic Development: An Inquiry into Profits, Capital, Credit, Interest, and the Business Cycle. Harvard University Press, H. Milford, Cambridge, MA, London.
Schwert, G.W. (2000). "Hostility in takeovers: in the eyes of the beholder?" Journal of Finance 55, 2599–2640.
Seligman, J. (1986). "Equal protection in shareholder voting rights: the one-share-one-vote controversy". George Washington Law Review 54, 687.
Shapiro, C., Willig, , R.D. (1990). "Economic rationales for the scope of privatization". In: Suleiman, E.N., Waterbury, J. (Eds.), The Political Economy of Public Sector Reform and Privatization. Westview Press, London, pp. 55–87.
Sherman, S. (2002). "Enron: uncovering the uncovered story". Columbia Journalism Review 40 (6), 22–28.
Shinn, J. (2001). "Private profit or public purpose? Shallow convergence on the shareholder model". Ph.D. Dissertation, Princeton University.
Shinn, J., Gourevitch, P. (2002). How Shareholder Reforms Can Pay Foreign Policy Dividends. Council on Foreign Relations, New York.
Shleifer, A., Summers, L.H. (1988). "Breach of trust in hostile takeovers". In: Auerbach, A.J. (Ed.), Corporate Takeovers: Causes And Consequences. National Bureau of Economic Research Project Report series. University of Chicago Press, Chicago and London.
Shleifer, A., Vishny, R.W. (1986). "Large shareholders and corporate control". Journal of Political Economy 94, 461–488.
Shleifer, A., Vishny, R.W. (1988). "Value maximization and the acquisition process". Journal of Economic Perspectives 2, 7–20.
Shleifer, A., Vishny, R.W. (1989). "Equilibrium short horizons of investors and firms". American Economic Review 80 (2), 148–153.
Shleifer, A., Vishny, R.W. (1997a). "A survey of corporate governance". Journal of Finance 52, 737–783.
Shleifer, A., Vishny, R.W. (1997b). "The takeover wave of the 1980s". In: Chew, D.H. (Ed.), Studies in International Corporate Finance and Governance Systems. Oxford University Press, New York, pp. 98–105.
Shockley, R., Thakor, A.V. (1997). "Bank loan commitments: data, theory, and tests". Journal of Money, Credit and Banking 29 (4), 517–534.
Short, H. (1994). "Ownership, control, financial structure and the performance of firms". Journal of Economic Surveys 8, 203–249.
Short, H., Keasey, K. (1999). "Managerial ownership and the performance of firms: evidence from the U.K." Journal of Corporate Finance: Contracting, Governance and Organization 5, 79–101.
Singapore (1998). Stock Exchange of Singapore, Listing Manual (as amended) and Best Practices Guide, Stock Exchange of Singapore: Singapore, 1999. (http://www.combinet.org/governance/finalver/listof.htm.)
Skeel, D. (2005). Icarus in the Boardroom. Oxford University Press, Oxford.
Smith, B.F., Amoako-Adu, B. (1995). "Relative prices of dual class shares". Journal of Financial and Quantitative Analysis 30, 223–239.
Spier, K.E. (1992). "Incomplete contracts and signalling". Rand Journal of Economics 23, 432–443.
Stapledon, G. (1996). Institutional Shareholders and Corporate Governance. Clarendon Press, Oxford.
Steer, P.S., Cable, J.R. (1978). "Internal organization and profit: an empirical analysis of large U.K. companies". Journal of Industrial Economics 27, 13–30.
Stein, J.C. (1988). "Takeover threats and managerial myopia". Journal of Political Economy 96, 61–80.
Stein, J.C. (1989). "Efficient capital markets, inefficient firms: a model of myopic corporate behavior". Quarterly Journal of Economics 104, 655–669.

Stein, J.C. (2003). "Agency, information and corporate investment". In: Handbook of the Economics of Finance Volume 1A. Elsevier, North-Holland.
Stock Exchange of Thailand (SET) (1998). The SET Code of Best Practice for Directors of Listed Companies, SET: Bangkok. (http://www.combinet.org/governance/finalver/listof.htm.)
Stoll, H.R., Whaley, R.E. (1983). "Transaction costs and the small firm effect". Journal of Financial Economics 12, 57–79.
Streeck, W., Kluge, N. (Eds.) (1999). Mitbestimmung in Deutschland. Tradition und Effizienz. Campus Verlag, Frankfurt/Main, New York.
Stulz, R.M. (1988). "Managerial control of voting rights: financing policies and the market for corporate control". Journal of Financial Economics 20, 25–54.
Stulz, R.M., Williamson, R. (2003). "Culture, openness and finance". Journal of Financial Economics 70, 313–349.
Svejnar, J. (1981). "Relative wage effects of unions, dictatorship and codetermination: econometric evidence from Germany". Review of Economics and Statistics 63, 188–197.
Svejnar, J. (1982). "Codetermination and productivity: evidence from the Federal Republic of Germany". In: Jones, D., Svejnar, J. (Eds.), Participatory and Self-Managed Firms. Heath, Lexington, MA.
Tabarrok, A. (1998). "The separation of commercial and investment banking: the Morgans vs. the Rockefellers". Quarterly Journal of Austrian Economics 1, 1–18.
Tallman, E.W. (1991). "The house of Morgan: an American banking dynasty and the rise of modern finance: review article". Federal Reserve Bank of Atlanta Economic Review 76, 28–32.
Temporary National Economic Committee (TNEC) (1940). The Distribution of Ownership in the 200 Largest Nonfinancial Corporations. U.S. Government Printing Office, Washington, D.C.
Thomas, R.S., Martin, K.J. (2000). "The determinants of shareholder voting on stock option plans". Wake Forest Law Review 35 (1), 31–73.
Tilly, R.H. (1989). "Banking institutions in historical and comparative perspective: Germany, Great Britain and the U.S. in the nineteenth and early twentieth century". Journal of Institutional and Theoretical Economics 145, 189–209.
Tinic, S.M. (1972). "The economics of liquidity services". Quarterly Journal of Economics 86, 79–93.
Tirole, J. (1986). "Hierarchies and bureaucracies". Journal of Law, Economics and Organization 2, 181–214.
Tirole, J. (2001). "Corporate governance". Econometrica 69, 1–35.
Twain, M., Warner, C.D. (1873). The Gilded Age. A Tale of Today. American Pub. Co., F.G. Gilman, Hartford, Chicago, IL.
Vafeas, N. (1999). "Board meeting frequency and firm performance". Journal of Financial Economics 53 (1), 113–142.
Van-Nuys, K. (1993). "Corporate governance through the proxy process: evidence from the 1989 Honeywell proxy solicitation". Journal of Financial Economics 34 (1), 101–132.
Veblen, T. (1923). Absentee Ownership And Business Enterprise In Recent Times; The Case Of America. B.W. Huebsch Inc., New York.
Viénot Report (1995). Conseil National du Patronat Français ("CNPF") & Association Française des Entreprises Privéss ("AFEP") (France), The Boards of Directors of Listed Companies in France, Viénot I. (www.ecgi.org.)
Warga, A., Welch, I. (1993). "Bondholder losses in leveraged buyouts". Review of Financial Studies 6, 959–982.
Warshow, H.T. (1924). "The distribution of corporate ownership in the U.S." Quarterly Journal of Economics 39, 15–38.
Warther, V.A. (1998). "Board effectiveness and board dissent: a model of the board's relationship to management and shareholders". Journal of Corporate Finance: Contracting, Governance and Organization 4, 53–70.
Weiner, J.L. (1964). "The Berle-Means dialogue on the concept of the corporation". Columbia Law Review 64, 1459–1467.
Weisbach, M.S. (1988). "Outside directors and CEO turnover". Journal of Financial Economics 20, 431–460.

Weston, J.F., Siu, J.A., Johnson, B.A. (2001). Takeovers, Restructuring & Corporate Governance. Prentice Hall, Upper Saddle River, NJ.

Williamson, O.E. (1971). "The vertical integration of production: market failure considerations". American Economic Review 61, 112.

Williamson, O.E. (1975). Markets and Hierarchies: Analysis and Antitrust Implications. The Free Press, New York.

Williamson, O.E. (1979). "Transaction cost economics: the governance of contractual relations". Journal of Law and Economics 22, 233–261.

Williamson, O.E. (1984). "Corporate governance". Yale Law Journal 93, 1197–1230.

Williamson, O.E. (1985a). The Economic Institutions of Capitalism. Free Press, New York.

Williamson, O.E. (1985b). "Employee ownership and internal governance: a perspective". Journal of Economic Behavior and Organization 6, 243–245.

Williamson, O.E. (1988). "Breach of trust in hostile takeovers: comment". In: Auerbach, A. (Ed.), Corporate Takeovers: Causes and Consequences. Chicago University Press, Chicago, pp. 61–67.

Wolfenzon, D. (1999). A Theory of Pyramidal Ownership. Leonard N. Stern School of Business, New York, unpublished.

Womack, J.P., Jones, D.T., Roos, D. (1991). The Machine that Changed the World: How Japan's Secret Weapon in the Global Auto Wars Will Revolutionize Western Industry. HarperPerennial, New York, NY.

Wormser, I.M. (1931). Frankenstein, Incorporated. Whittlesey House McGraw-Hill Book Company Inc., New York, London.

Wymeersch, E. (1998). "A status report on corporate governance rules and practices in some continental European states". In: Hopt, K.J., Kanda, H., Roe, M.J., Wymeersch, E., Prigge, S. (Eds.), Comparative Corporate Governance. The State of the Art and Emerging Research. Clarendon Press, Oxford, pp. 1046–1199.

Wymeersch, E. (2003). "Do we need a law on groups of companies?" In: Hopt, K.J., Wymeersch, E. (Eds.), Capital Markets and Company Law. Oxford University Press, Oxford.

Yermack, D. (1997). "Good timing: CEO stock option awards and company news announcements". Journal of Finance 52, 449–476.

Yilmaz, B. (2000). "Strategic voting and proxy contests". Rodney L. White Center for Financial Research, Working Paper no. 05-00, Wharton School, University of Pennsylvania, Philadelphia, PA.

Zender, J.F. (1991). "Optimal financial instruments". Journal of Finance 46, 1645–1663.

Zingales, L. (1994). "The value of the voting right: a study of the Milan stock exchange experience". Review of Financial Studies 7, 125–148.

Zingales, L. (1995). "What determines the value of corporate votes?" Quarterly Journal of Economics 110, 1047–1073.

Zingales, L. (1998). "Corporate governance". In: Newman, P. (Ed.), The New Palgrave Dictionary of Economics and the Law. Macmillan, New York, NY.

Zwiebel, J. (1995). "Block investment and partial benefits of corporate control". Review of Economic Studies 62, 161–185.

Chapter 13

EMPIRICAL STUDIES OF CORPORATE LAW

SANJAI BHAGAT[*]
School of Business, University of Colorado at Boulder

ROBERTA ROMANO[**]
School of Law, Yale University

Contents

[*] Professor of Finance. University of Colorado at Boulder.
[**] Oscar M. Ruebbausen Professor of Law, Yale University.

Handbook of Law and Economics, Volume 2
Edited by A. Mitchell Polinsky and Steven Shavell

DOI: 10.1016/S1574-0730(07)02013-0

Abstract

This chapter reviews the empirical literature, especially the event study literature, as it relates to corporate and securities law. Event studies are among the most successful uses of econometrics in policy analysis. By providing an anchor for measuring the impact of events on investor wealth, the methodology offers a fruitful means for evaluating the welfare implications of private and government actions. This chapter begins by briefly reviewing the event study methodology and its strengths and limitations for policy analysis. It then discusses one of the limitations of more conventional empirical work (cross-sectional analysis), the problem presented by the fact that the characteristics of firms that are studied in relation to each other (such as ownership and mechanisms of corporate governance) or to firm performance are not exogeneous but self-selected by firms. Thereafter it reviews in detail how event studies have been used to evaluate the wealth effects of corporate litigation. Subsequently, we focus on the methodology's application to corporate law and corporate governance issues, supplemented with discussion of other relevant empirical work as well. Event studies are emphasized because they have played an important role in the making of corporate law and in applied corporate finance and corporate law scholarship. The reason for this input is twofold. First, there is a match between the methodology and subject matter: the goal of corporate law is to increase shareholder wealth and event studies provide a metric for measurement of the impact upon stock prices of policy decisions. Second, because the participants in corporate law debates share the objective of corporate law, to adopt policies that enhance shareholder wealth, their disagreements are over the means to achieve that end. A further reason for emphasizing event study data is that they avoid the endogeneity concerns that can limit the results of other modes of empirical research in this area.

Keywords

Corporate law, corporate litigation, corporate governance, event study

JEL classification: G34, K20, K22, K40, K44

1. Introduction

This chapter reviews the empirical literature, especially the event study literature, as it relates to corporate and securities law. Event studies are among the most successful uses of econometrics in policy analysis. The methodology, which studies the movement of stock prices due to specific events (unexpected actions by managers or policy-makers that are expected to affect firm values) was originally developed to test the hypothesis that the stock market was efficient—that publicly available information is impounded immediately into stock prices such that an investor cannot earn abnormal profits by trading on the information after its release. As evidence accumulated that the stock market was efficient, the methodology came to be used instead to value the event under study. It is through this latter usage that event studies have influenced policy analysis, particularly in corporate and securities law. This is no doubt because there is a natural fit between the methodology and those fields of law: the benchmark for evaluating the benefit of corporate and securities laws is whether they improve investor welfare, and this can be ascertained by what event studies measure, whether stock prices have been positively affected.

The event study methodology is well-accepted and extensively used in finance. Event study results have been used in several hundred scholarly articles in leading academic finance journals to analyze corporate finance issues, such as stock repurchases and stock splits and the relation between stock prices and accounting information, by examining the impact of earnings releases. Because the event study technique may be less familiar to non-financial economists than other techniques of empirical analysis, this chapter draws on our earlier work, Bhagat and Romano (2002a, 2002b) to begin by briefly reviewing the event study methodology and its strengths and limitations for policy analysis. It then highlights a principal limitation of other modes of empirical research involving corporate law, the concerns implicated by an endogeneity problem, that firms' ownership and governance characteristics are not exogeneously given but are chosen by managers and investors. Thereafter we review in detail how event studies have been used to evaluate the wealth effects in corporate litigation, corporate law and corporate governance, integrating into the discussion, where relevant, research findings using other empirical approaches. The empirical literature relevant to issues in corporate and securities law is vast; the fact that the event study methodology is well-suited to evaluating the policy objectives of legal regimes undoubtedly helps explain its scope. As a consequence, the chapter is unavoidably selective in coverage and does not discuss many important topics and individual contributions to the field.

2. A guide to event studies

The price of a stock reflects the time- and risk-discounted present value of all future cash flows that are expected to accrue to the holder of that stock. According to the semi-strong version of the efficient market hypothesis, all publicly-available information is

reflected completely and in an unbiased manner in the price of the stock, such that it is not possible to earn economic profits on the basis of this information.[1] Therefore, only an unanticipated event can change the price of a stock. This change should equal the expected changes in the future cash flows of the firm or the riskiness of these cash flows. Thus, an event is said to have an impact on the financial performance of a firm if it produces an abnormal movement in the price of the stock. Broad stock market movements are usually subtracted from the stock's price movement in estimating the abnormal return. Event studies apply conventional econometric techniques to measure the effect of specific events, such as actions by firms, legislatures, and government agencies, on the stock price of affected firms. Their advantage for policy analysis is that they provide an anchor for determining value, which eliminates reliance on ad hoc judgments about the impact of specific events or policies on stock prices.

2.1. Mechanics of event studies

An event study has four component parts: defining the event and announcement day(s); measuring the stock's return during the announcement period; estimating the expected return of the stock during this announcement period in the absence of the announcement; and computing the abnormal return (actual return minus expected return) and measuring its statistical and economic significance.

In order to conduct an event study, the researcher first defines the event under investigation. Events are usually announcements of various corporate, legal, or regulatory action or proposed action. Examples of events that have been studied are: takeovers, equity offerings, change in state of incorporation, adoption of antitakeover provisions, filing of lawsuits against corporations, deaths of corporate executives, and product recalls. After defining the event, the researcher searches for the first public announcement

[1] The efficient market hypothesis has been subjected to extensive empirical testing; perhaps the most intensive and extensive testing of any hypothesis in all of the social sciences. Most tests find evidence consistent with the efficient market hypothesis. Some studies find that the stock price responds within minutes of a corporate announcement such as a stock offering (see Barclay and Litzenberger, 1988). Most finance scholars hold the view that the stock market in the U.S. is semi-strong form efficient (Welch, 2000). But controversy regarding the efficient market hypothesis lingers. This controversy is based on issues regarding the definition and measurement of risk, and the relationship between risk and return. There is, however, agreement that these issues do not invalidate the event study methodology; see Fama (1990); and Brown and Warner (1985). Some legal scholars consider the stock market to be inefficient (see, e.g., Stout, 2005). But careful scrutiny of the efficient market anomalies have raised concerns about the asset pricing models used to construct the expected returns rather than the efficiency of the market (see Schwert, 2003). It should further be noted that finance theory does not depend on whether the average investor is rational (a criticism directed by users of the behavioral finance literature, e.g., Stout, 2005); it depends, as one finance scholar puts it, on the existence of "sharks," sophisticated investors who seek to profit from arbitraging pricing anomalies (Ross, 2005). There are a few fascinating examples in which arbitrage is ineffective at eliminating pricing differentials for a period of time (e.g., Lamont and Thaler, 2003), but these micro examples of violations of the law of one price are not very important for the question of market efficiency, occurring as they do, in isolated examples of individual stocks (Ross, 2005), and not always offering an exploitable arbitrage opportunity (e.g., Lamont and Thaler, 2003).

of the event. Identification of the first public announcement of the event is critical since, under the semi-strong form of the efficient market hypothesis, the impact of the event on the value of the firm would occur on the announcement date. Historically, the *Wall Street Journal Index* has been a popular source for announcement dates. More recently, computer accessible databases such as *Lexis-Nexis* and the *Thompson Financial Securities Data* are being increasingly used.

Conceptually, the announcement date is straightforward: It is the "day" the public is first informed of the event.[2] However, identification of this date can sometimes be nontrivial. Consider the announcement of a tender offer. It is possible and probable that news of the tender offer may have leaked to some market participants prior to the first public announcement. If such is the case then some impact of the tender offer on the firm's share price would occur prior to the public announcement. Some researchers have attempted to address this issue by considering the period several weeks (or months) through the announcement day as the announcement period. However, this obvious solution has two problems, one conceptual and the other technical. Conceptually, it is unclear if the leakage occurs over a few days, weeks, or months. Technically, as we increase the length of the announcement period, the noise-to-signal ratio increases, and it becomes increasingly difficult to measure the impact of the tender offer on share price with precision; we will discuss this later in the chapter. Aside from news leakage issues, at the time the tender offer is announced there is uncertainty over whether it will be successful, and if successful, over the terms of the final offer. Sometimes the final resolution may not be known for months or even years.

Finally, some events may have several distinct event dates. For example, the enactment of a statute involves many different events, each of which may provide new information to investors regarding the likelihood of passage: when a bill is introduced, when a committee holds hearings on the bill, when one legislative chamber votes on the bill, when a conference committee approves a final bill, and when the executive signs the bill (if there is uncertainty over whether or not the bill will be vetoed). In this context, rather than treat the entire interval from bill introduction to executive signature as the event and run into the problems discussed above, the researcher can adapt the methodology to permit each event date to be identified separately; however, in doing so the researcher's bias and priors on what is a significant or relevant event enters the analysis.

After defining the event and announcement period, stock returns are measured for this period. If daily data are being used, this is straightforward: the return is measured using closing prices. Often there is uncertainty if the announcement is made before or

[2] Currently, most event studies consider daily returns, hence the announcement period is typically a day. However, historically, some event studies have considered monthly returns—where the announcement need only be identified for a particular month; see the classic study by Fama et al. (1969). More recently, announcements have been identified to the nearest minute, and returns have been computed over minute and trade intervals such that the event study is conducted using intra-day data; see Barclay and Litzenberger (1988).

after the close of trade on the exchange. To address this, the returns from the next day are often included.

Calculation of the third component is more complicated. While it is straightforward to measure the actual return for the announcement period, determination of the impact of the event itself on the share price is less so. To measure this impact, the *expected return* must be subtracted from the actual announcement period return. This expected return is the return that would have accrued to the shareholders in the absence of this or any other unusual event. The finance literature has considered several models of expected returns. These models can broadly be classified as statistical models or economic models:

Statistical models

The constant expected returns model:

$$R_{it} = \mu_i + \varepsilon_{it}, \tag{1}$$

where, R_{it} is the return for stock i over time period t, μ_i is the expected return for stock i, and ε_{it} is the usual statistical error term.

The market model:

$$R_{it} = a_i + b_i * R_{\mathrm{m}t} + \varepsilon_{it}, \tag{2}$$

where, a_i and b_i are firm-specific parameters, and $R_{\mathrm{m}t}$ is the market return for the period t.

Economic models:

Capital Asset Pricing Model (CAPM):

$$R_{it} = R_{\mathrm{f}} + \beta_i * (R_{\mathrm{m}t} - R_{\mathrm{f}}) + \varepsilon_{it}, \tag{3}$$

where, R_{f} is the riskfree rate and β_i is the beta or systematic risk of stock i.

Arbitrage Pricing Theory:

$$R_{it} = \delta_0 + \delta_{i1} F_{1t} + \delta_{i2} F_{2t} + \cdots + \delta_{in} F_{nt} + \varepsilon_{it}, \tag{4}$$

where, $F_1, F_2, \ldots, F_n$ are the returns on the n factors that generate returns, and δ are the factor loadings.

The statistical models are simple models of price formation that are not grounded in a specific economic theory. The economic models are derived from specific economic theories of asset price formation. One can think of the economic models as placing certain restrictions on the statistical models (that is, on the slopes and intercepts being estimated).

Since several studies have found evidence inconsistent with the economic models, in particular CAPM, the use of such restrictions is not appropriate. Hence, most researchers have begun to rely on the statistical models to estimate the expected returns during the announcement period. For estimation of the market model, researchers most commonly use for the market portfolio, all of the stocks in the University of Chicago Center for Research in Securities Prices (CRSP) data base, the best source for stock

return data; if all of the firms under study are small, however, using the CRSP portfolio or an index such as the S&P 500, whose average firm size is large, for the market adjustment, may produce biased estimates of the sample firms' abnormal return (see, e.g., Karpoff and Malatesta, 1995). The statistical models are usually estimated using between 100 and 200 daily returns in the period preceding the announcement period. The unexpected announcement period return, also known as the *abnormal return*, is computed as the actual return minus the estimated expected return. This abnormal return is the estimated impact of the event on the share value.

The fourth and final step is to compute the statistical significance of this abnormal return. The standard error of the residuals from the estimated statistical model can be used as an estimate of the standard error for the announcement period abnormal return. However, since individual stock returns are quite volatile, this standard error can be quite high relative to the abnormal return. Event studies usually consider a sample of firms that have made or been the subject of the same type of announcement; each firm's announcement typically has been made on a different calendar day. Another benefit of this approach is that it increases the likelihood that no other information besides the event under study will be valued, since any additional unexpected information disclosed on one firm's announcement date will wash out with that on other firms' announcement days.

The abnormal returns of this sample of firms is averaged to obtain the *average abnormal return*. This average abnormal return is the estimated impact of the event on the share value. Next, the residuals from the estimated statistical model for these firms are averaged in *event time*. Usually the announcement day is defined as *event day* 0. t days before (after) the announcement day is defined as *event day* $-t$ ($eventday + t$). Finally, the standard error of these averaged residuals is used as an estimate of the standard error of the average abnormal return. Under the null hypothesis that the event under study has no impact on firm value, the expected average abnormal return is zero. Additionally, assuming that the announcement period returns for the sample firms are independently and identically distributed, then by the Central Limit Theorem the average abnormal return is normally distributed with mean zero.

The above estimate of the standard error of the average abnormal return would be appropriate if the announcement period abnormal return had the same variance as the estimation period residuals. However, substantial evidence in the finance literature suggests that stock returns in the announcement period are typically more volatile. Brown and Warner (1985) have suggested the use of cross-sectional test statistics when there is an increase in return variance during the announcement period. The standard error of the announcement period returns for the sample firms is used as an estimate of the standard error of the average abnormal return. Non-parametric tests, such as the Fisher sign test and the Wilcoxon signed rank test, are also conducted on the announcement period returns; the usual null hypothesis is that the median announcement period return is zero.

2.2. *Statistical power of event studies*

If an event changes firm value by a specific amount, say, 1 percent, can the event study technique detect it with some statistical precision? Equally important, from a statistical, financial and legal viewpoint: If an event has no impact on firm value, that is, the announcement period abnormal return is zero, can the event study technique provide this inference with some statistical precision? These questions can be addressed by considering the statistical power of event studies.

The power of a test statistic is considered in the context of a null hypothesis and an alternate hypothesis. (Hopefully, the alternate hypothesis would be economically meaningful.) In the context of event studies, the usual null hypothesis is that the event has no impact on firm value. An interesting alternate hypothesis could be that the event increases firm value by 1 percent. Under the assumption that the alternate hypothesis is true, the power of the event study in this context is the probability of observing a statistically significant test statistic. Brown and Warner (1985) and MacKinlay (1997) have studied the power of test statistics typically used in event studies. These authors show that the power of the event study technique improves as the number of firms in the sample increase, as the number of days in the announcement window decrease, and as the alternative of a larger abnormal return is considered against the null hypothesis of zero abnormal return.

The following numerical examples from MacKinlay (1997, Table 2) illustrate the power of the event test methodology, and how the power can be enhanced.

For a one day announcement window, a sample size of 25 firms, and a two-sided test with a 5 percent significance level, the probabilities of detecting an abnormal return of 0.5 percent, 1.0 percent and 2.0 percent, are 24 percent, 71 percent and 100 percent, respectively.

- If the sample size were increased to 50 firms, the probabilities of detecting an abnormal return of 0.5 percent, 1.0 percent and 2.0 percent, are 42 percent, 94 percent and 100 percent, respectively.
- If the sample size were increased to 100 firms, the probabilities of detecting an abnormal return of 0.5 percent, 1.0 percent, and 2.0 percent, are 71 percent, 100 percent and 100 percent, respectively.
- For a two day announcement window (or equivalently, doubling of the standard deviation of the event day abnormal return), and a sample size of 25 firms, the probabilities of detecting an abnormal return of 0.5 percent, 1.0 percent and 2.0 percent, are 10 percent, 24 percent and 71 percent, respectively.
- For this two day announcement window and a sample size of 50 firms, the probabilities of detecting an abnormal return of 0.5 percent, 1.0 percent and 2.0 percent, are 14 percent, 42 percent and 94 percent, respectively.
- For this two day announcement window and a sample size of 100 firms, the probabilities of detecting an abnormal return of 0.5 percent, 1.0 percent and 2.0 percent, are 24 percent, 71 percent and 100 percent, respectively.

The above findings suggest that the power of the event study diminishes as the sample size decreases. An important question is can an event study be conducted with just one firm, that is, is a sample size of one acceptable? This question is especially relevant in court cases or regulatory injunctions involving only one firm. Conceptually, a sample of one is a rather small sample but this by itself does not invalidate the event study methodology. However, the statistical power with a sample of one is likely to be quite low. First, the variability of (abnormal) returns of a portfolio with just one stock in it is significantly higher than a portfolio with even a few, say five, stocks in it. Any standard finance or investment textbook will have a graph depicting the sharp drop in variance of portfolio returns as the number of stocks in the portfolio increases from one, to five, to ten; after about fifty stocks in the portfolio the decrease in variance is quite small. Second, it is plausible that the announcement period return of an announcing firm will be affected by other information unrelated to the event under study. If a sample of one is considered, it is quite difficult to determine the separate effects on firm value of the announcement and of the unrelated information item(s). If the sample has several firms, then the effect on firm value of such unrelated information is likely to cancel out. As the sample size increases the effect on firm value of such unrelated information (goes to zero) becomes less and less significant.

The above findings also suggest that the power of the event study methodology diminishes substantially as the event period is increased from one to just two days. During the past decade an increasing number of finance studies have considered abnormal returns for long-horizon windows of several *years*. Such studies have considered abnormal returns over *twelve to sixty months* after the announcements of various corporate events like mergers, share repurchases, initial public and seasoned equity offerings, spin-offs, stock splits and dividends. Examples of such studies include Ikenberry, Lakonishok, and Vermaelen (1995), Loughran and Ritter (1995), Brav and Gompers (1997), McConnell, Ozbilgin, and Wahal (2001), Desai and Jain (1999).

There are two reasons for studying the long-horizon window of several years after an announcement. First, the market may be unable to fully understand and incorporate the impact of the announcement on the company's value. Over time the market gets the opportunity to fully understand and incorporate the impact of the announcement on the company's value. Under this explanation, no new information related to the first announcement is released in this post-announcement period; hence this reason presumes a semistrong form *inefficient* market. Second, new information pertinent to the initial announcement may become known to the market participants in the months or years subsequent to the announcement. For example, the initial announcement could be a takeover offer announcement. Before the offer is finalized and completed several events could occur that might change the likelihood of the success of the initial offer. Examples of such events include the arrival of a second bidder, litigation by target management, and regulatory objections (see Bhagat et al., 2005). In this scenario, one way to estimate the full impact of the initial event would be to consider the period from the initial announcement through final resolution—a period that could extend several years in some cases.

Kothari and Warner (1997), Barber and Lyon (1997), and Lyon, Barber, and Tsai (1999) have raised serious concerns about the specification and power of the event study methodology when long-horizon windows of several years are considered. Kothari and Warner find that the event study test statistics used in the above-mentioned studies are generally misspecified in the sense that they reject the null hypothesis of normal performance when there is no abnormal performance too frequently given the significance level. Lyon, Barber, and Tsai (1999) ways to construct properly specified test statistics. However, these authors caution that while these test-statistics appear to be well-specified for random samples, they are not well-specified for non-random samples. Given that tests of most interesting finance and legal hypotheses are likely to lead to the construction of non-random samples, the concern with the misspecification of the long-run test statistics remains. Finally, Lyon, Barber, and Tsai (1999) document the power of the long-horizon test-statistic to detect abnormal performance when it is actually present. Using state-of-the-art techniques, for a twelve-month buy-and-hold abnormal return, a sample size of 200 firms, and a one-sided test with a 5 percent significance level, the probabilities of detecting an abnormal return of 5 percent, 10 percent and 20 percent, are 20 percent, 55 percent and 100 percent, respectively. As the horizon increases beyond twelve months, and the sample size decreases, the power of the technique would further diminish. For these reasons, these authors (p. 198) conclude that "the analysis of long-run abnormal returns is treacherous." The problems with the specification of the methodology of long-horizon event studies, identified by these authors, have still not been resolved (see the literature review of Kothari and Warner, 2007, updating the earlier papers).

2.3. Cross-sectional determinants of the stock market's reaction

Some researchers have sought to provide insight into the cross-sectional determinants of the stock market's reaction to the announcement of an event by examining the relation between the size of the abnormal return (AR) identified in an event study and characteristics specific to the event observations, that is, cross-sectional differences in the firms in the study. This approach can be used, for instance, where there are multiple hypotheses for the source of a wealth effect. The AR is the dependent variable in an ordinary least squares regression on the firm characteristics of interest:

$$\mathrm{AR}_j = \delta_0 + \delta_1 x_{1j} + \cdots + \delta_M x_{Mj} + \eta_j, \tag{5}$$

where AR_j is the jth abnormal return observation, x_{mj}, $m = 1, \ldots, M$, are M characteristics for the jth observation and η_j is the zero mean disturbance term that is uncorrelated with the x's. δ_m, $m = 0, \ldots, M$ are the regression coefficients.

This approach has been used in a variety of contexts. We note here an illustration from the methodology's application to assessing the wealth effects of corporate litigation discussed in Section 4.1 below. Bhagat, Brickley, and Coles (1994) provide an example of its use in determining the source of the significant negative wealth effects experienced by corporate defendants. They find that the negative abnormal returns from litigation

are significantly related to variables proxying for the defendant's proximity to financial distress.

An interpretational concern involving cross-sectional models is whether the abnormal return is related to the firm characteristics not only through the wealth effect identified in the event study but also through investors' anticipation of the event. Namely, investors may expect that firms with the specified characteristics will be subject to the event under study. In this case, the linear specification will not uncover a relation between the variables. Moreover, the greater the connection between the specified characteristics and the occurrence of the event—that is, the more highly the event is anticipated—the less likely a relation with be found in the cross-section because the information effect (the AR) will be that much smaller (Bhagat and Jefferis, 1991; and Prabhala, 1997). MacKinlay (1997) provides an overview and further references. The issue also implicates event studies in general, for if the anticipation is sufficiently great, there will be no announcement effect; given this possibility, some researchers have proposed the use of a conditional approach instead of the conventional approach that we have discussed (for example, Acharya, 1988). However, Prabhala (1997) shows that the significance test for the existence of an information effect in the traditional methodology is, in fact, well-specified. He also shows the circumstances under which the regression coefficients on firm characteristics in traditional cross-sectional models are proportional to the true cross-sectional parameters, and hence the associated t-statistics may be interpreted as a conservative (lower bound) estimate of the parameters' true statistical significance. We therefore conclude that the principal use of cross-sectional models will continue to be for refinement of researchers' theories for undertaking their event studies by explaining the results of the standard model, that is, for relating the size and sign of the abnormal returns to specified firm and event characteristics.

2.4. Assessing the usefulness of the event study methodology for corporate law research

The standards for conducting an event study are well established. A researcher can increase the power of an event study by increasing the sample size, or/and narrowing the public announcement to as short a time-frame as possible. Users of event studies for policy analysis in corporate law should therefore keep those factors in mind—sample size and event interval—when evaluating the results.

How large should the sample size be? In general, the larger the better. This said, the recommended sample size would depend on the magnitude of the abnormal return that one is trying to detect. If the abnormal return is about 1 percent (and the announcement window can be narrowed to one day) then a sample of 100 firms would be sufficient. If the abnormal return is only 0.5 percent (and the announcement window can be narrowed to one day) then we would recommend a sample of 200 firms. On the other hand, in general, a sample of just one firm would be quite inadequate in detecting an abnormal return of even 2 percent.

Regarding the length of the announcement window: the shorter the better. If one is using daily return data, an announcement window of one day is quite feasible and the window that we recommend. However, in going from one to two or three days, the loss in statistical power is not serious. But it is very difficult to have much confidence in the results of event studies that consider long-horizon returns of several years.

Many topics of interest to legal researchers involve events that will produce a data set that does not fall into these extreme cases. For instance, if the topic of investigation is the wealth effect of a specific state law, it may be impossible to identify a one-day event interval. Given the nature of the legislative process, statutory changes typically occur over an interval significantly longer than one day, encompassing at least several months. In this setting, the researcher should try to narrow the event interval as best as he or she can: for instance, by examining the impact on returns only of specific event days (introduction of the bill, committee hearing, chamber vote) over the longer legislative interval. But identification of a single event day is not always possible. In addition, the number of firms affected by one state statute is likely to be substantially below 100 in all but a few states.

Inability to increase sample size or narrow the event interval does not indicate that the methodology cannot or should not be used: rather, it means that interpretation of results, such as a finding of insignificance, should be undertaken with care. For a sample of 50 firms and an event date consisting of a one week interval, for example, the event would have to produce an abnormal return of about 4 percent to be reliably detected, although there may be a further question whether a smaller level of abnormal returns would be considered economically significant.

3. Econometric issues: endogeneity in corporate governance and performance studies

Bhagat and Jefferis (2002) note that a vast theoretical and empirical literature in corporate finance considers the inter-relationships between corporate governance, takeovers, management turnover, corporate performance, corporate capital structure, and corporate ownership structure. In the following sub-sections we review the theoretical literature that provides support for relationships among subsets of these variables and the problem those relationships pose for empirical analysis.

3.1. Corporate control, performance, and governance

The interpretation of takeovers and managerial turnover as mechanisms for discipline may be motivated by incentive-based economic models of managerial behavior. Broadly speaking, these models fall into two categories. In agency models, a divergence in the interests of managers and shareholders causes managers to take actions that are costly to shareholders. Contracts cannot preclude this activity if shareholders are unable to observe managerial behavior directly, but ownership by the manager may be used to

induce managers to act in a manner that is consistent with the interest of shareholders.[3] Performance is reflected in managerial payoffs, which may be interpreted as including takeovers and managerial turnover. Grossman and Hart (1983) describe this problem.

Adverse selection models are motivated by the hypothesis of differential ability that cannot be observed by shareholders. In this setting, ownership may be used to induce revelation of the manager's private information about cash flow or his ability to generate cash flow, which cannot be observed directly by shareholders. Performance provides information to the principal about the ability of the manager, and is therefore reflected in managerial payoffs, which may include dismissal for poor performance. A general treatment is provided by Myerson (1979).

In this setting, takeover defenses may be interpreted as a characteristic of the contract that governs relations between shareholders and managers. The presence of takeover defenses is affected by the same unobservable features of managerial behavior or ability that are linked to ownership and performance.

3.2. *Corporate governance and performance*

Corporate governance could affect firm performance, but firm performance could also affect governance. The factors that determine governance structure are not well understood, but governance, for example, board composition, is known to be related to industry (Agrawal and Knoeber, 2001) and to a firm's ownership structure (firms with high inside ownership have less independent boards; see Bhagat and Black, 2002). If board composition is endogenous, ordinary least squares (OLS) coefficient estimates can be biased (because the error terms are correlated with the endogeneous variable). Simultaneous equations methods can address endogeneity, but are often more sensitive than OLS to model misspecification; for an example of the sensitivity of results depending on the model used to examine the relationship among board composition, insider ownership and performance, in which "relatively minor changes in [the full model and first-stage regression] have profound effects on overall results" see Barnhart and Rosenstein (1998, p. 14).

3.3. *Corporate ownership and performance*

Similar endogeneity concerns are implicated by the relation between corporate ownership and performance. For reasons related to performance-based compensation and insider information, firm performance could be a determinant of ownership. For example, superior firm performance leads to an increase in the value of stock options owned

[3] This suggests a positive relationship between ownership and performance. However, as pointed out by Stulz (1988), ownership has both an incentive effect through a stake in the firm's cash flows and an entrenchment effect through control of votes. As ownership gets large enough, there is no way to take a corporation over. Recent evidence in Himmelberg, Hubbard, and Palia (1999) suggests that econometric estimation of the effect of managerial ownership may be quite difficult for the reasons noted in this section.

by management which, if exercised, would increase their share ownership. In addition, if there are serious divergences between insider and market expectations of future firm performance, then insiders have an incentive to adjust their ownership in relation to the expected future performance; Seyhun (1998) provides evidence on this. Himmelberg, Hubbard, and Palia (1999) argue that the ownership structure of the firm may be endogenously determined by the firm's contracting environment which differs across firms in observable and unobservable ways. For instance, if the scope for perquisite consumption is low in a firm then a low level of management ownership may be the optimal incentive contract.

The endogeneity of management ownership has also been noted by Jensen and Warner (1988, p. 13): "A caveat to the alignment/entrenchment interpretation of the cross-sectional evidence, however, is that it treats ownership as exogenous, and does not address the issue of what determines ownership concentration for a given firm or why concentration would not be chosen to maximize firm value. Managers and shareholders have incentives to avoid inside ownership stakes in the range where their interests are not aligned, although managerial wealth constraints and benefits from entrenchment could make such holdings efficient for managers."

The primary responsibility of the corporate board of directors is to engage, monitor, and, when necessary, replace company management. The central criticism of many modern public company boards has been their failure to engage in the kind of active management oversight that results in more effective corporate performance. It has been suggested that substantial equity ownership by the outside directors creates a personally-based incentive for active monitoring. An integral part of the monitoring process is the replacement of the CEO when circumstances warrant. An active, non-management obligated board will presumably make the necessary change sooner rather than later, as a poorly performing management team creates more harm to the overall enterprise the longer it is in place. On the other hand, a management dominated board, because of its loyalty to the company executives, will take much longer to replace a poor performing management team because of strong loyalty ties. Consequently, it may be argued that companies where the CEO is replaced expeditiously in times of poor performance may have more active and effective monitoring boards than those companies where ineffective CEOs remain in office for longer periods of time. Bhagat, Carey, and Elson (1999) find that when directors own a greater dollar amount of stock, they were more likely to replace the CEO of a company performing poorly.

The above discussion focuses on the costs of diffuse share-ownership; that is, the impact of ownership structure on performance. Demsetz (1983) argues that since we observe many successful public companies with diffuse share-ownership, clearly there must be offsetting benefits, for example, better risk-bearing. Sometimes, as in the case of leveraged buyouts, when the benefits are substantially less than the costs of diffuse share-ownership, we do observe companies undergoing rapid and drastic changes in their ownership structure. In other words, ownership structure may be endogenous.

3.4. Corporate governance and ownership structure

The corporate charter is a contract that governs relations between managers and shareholders. Most studies of management-sponsored antitakeover amendments adopted by the shareholders focused mainly on the wealth effects associated with the amendments, as discussed in Section 4, and secondarily on the ownership structure of the firms that adopt them. There are patterns associating ownership and takeover defenses. Jarrell and Poulsen (1987), for instance, report above-average insider holdings and below-average institutional holdings in a large sample of firms enacting amendments. It is also plausible that these corporate characteristics are endogenously determined. Shareholder support for amendments involving takeover defenses has been attributed to free-rider problems (Jarrell, Brickley, and Netter, 1988). Bhagat and Jefferis (1991) argue that the transaction costs that give rise to the free-rider problem are, at least in part, an endogenous consequence of strategic behavior by managers using the proxy process that might be eliminated through either changes in the charter or proxy reform. The next section provides a model for empirical research that takes into account the endogeneity across corporate governance, and in particular takeover defenses, ownership structure and performance.

3.5. Simultaneous equations estimation

Given the above considerations regarding the endogeneity among corporate governance, ownership, takeovers, and performance, we propose the following system of equations as appropriate for modeling the interactive effect.

$$\text{Performance} = f_1(\text{Ownership, Governance, Takeover, } \mathbf{Z}_1, \varepsilon_1) \tag{1}$$

$$\text{Governance} = f_2(\text{Ownership, Performance, Takeover, } \mathbf{Z}_2, \varepsilon_2) \tag{2}$$

$$\text{Ownership} = f_3(\text{Performance, Governance, Takeover, } \mathbf{Z}_3, \varepsilon_3) \tag{3}$$

$$\text{Takeover} = f_4(\text{Performance, Governance, Ownership, } Z_4, \varepsilon_4) \tag{4}$$

In Equations (1) through (4) the $\mathbf{Z}_i$ are vectors of instruments that affect the dependent variable. The error terms ε_i are associated with exogenous noise and the unobservable features of managerial behavior or ability that explain cross-sectional variation in ownership, performance and governance. Identification requires some combination of exclusion restrictions, assumptions about the joint distribution of the error terms, and restrictions on the functional form of the f_i. Maddala (1983) discusses restrictions that identify the model when the ε_i are normally distributed. Identification in *single equation* semiparametric index models, where the functional form of f_i is unknown and the explanatory variables in that equation are continuous, known functions of a basic parameter vector is discussed by Ichimura and Lee (1991). Estimation of a system of the form (1)–(4) in the absence of strong restrictions on *both* the f_i and the joint distribution of error terms is, to the best of our knowledge, an unsolved problem.

We are unaware of a model of takeover defense that implies specific functional forms for the f_i. If these functions are linear, identification may be attained through either strong distributional assumptions or exclusion restrictions. Maddala (1983) and Amemiya (1985) discuss restrictions on the ε_i that identify the model in the absence of exclusion restrictions. But these restrictions are inconsistent with incentive-based explanations of takeover defenses, since unobservable characteristics of managerial behavior or type will be reflected in all of the ε_i. Using panel data and firm-fixed effects it would be possible to control for unobservable characteristics of managerial behavior or type; however, a system such as in (1)–(4) would have to be specified and estimated. Aside from the non-trivial data collection effort required to estimate such a system, this system would not be identified when $\mathbf{Z}_2 = \mathbf{Z}_3 = \mathbf{Z}_4$. Exclusion restrictions are therefore the most likely path to identification.

However, exclusion restrictions would be difficult to justify. Intuitively, variables that affect the likelihood of a takeover will be reflected in the structure of takeover defenses. A detailed microeconomic model, based on specific assumptions about preferences and production possibilities, might yield exclusion restrictions. But we are unaware of any candidates and suspect that the same features of the data that yield identification (for example, a Cobb-Douglas production technology) would render the model inconsistent with the data; see Griliches and Mairesse (1999).[4] In the absence of distributional assumptions or functional form restrictions, the econometric model (1)–(4) is not identified when $\mathbf{Z}_2 = \mathbf{Z}_3 = \mathbf{Z}_4$.

The difficulties presented by a complete simultaneous equation model is one reason why event studies are the more preferable form of empirical research in corporate law.[5] Our review of the empirical literature will consequently primarily focus on stock price studies; we will touch on econometric analyses of corporate governance, ownership and performance, but many of those studies are subject to the caveat that they do not control for the endogeneity concerns identified in this section.

4. Empirical research in corporate law

4.1. Shareholder wealth implications of corporate lawsuits

In the 1980s–1990s, business frequently complained about a litigation explosion and the costs associated with legal disputes, raising concerns that the U.S. legal system af-

[4] In a recent paper, Coles, Meschke, and Lemmon (2003) construct a structural model of the firm and calibrate the exogenous parameters of the model to data. While these authors have made a significant contribution to addressing the endogeneity concerns in this literature, the problems of estimating firm production functions, as noted in Griliches and Mairesse (1999), are still relevant.

[5] Event studies of firm choices, such as corporate governance mechanisms or corporate domicile discussed in Section 4, that are voluntarily undertaken by firms, avoid the endogeneity concern because they study the wealth effect of the choices of firms that have (voluntarily) made the choice. No claims are made in an event study that the price effect would be the same for firms that have not taken the particular decision, as it is for the firms under study.

fected firms' competitiveness in global markets. Surveying corporate legal department budgets, Economic Analysis Group, Ltd., Craig Consulting Co., and Endispute, Inc. estimated that salaries to in-house lawyers and fees to outside counsel for the 1000 largest public companies hit $20 billion in 1991.[6] Large liability or settlement payments undoubtedly dwarf direct legal costs. Indeed, some mass torts, such as the breast-implant cases against Dow Corning and the Dalkon Shield cases against A.H. Robins, have threatened the existence of defendant firms, forcing them into insolvency proceedings.

It is, however, possible that estimates of business' legal costs are overstated, reflecting political agendas or overreaction to media coverage of a few spectacular cases. Many large publicized damage awards, for example, are overturned on appeal or significantly reduced in a settlement (Shanley and Peterson, 1987). In addition, much corporate litigation involves contract disputes between firms.[7] But concerns over litigation continued in the 1990s: tort reform was one of ten points in the Republican party's "Contract with America" 1994 campaign platform under which it gained a majority in the House of Representatives for the first time in 40 years, and successful litigation initiatives against tobacco companies that produced a settlement of over $200 billion have led to other industry targets, such as health care providers and fast food restaurants.

Event studies can be used to identify and measure the costs of lawsuits against firms, and they have been used to evaluate the costs of interfirm litigation. The results are quite uniform: when the costs and benefits to both parties are computed, litigation is not a positive net present value event for both firms considered together. This result is not surprising: it is an impetus motivating the successful move to greater use of alternative dispute resolution, particularly in the corporate context.

4.1.1. Wealth effects of corporate litigation

The primary focus in the literature has been on "leakages" in the litigation process: negative wealth effects upon netting the parties' gains and losses. For example, Cutler and Summers (1988) examine the Pennzoil/Texaco lawsuit, which involved a claim of tortious interference of a merger contract, and find significant costs to both parties from the dispute, with the losses for the losing defendant Texaco, being larger than the gains for the winning plaintiff Pennzoil. The combined drop in value for the two firms was $2 billion. They attribute the loss mainly to an increase in the probability of financial distress for Texaco. Engelmann and Cornell (1988) study the wealth implications around filings, settlements, and verdicts for a sample of five interfirm disputes. They too observe combined wealth losses, or leakages, to the litigating parties. Bhagat, Brickley, and Coles (1994) examine the market reaction to lawsuit filings and settlements for a

[6] An article in Forbes, citing statistics from a Rand study on tort litigation, estimated the direct costs of all lawsuits, including those involving business, to be as high as $117 billion a year (Spencer, 1992, p. 40). Another estimate (id., p. 41) placed litigation costs as high as 2.5 percent of GNP.

[7] For example, a Rand study of Fortune 1000 companies found that contract disputes between firms constituted the largest single category of federal civil suits (Dungworth and Pace, 1990).

much larger sample of 550 interfirm disputes. They observe combined wealth losses arising from lawsuit filings and find that these leakages are a result of increased probability of financial distress for the defendant. In addition, they find that defendant firms gain upon the announcement of a settlement.

Ellert (1975) examines the market responses to announcements of legal challenges to mergers under Section 7 of the Clayton Act by the Federal Trade Commission and Department of Justice over the period 1950–1972. During the month of the announcement of the suit, the market adjusts defendant firm value downward by about two percent. Bizjak and Coles (1995) analyze a more homogeneous but still large sample of interfirm disputes—private antitrust suits. To our knowledge, this is the only study to find a positive stock market reaction to plaintiffs upon any sort of lawsuit filing. They also find that the joint wealth effects associated with the announcement of a filing tend to be negative and that leakages in antitrust disputes are attributable to court-imposed behavioral restraints, the likelihood of follow-on suits, and an increased likelihood of financial distress. Moreover, they confirm that factors which affect the costs of litigation also affect behavior in suit, settlement, and trial. In their sample of antitrust lawsuits, the parties are more likely to settle when the suit involves potential restrictions on the defendant's business practices and when there is the potential for financial distress.

Event studies have also been used to address the validity of the government's antitrust actions against various corporations. The argument goes that for a corporation exercising market power, the government's antitrust action against it will lower its share price *and* increase the share price of its competitors. The competitors will experience a positive reaction since the government's antitrust action increases the odds that these competitors will be competing in an industry without a dominant company that might be exercising market power. Bittlingmayer and Hazlett (2000) use this intuition to evaluate the U.S. Department of Justice's recent antitrust action against Microsoft. They find evidence inconsistent with the joint hypothesis that Microsoft's behavior has been anticompetitive and that antitrust enforcement enhances economic efficiency.

Finally, Bhagat, Bizjak, and Coles (1998) analyze a large sample of lawsuits in which at least one side, plaintiff or defendant, is a corporation. To estimate the implications of litigation for shareholder wealth, they examine the abnormal stock market reaction to filing and settlement announcements. They find that the average wealth loss for a defendant is 0.97 percent of the market value of the equity, or $15.96 million. They further test whether characteristics of the suit, such as legal issue, type of opponent, and firm characteristics (such as firm size and proximity to bankruptcy) have power to explain cross-sectional variation in these wealth effects.

Bhagat, Bizjak, and Coles (1998) find that no matter who brings a lawsuit against a firm, be it a government entity, another firm, or private citizen, defendants experience economically-meaningful and statistically-significant wealth losses upon the filing of the suit. Furthermore, they find some evidence that the identity of the plaintiff has an influence on the wealth effects upon filing. Defendants involved in government suits suffer larger declines in shareholder wealth (−1.73 percent) than defendants involved in

lawsuits with other firms (−0.75 percent) or with private parties (−0.81 percent).[8] This result is consistent with the notion that government agencies have more leverage and resources at their disposal to use in a legal battle and/or the type of suit most frequently filed by government agencies, such as an environmental action, is typically more serious. Indeed, they do find that certain types of litigation are more costly for defendants. Environmental suits (−3.08 percent), product liability suits (−1.46 percent), and violations of securities laws (−2.71 percent) result in significantly greater wealth losses for defendant firms, compared to disputes involving antitrust or breach of contract issues. It appears that, at least for some types of suits, the actual or potential lawsuit is associated with a large decline in shareholder wealth and a corresponding nontrivial deterrent effect. The results of these and other studies that consider the impact of litigation on corporate value are summarized in Table 1.

Bhagat, Bizjak, and Coles (1998) also find that the defendant wealth effect on announcement of a filing is significantly positively related to the size of the firm and, in some specifications, significantly negatively related to the firm's proximity to bankruptcy. One possible explanation for this effect of firm size is that larger firms can have more bargaining power or more resources to devote to the legal dispute (e.g., because of better access to capital markets or "deep pockets"). The results on proximity to bankruptcy are consistent with other work that has identified potential bankruptcy costs as an important indirect cost of a legal dispute (Bhagat, Brickley, and Coles, 1994; Bizjak and Coles, 1995; and Cutler and Summers, 1988).

For plaintiff firms, they find no significant wealth effects associated with lawsuit filings. They also find that the identity of the defendant—that is, whether the defendant is another firm, a government agent, or private citizen—and the legal issue are not related to the stock price change of the plaintiff when a suit is filed. They are, accordingly, unable to detect in the data evidence of strong incentives for plaintiffs to sue.

Bhagat, Bizjak, and Coles (1998) results indicate that when a defendant firm settles a suit with another firm there is a significant wealth increase. It is surprising that, in contrast, they can detect no significant wealth change for defendants upon announcement of a settlement when the opponent is a governmental entity or noncorporate private party.[9] In addition, the wealth effect of a settlement for the defendant is unrelated to the legal issue. For plaintiff firms the wealth implications of settlements appear to be trivial. On average, they find no significant wealth gains or losses to plaintiff firms who settle a lawsuit, and neither legal issue nor the identity of the opposing party has power to explain variation in those returns. These data suggest that lawsuits are not positive net present value undertakings for plaintiffs, since the absence of positive abnormal returns

[8] Note that a related finding in event studies of government legal and regulatory actions against firms is that the market appears to impose a higher sanction than actual criminal sanctions and the reputational losses are of equal magnitude for civil fines as for criminal fines (see Bhagat and Romano, 2002a, pp. 161–162).

[9] In a recent paper, Haslem (2003) documents a significant negative market response to settlements for defendant corporations. He also finds a marginally positive response if the defendant corporation litigates—regardless of the outcome.

Table 1

Panel A. Announcement period abnormal returns for **defendant** corporations by opponent type

Plaintiff	Study	Sample period	Sample size	Announcement window: (event days)	Announcement return (%)	Z-statistic
Another firm	BBC (1998)	1981–1983	239	Filing (−1, 0)	−0.75**	−3.31
Government	BBC (1998)	1981–1983	110	Filing (−1, 0)	−1.73**	−4.99
Private non-firm	BBC (1998)	1981–1983	221	Filing (−1, 0)	−0.81**	−2.67
Another firm	BC (1995)	1973–1983	343	Filing (−1, 0)	−0.60**	−3.17
Stakeholders	KL (1993)	1978–1987	19	Allegation (−1, 0)	−1.34	−1.21
Stakeholders	KL (1993)	1978–1987	25	Filing (−1, 0)	−1.67*	−2.35
Government	KL (1993)	1978–1987	13	Allegation (−1, 0)	−5.05**	−4.77
Government	KL (1993)	1978–1987	17	Filing (−1, 0)	−0.93	−1.14
Stakeholders	KL (1999)	1979–1995	80	Filing (−1, 0)	−1.02**	−2.86
Consumers	PR (2002)	1985–1995	15	Filing (−1, 1)	−1.93**	−3.31
Another firm	BBC (1998)	1981–1983	12	Settlement (−1, 0)	3.66**	3.29
Government	BBC (1998)	1981–1983	4	Settlement (−1, 0)	−0.68	−0.22
Private non-firm	BBC (1998)	1981–1983	12	Settlement (−1, 0)	−1.06	−1.72
Stakeholders	KL (1993)	1978–1987	13	Settle/Verdict (−1, 0)	−0.17	−0.49
Government	KL (1993)	1978–1987	10	Settle/Verdict (−1, 0)	1.48	1.20
Stakeholders	KL (1999)	1979–1995	15	Verdict-Defense (−1, 0)	−0.36	−0.51
Stakeholders	KL (1999)	1979–1995	193	Verdict-Plaintiff (−1, 0)	−0.62*	−2.74
Stakeholders	KL (1999)	1979–1995	4	Settlement (−1, 0)	−2.43	−1.35
Consumers	PR (2002)	1985–1995	25	Verdict-Plaintiff (−1, 1)	0.33	0.73
Government	H (2003)	1994–1998	13	Settlement (−1, 3)	−0.25	nr
Another firm	H (2003)	1994–1998	285	Settlement (−1, 3)	−0.65	nr
Private non-firm	H (2003)	1994–1998	439	Settlement (−1, 3)	0.05	nr

Panel B. Announcement period abnormal returns for **plaintiff** corporations by opponent type

Defendant	Study	Sample period	Sample size	Announcement window: (event days)	Announcement return (%)	Z-statistic
Another firm	BBC (1998)	1981–1983	172	Filing (−1, 0)	−0.25	−0.60
Government	BBC (1998)	1981–1983	26	Filing (−1, 0)	−0.44	−0.80
Private non-firm	BBC (1998)	1981–1983	51	Filing (−1, 0)	0.71	0.34
Another firm	BC (1995)	1973–1983	86	Filing (−1, 0)	1.24**	4.26
Another firm	BBC (1998)	1981–1983	8	Settlement (−1, 0)	−0.77	−1.26

Panel C. Announcement period abnormal returns for **defendant** corporations by type of legal issue

Legal issue	Study	Sample period	Sample size	Announcement window: (event days)	Announcement return (%)	Z-statistic
Antitrust	BBC (1998)	1981–1983	62	Filing (−1, 0)	−0.81	−1.52
Breach of contract	BBC (1998)	1981–1983	48	Filing (−1, 0)	−0.16	−0.59
Corp. governance	BBC (1998)	1981–1983	154	Filing (−1, 0)	0.08	0.64
Environment	BBC (1998)	1981–1983	27	Filing (−1, 0)	−3.08**	−5.32

Table 1
(*Continued*)

Panel C. Announcement period abnormal returns for **defendant** corporations by type of legal issue

Legal issue	Study	Sample period	Sample size	Announcement window: (event days)	Announcement return (%)	Z-statistic
Exclusive dealing	BBC (1998)	1981–1983	27	Filing (−1, 0)	−0.14	0.28
Patent infringement	BBC (1998)	1981–1983	33	Filing (−1, 0)	−1.50*	−2.42
Product liability	BBC (1998)	1981–1983	38	Filing (−1, 0)	−1.46**	−3.12
Disclosure laws	BBC (1998)	1981–1983	46	Filing (−1, 0)	−2.71**	−4.49
Antitrust-horizontal	BC (1995)	1973–1983	117	Filing (−1, 0)	−1.45**	−4.88
Antitrust-vertical	BC (1995)	1973–1983	105	Filing (−1, 0)	0.27	1.29
Fraud of stakeholders	KL (1993)	1978–1987	19	Allegation (−1, 0)	−1.34	−1.21
Fraud of stakeholders	KL (1993)	1978–1987	25	Filing (−1, 0)	−1.67*	−2.35
Fraud of government	KL (1993)	1978–1987	13	Allegation (−1, 0)	−5.05**	−4.77
Fraud of government	KL (1993)	1978–1987	17	Filing (−1, 0)	−0.93	−1.14
Fin. reporting fraud	KL (1993)	1978–1987	4	Allegation (−1, 0)	−4.60**	−2.00
Fin. reporting fraud	KL (1993)	1978–1987	7	Filing (−1, 0)	−4.56*	−1.99
Punitive damages	KL (1999)	1979–1995	80	Filing (−1, 0)	−1.02*	−2.86
Product liability	PR (2002)	1985–1995	15	Filing (−1, 1)	−1.93**	−3.31

*Significant at 0.05 level.
**Significant at 0.01 level.
Event day 0 is the publication date of the filing, allegation, or settlement.
BBC (1998): Bhagat, Bizjak, and Coles (1998).
BC (1995): Bizjak and Coles (1995). Horizontal antitrust issues include horizontal price-fixing, merger/joint-venture, asset accumulation, predatory pricing, and monopolization. Vertical antitrust issues include resale price maintenance, exclusive dealing, tying, territorial restrictions, dealer termination, and refusal to deal.
KL (1993): Karpoff and Lott (1993). Fraud of stakeholders occurs when the firm is accused of cheating on implicit or explicit contracts with suppliers, customers, or employees. Fraud of government occurs when the firm is accused of cheating on implicit or explicit contracts with government agencies. Financial reporting fraud occurs when the firm is accused of misrepresenting the firm's financial condition.
KL (1999): Karpoff and Lott (1999). Punitive damages are sought in cases involving product liability, fraud, business negligence, breach of contract, insurance claims, employment claims, asbestos claims, and vehicular accident claims.
PR (2002): Prince and Rubin (2002). Product liability claims involving auto manufacturers.
H (2003): Haslem (2003).
nr: not reported.

on settlement cannot be explained by investor anticipation upon the lawsuit filing (there was no significant positive gain at the earlier date).

Two caveats are in order regarding the findings concerning the wealth effects of corporate litigation discussed in this section. First, the announcement-period abnormal return understates the expected decline in shareholder wealth. The reason is that information about the forthcoming suit may already have reached the market (prior to the announcement in the press) and therefore already be reflected in the market price of the firm's stock. Most of the studies have attempted to reduce the severity of this problem by excluding cases where there was indication in published news reports that informa-

tion about the suit had previously reached the public. Second, event studies of litigation report the average market response associated with the filing or settlement of a lawsuit. Under what circumstance would a court, corporate manager or corporate legal counsel use such information? Virtually, no litigation situation is an average situation. Each suit represents a unique set of costs and benefits, and managers deciding whether to launch or defend a suit will consider the specific costs and benefits of their situation, rather than the average market response to a collection of suits that may or may not share similar characteristics. However, it is precisely information in a wide spectrum of suits that is most useful for the ex ante formulation of public policy and corporate strategy.

4.1.2. Corporate litigation brought by shareholders: derivative and securities lawsuits

Studies have investigated the wealth effects of corporate litigation involving suits brought by shareholders against corporate officers and directors for fiduciary breach (e.g., Fischel and Bradley, 1986; Romano, 1991). These suits are referred to as derivative suits, as the shareholder brings the action in the name of the corporation rather than herself; the right to sue is derived from the loss experienced by the corporation (technically the shareholder sues the board of directors for not pursuing the fiduciary claims against the individuals accused of misconduct, since corporate law places the litigation decision in the board; under specific circumstances when the board refuses to take action, courts permit the plaintiffs to proceed with the suit in place of the corporation). Shareholders may sue officers and directors for fiduciary breach in their own right when the misconduct affects their rights as shareholders (voting rights, dividend rights, etc.); these suits may be brought individually or in a representative capacity as a class action. Plaintiffs prefer to bring that type of suit when possible, in order to avoid a variety of procedural barriers that apply to derivative suits.

Because the efficacy of shareholder litigation as a device to monitor is hampered by collective action problems—the cost of bringing the lawsuit will typically be greater than the shareholder's pro rata benefit, although less than the aggregate gain across all owners—the law provides financial incentives to attorneys to prosecute cases. Successful plaintiffs are awarded counsel fees, even in the absence of a monetary recovery to the plaintiff, as is often the case in derivative claims, and the calculation of the fee award includes compensation for risk beyond hours worked. This resolution of the collective action problem creates an agency problem in that the attorney's incentives need not coincide with the shareholders' interest (Coffee, 1985). Compounding that agency problem is the interaction between the legal regime on indemnification and directors' and officers' liability insurance: individual expenditures on settlements or judgments in derivative litigation may not be indemnified, while liability insurance policies exclude deliberate dishonesty or fraud. Individual directors and officers have a powerful incentive to settle, even if the case has no merit, as that will avoid the possibility of an adjudication of fraud invalidating the insurance policy, however remote, and thereby guarantee no out-of-pocket expenditures.

As is true of most civil litigation, the majority of shareholder suits settle (Romano, 1991, p. 60). The attorney-client agency problem, when coupled with the incentives of

defendants to settle, appears to produce two problematic trends: frivolous claims tend to be overcompensated and meritorious claims undercompensated, and the plaintiffs' bar is the principal beneficiary of the system (see Romano, 1991). The event study data regarding the wealth effects of the litigation are, however, mixed. Romano (1991) finds no significant price effect for derivative lawsuit filings.[10] One explanation of the difference between this finding and that of other corporate litigation summarized in Table 1 is that the market anticipates the outcomes of derivative suits: the suits typically result in no or very low monetary rewards. In Romano's sample, for instance, most derivative suits did not produce a monetary recovery, and for those that did, the value was less than 0.5 percent of firm assets, or $0.15 per share net of attorneys' fees (pp. 61–62). Romano also finds no significant price effect for lawsuit dispositions: dismissed derivative suits have insignificant negative returns, but so do settled suits. Fischel and Bradley (1986) find a significant negative reaction to suit terminations and an insignificant positive reaction to judicial decisions not to dismiss. Although this result might suggest that the market views derivative suits positively, they conclude that the suits have no significant wealth effects because when returns are cumulated around the filing date they are insignificant.

The event study methodology is not directed at measuring any potential deterrent effect of shareholder lawsuits. Such a third-party effect would not be incorporated in a sued firm's stock price. Because in order for lawsuits to deter misconduct generally, managers who are sued need to suffer a penalty or sanction (there must be specific deterrence), Romano (1991) investigated whether top management of sued firms experienced a decline in compensation, an increased frequency of termination, or a decrease in directorships held on other companies' boards, compared to management of firms, matched by industry and size, that were not sued. Romano failed to find any significant differences across management on all of the dimensions she measured, compensation, employment, and directorships, and therefore concluded that derivative suits do not provide specific deterrence. This finding raises the possibility that the litigation does not serve as a mechanism of general deterrence. Romano's study suggests that shareholder litigation does not provide much in the way of benefit (compensatory or deterrent) to investors, as opposed to their attorneys, and therefore lent credence to the view that many of the claims of misconduct underlying such suits are insubstantial and that the procedures that have been devised to restrict the litigation should be expanded rather than reduced.

To the extent that litigation patterns have shifted since Romano's study, the implication regarding a need for reform may no longer be accurate. Romano found that litigation over acquisitions and defensive tactics, which constituted approximately 40 percent of the disputes in her sample, increased fivefold over the period of her study, the late 1960s through the beginning of 1987. Thompson and Thomas (2003) more recently examined shareholder litigation and report that over 80 percent of the shareholder suits

[10] Romano does find a significantly negative price effect for the filing of a class action, which is in keeping with the findings in the literature reported in the previous section.

filed in 1999–2000 in the Delaware Chancery court, which is the leading corporate law jurisdiction, involved acquisitions. As a consequence, most of the claims in Thompson and Thomas' sample were brought as class actions, rather than derivative suits. Thompson and Thomas find that the greatest benefits (increased premiums) occur in the class action cases of acquisitions by controlling shareholders. In comparing the derivative and class action claims, characteristics they consider to be indicia of nonmeritorious claims—multiple lawsuits, with identical language in the complaints, filed by a small set of law firms very shortly after the announcement of the event on which the litigation is based, and few substantive motions filed after the initial complaint—are more frequently observed in the class action acquisitions cases than the derivative suits. They therefore conclude, in contrast to the implications of Romano's earlier study, that relaxing the procedures for derivative actions and tightening those for class actions would be beneficial. Weiss and White (2004) provide further support for such a conclusion: they analyze 104 merger-related class actions brought in Delaware from 1999–2001, and describe an out-of-control process of "opportunistic filings" and collusive settlements in which attorneys are awarded fees in amounts in relation to recoveries that the authors consider suggest "little value" added by the attorneys to the recoveries.

Another class of suits brought by shareholders against corporations are securities class actions (lawsuits brought by investors with losses on purchases or sales of stock for violations of the federal securities laws). As in the derivative suit context, the litigation environment implicates concerns over collusive settlements, in which defendants pay damages on frivolous claims to avoid the higher cost of litigation and an organized plaintiff's bar takes a large share of the recovery (which is typically paid by insurers). And, as in the state law claim context, as claims paid increase, liability insurance becomes more expensive, and exposed firms lobby legislators for relief.[11] One question motivating empirical research on securities litigation is related to those concerns, whether suits are frivolous or, as some researchers have put it, whether settlements reflect the merits of claims (Alexander, 1991). For example, Alexander (1991) compares settlement value and potential damages for litigation following initial public offerings of a small number of high technology firms. She concludes that the settlement amounts do not depend on the merits of the cases. Other studies suggesting instead that settlements are related to the merits are Francis, Philbrick, and Schipper (1994), who find a positive correlation between settlements and potential damages, and Skinner (1995), who finds more untimely disclosures of adverse earnings produce less favorable litigation outcomes.

Another question is whether the market responds efficiently to information about these lawsuits. Griffin, Grundfest, and Perino (2004) seek to illuminate this question

[11] In the state law context, statutes permitting firms to limit the monetary liability of directors for negligence were enacted in the mid-1980s as a response to a perceived crisis in the market for directors' and officers' liability insurance which was, arguably, partly due to judicial expansion of fiduciary liability (Romano, 1990). In the federal context, Congress passed securities litigation reform in the mid-1990s, largely in response to concern over frivolous litigation brought against firms in the high technology sector, whose stock is volatile; many of these firms went public in the beginning of the decade and were sued under the securities law when their prices dropped subsequent to the offering.

by examining the relation between abnormal returns at the date of a corrective disclosure that could serve as the basis for litigation (such as an announcement of a financial restatement, which is referred to as the date of the end of the class action period), the date of the filing of a class action complaint, and the date at which the alleged violation occurred (such as the date of the original fraudulent financial statement, which is referred to as the date of the beginning of the class action period), for a sample of several thousand federal class actions filed from 1990–2003. They find significant predictable responses on the three event dates (negative on the class action period ending and complaint filing dates and positive on the class action period beginning date, with the response on the filing date conditional on the response on the class period ending date), but not after the dates at which the allegedly false information is revealed to the market. They also find that the responses are related to litigation characteristics (such as the content of the complaint—whether accounting violations are alleged—and the outcome of the litigation), and in particular, that the response on the class action period ending date reflects the subsequent filing and settlement amount. Griffin et al. therefore conclude that the market is "reasonably efficient" in this context.

At about the same time as researchers were focusing attention on securities litigation, the issue moved onto the national policy agenda as frivolous securities litigation became a central concern of Congress, culminating in the 1995 enactment of the Private Securities Litigation Reform Act (see e.g., H.R. Conference Report, 1995). Johnson, Nelson, and Pritchard (2007) study lawsuits filed before and after the 1995 legislation, and conclude that it appears to have been somewhat successful in weeding out nonmeritorious claims: complaints filed after 1995 appear to be more likely to raise serious accounting and insider trading issues than pre-1995 claims (allegation of an accounting issue is correlated with variables proxying for accounting wrongdoing, such as earnings restatements and abnormal discretionary accruals that signal earnings management, only for complaints filed after the 1995 statute, and insider trading allegations are correlated with net sales by insiders only for post-enactment complaints).[12] The conclusion regarding the statute's success is less robust when examining litigation outcomes (strength of claims) rather than incidence because the variables that proxy for the merits of the claim have only mixed power in differentially explaining settlement amounts

[12] The sample consists of 119 firms in the computer hardware and software industry, an industry sector that is a frequent target of securities lawsuits, that were sued from 1991–2000, and a matched control sample of firms experiencing similar price drops but that were not sued. The authors find that variables measuring the extent of damages that other studies have shown predict lawsuit filing, such as market capitalization and share turnover, did not change in significance pre- and post-legislation. As Johnson, Nelson, and Pritchard (2007) note, because damage variables determine the attorneys' recovery—greater potential damages have a higher fee award potential—those variables will always be correlated with lawsuit incidence even though they are not likely to be correlated with the merits of a claim (the likelihood of fraud). The empirical research on what firm and market characteristics explain the incidence of securities litigation is considerable; for a study whose model specification explains 41 percent of the litigation incidence in three industries in sectors with a high rate of litigation see Johnson, Kasznik, and Nelson (2000).

above nuisance value before and after the legislation: earnings restatements are positively correlated with what the authors consider non-frivolous settlements (above $2 million) post-enactment, but the level of insider trading correlates with the likelihood of non-frivolous settlement only prior to enactment. See also Perino (2003), for additional data suggesting that the statute's success at reducing nonmeritorious suits has been mixed.

4.2. Empirical research and the debate over state competition for corporate charters

In the United States, corporate law is largely a matter for the states. State corporation codes consist primarily of enabling provisions that supply standard contract terms for corporate governance. Firms choose their state of incorporation, a statutory domicile that is independent of physical presence. Midstream domicile changes require the approval of a majority of the shareholders. Firms consequently can particularize their governance arrangements both by the choices made in their charters under state law and by their choice of domicile.

One small state, Delaware, has come to dominate the incorporation process, serving as the domicile for the majority of publicly traded corporations. Its profits from providing corporate charters are considerable: for example, franchise fees averaged 17% of total tax revenues over the past 30 years (Romano, 2002, Table 4.1). Delaware's success has fueled an ongoing debate among corporate law commentators, mirroring the more general U.S. political debate over the benefits of federalism: are the aims of corporation codes—protecting the interest of the shareholders—best achieved by firms' ability to choose among domiciles compared to a centralized national regime.

A little over 25 years ago the unquestioned consensus among corporate law scholars followed the position best articulated by Cary (1974), that states were competing in a race "for the bottom," in which Delaware led the pack to produce corporate laws that decidedly favored managers' over shareholders' interests.[13] But today Cary's position is no longer accepted as a self-evident proposition. Indeed, even adherents of Cary's position in the contemporary discourse advocate federal law as an option in addition to state law, rather than preemption of state law (Bebchuk and Ferrell, 2001). What accounts for such a seismic shift?

Winter (1977) first articulated the flaw in Cary's position from the omission of markets from the analysis of firm behavior. As he explained, were managers to choose to incorporate in states whose codes disadvantaged shareholders, they would encounter a higher cost of capital and ultimately a lower job retention rate, compared to competitors operating under codes more favorable to shareholders. While Cary's position can be amended to join Winter's argument by asserting that markets are imperfect at disciplining managers when it comes to domicile choice, Winter's insight motivated

[13] For example, 80 law professors signed a letter endorsing a national corporation law in 1976 (Romano, 1993a, p. 14, n. 2).

empirically-oriented researchers to study the effect of incorporation choices on firm value, for the purpose of arbitrating the debate.

The event study methodology meshed neatly with Winter's analysis of the issue. This is because a good proxy for ascertaining whether the legal regime decisions made by firms under competition benefit investors is the effect upon shareholder wealth of a change in domicile. If a change in domicile increases firm value, it would be exceedingly difficult to maintain that charter competition, and particularly, Delaware's legal regime, is harmful to shareholders, as the overwhelming majority of firms reincorporate in Delaware (Romano, 1985; Daines, 2001).

There have been eight event studies investigating the effect on stock prices of a change in incorporation state. The event day 0 is identified as the date of the proxy mailing announcing the proposed reincorporation. All of the studies find positive abnormal returns, with four (Bradley and Schipani, 1989; Romano, 1985; Wang, 1995; Hyman, 1979) finding a significant positive stock return at the time of the announcement of the domicile change (although one of these, the earliest study by Hyman, employs a variant of the event study methodology and uses a difference-in-means test between price changes of reincorporating firms and the S&P index), one (Heron and Lewellen, 1998) finding a significant positive return for only a reduced subset of reincorporations on the announcement date, with different results—significant positive and negative returns for different subsets of reincorporations—on the subsequent shareholder meeting date, another (Dodd and Leftwich, 1980) finding a significant positive return over two years prior to the reincorporation, and two (Netter and Poulsen, 1989; Peterson, 1988) finding positive returns significant at 10 percent (albeit in one of these, the study by Peterson, the finding holds only for a subset of reincorporations) (see Table 2). As indicated in the table, the sample size in many of the studies finding significant positive abnormal returns is large (over 100 firms), whereas some of the studies that report a significant abnormal return at only a 10 percent significance level have small samples (less than 40 firms). Hence the difference could be attributed to the more limited power of the test for small samples, as discussed in Section 2. The event study literature thus suggests that Winter's core insight is accurate: competition for corporate charters benefits investors. One certainly cannot read the event study literature and conclude that firms reincorporating are reducing their shareholders' wealth, as Cary's position implies.

Bebchuk, Cohen, and Ferrell (2002, p. 1791), who are critics of state competition for corporate charters, criticize the conclusion that the event study data indicate that shareholders benefit from competition by contending that the significantly positive abnormal returns are "rather small" or "modest" (on average 1.28 percent). However, an investment project that generates positive abnormal returns of even 1 percent is considerable for competitive capital markets: for example, the magnitude of the price effect of announcements of capital expenditures, joint ventures, product introductions and acquisitions is less than 1 percent (Andrade, Mitchell, and Stafford, 2001, p. 119). That positive abnormal return is also three times greater than the negative abnormal return found in the most comprehensive event study of takeover statutes (Karpoff and Malatesta, 1989).

Table 2
Announcement period abnormal returns for firms changing their state of incorporation

Study	Sample period	Sample size	Announcement window: (event days)	Announcement return (%)	Z-statistic
BS (1989)	1986–1988	32	0	1.04*	2.21
Romano (1985)	1961–1983	150	(−1, +1)	4.18**	11.92
Romano (1985)	1961–1983	63 (m&a)	(−1, +1)	6.94**	11.44
Romano (1985)	1961–1983	43 (to)	(−1, +1)	0.05	0.77
Wang (1995)	1986–1994	145	(−1, +1)	0.97*	1.99
Wang (1995)	1986–1994	94 (Del)	(−1, +1)	1.12	1.65
Wang (1995)	1986–1994	51 (non-Del)	(−1, +1)	0.69	0.87
HL (1998)	1980–1992	294	(0, +3)	−0.15	−0.51
HL (1998)	1980–1992	45 (to)	(0, +3)	−0.51	−1.08
HL (1998)	1980–1992	59 (ll)	(0, +3)	1.20	1.66
HL (1998)	1980–1992	49 (oth)	(0, +3)	−1.23	−0.72
DL (1980)	1927–1977	140	0 (month)	1.58	nr
Hyman (1979)	1968–1976	26	Announcement week	2.73*	2.01
NP (1989)	1986–1987	36	(−1, +1)	0.93	1.61
NP (1989)	1986–1987	19 (Cal)	(−1, +1)	0.96	1.57
NP (1989)	1986–1987	17 (non-Cal)	(−1, +1)	0.89	0.68
Peterson (1988)	1969–1984	30	−1	0.27	1.35
Peterson (1988)	1969–1984	14 (to)	−1	−0.16	−0.20
Peterson (1988)	1969–1984	16 (no to)	−1	0.65	2.04

*Significant at 0.05 level.
**Significant at 0.01 level.
Event day 0 = proxy mailing date announcing meeting with reincorporation vote.
nr: test statistic not reported.
BS (1989): Bradley and Schipani (1989).
Romano (1985) subsamples: m&a: reincorporations accompanied by acquisition programs; to: reincorporations accompanied by takeover defenses.
Wang (1995) subsamples: Del: reincorporations into Delaware; non-Del: reincorporations to states other than Delaware.
HL (1998): Heron and Lewellen (1998), subsamples: to: reincorporations accompanied by takeover defenses; ll: reincorporations accompanied by director liability limits; oth: reincorporations not accompanied by takeover defenses or director liability limits.
DL (1980): Dodd and Leftwich (1980).
Hyman (1979): calculates AR as difference in mean changes in stock price compared to S&P index.
NP (1989): Netter and Poulsen (1989), subsamples: Cal: reincorporations from California; non-Cal: reincorporations from states other than California.
Peterson (1988) subsamples: to: reincorporations accompanied by takeover defenses; no to: reincorporations not accompanied by takeover defenses.

Because reincorporations are typically accompanied by changes in business plans (Romano, 1985, p. 250), there is, however, a question whether the positive stock price effects are evidence of the market's assessment of the change in business plan rather than the change in domicile. The issue is whether there is a confounding effect, that muddies the interpretation of stock price effects, requiring a more probing examination

of the findings. To investigate whether the positive price effect was a function of investors' responses to other changes in business plan accompanying the reincorporation, Romano (1985) compared the returns of the firms in her sample grouped by the type of activity accompanying or motivating the reincorporation—engaging in a mergers and acquisitions (m&a) program, undertaking takeover defenses, and a miscellaneous set of other activities including reducing taxes. Although one might have expected the impact to vary across firms, with the antitakeover reincorporations experiencing negative returns, as prominent commentators have viewed takeover defenses as adverse to shareholders' interest (e.g., Easterbrook and Fischel, 1981), and the m&a reincorporations positive returns, as research finds firms experience significantly positive abnormal returns on announcing an m&a program (Schipper and Thompson, 1983), not only was the sign on both groups' abnormal return positive but there was also no significant difference across the groups (p. 272). This finding suggests that the significant positive returns upon reincorporation are due to investors' positive assessment of the change in legal regime, and not a confounding of the impact of reincorporating firms' other future projects.

In contrast to Romano's study, Heron and Lewellen (1998, pp. 557–559) find a different price reaction depending on whether the reincorporation is undertaken to limit directors' liability (positive) or to erect takeover defenses (negative). However, the event date they use that produces the result is problematic. The takeover defense firms' abnormal returns are significantly negative only on the shareholder meeting day (except for a further subdivision of the subset of firms adopting takeover defenses, 32 creating a new poison pill, compared to other takeover defense groupings of 45, 168, and 83 firms), and Brickley's (1986, pp. 346–347) investigation of the event study methodology found that, in contrast to random samples of proxy mailing dates, random samples of annual meeting dates—that is, a sample on which there is no *a priori* reason to find a significant price effect—produce significant abnormal returns.[14]

Firms sometimes propose reincorporation at special meetings, which might limit the relevance of Brickley's finding for at least a subset of Heron and Lewellen's sample. Still the theoretical basis for expecting a price effect on the meeting date is weak: the number of contested management proposals is exceedingly small so that minimal uncertainty exists over the voting outcome to be resolved on the meeting date that could affect the returns of a portfolio of firms on that date. But even if we assume that most of the Heron and Lewellen sample firms held special meetings for the reincorporation and that they can identify the new information available on the meeting date and construct a model of the market's expectation concerning passage of the proposal and how that expectation changed after the proxy mailing date, the negative return on the mailing date for firms adopting takeover defenses upon reincorporation is not probative for judging

[14] Brickley's (1986, pp. 347–348) explanation of the finding of abnormal returns on randomly selected meeting dates in contrast to mailing dates is that annual meeting dates are known in advance and often contain important management announcements (such as earnings forecasts), which can produce abnormal returns because "risk and expected returns can increase around predictable events likely to contain information."

state competition adversely. The defenses they examine could have been adopted by the reincorporating firms in their original home states and were not solely available to them in Delaware. In this regard, a negative price reaction could be interpreted by investors as disappointment that the reincorporating firms would not obtain the maximum benefit—facilitation of a takeover bid—from the domicile change (see Daines, 2001). But such a reaction should still have been incorporated into the stock price at the proxy mailing rather than meeting date.

In short, it may well be that the reincorporations accompanied by takeover defenses are exercises in managerialism, as Heron and Lewellen imply, but managers do not need to reincorporate in Delaware to adopt defenses. The only significant concern that could affect an assessment of state competition is the possibility that by bundling defenses into one vote on the reincorporation rather than as multiple votes on separate charter amendments, management is able to garner support for a proposal that it would not otherwise be able to obtain. The argument is that issue bundling coerces shareholders to accept value-decreasing takeover defenses in order to obtain the value-increasing effect of the new legal regime. While we think that it is improbable that shareholders would vote for the reincorporation if they would not have voted for the defenses separately, identification of such a concern, were it correct, would indicate that state competition, through which firms can move to Delaware, is in fact perceived favorably by investors, as it would mean that they consider the benefit of a Delaware domicile to be greater than the loss from a takeover defense. In accord with that inference, the abnormal returns of firms reincorporating in Delaware from California in the Netter and Poulsen (1989) study, which firms tended to be in the bundling category—adopting takeover defenses at the same time as reincorporating—are positive (only marginally significant at 10 percent) and no different from those of the firms in their sample that migrated from other states.[15] Consistent with that finding, and mitigating the bundling concern, the wealth effects of many of the charter-level defenses adopted by the firms in Heron and Lewellen's study (such as fair-price provisions) are inconsequential (see e.g., DeAngelo and Rice, 1983; Linn and McConnell, 1983; Jarrell and Poulsen, 1987).

Bebchuk (1992, pp. 1449–1450) contends, analogously to the issue bundling concern posed by Heron and Lewellen's data but more generally, that event studies are not probative on whether state competition benefits shareholders because state competition may produce some code provisions that are harmful to shareholders even if the overall package of statutory provisions is not, and hence we would not detect any statistically significant price effect upon reincorporation. However, the power of the premise that shareholders are being forced to choose between bundles of offsetting good and bad statutes depends on finding insignificant returns on reincorporation, which would imply that the codes are in equipoise between wealth-increasing and wealth-decreasing provisions. Yet, as noted above, event studies report significant positive stock price effects.

[15] In the time frame of the Netter and Poulsen study, and for roughly 2/3 of the sample period of the Heron and Lewellen study, California did not permit staggered boards; the law was changed in 1989.

Moreover, from the perspective of shareholders, and hence from the perspective of assessing the efficacy of the output of state competition, it is the net wealth effect of a code on investors that is important.

A related contention of Bebchuk, Cohen, and Ferrell (2002) is that the positive price effects of a reincorporation in Delaware are due to network effects unrelated to the content of the legal regime, that is, that investors value the presence of a stock of legal precedents, even though the code (and precedents thereunder) are adverse to their interest (that is, they favor managers over shareholders). The Bebchuk, Cohen, and Ferrell (2002) insight, which adopts the view of corporate law advanced by Klausner (1995), is a valuable contribution to this debate. However, this thesis would suggest that the value of the stock of precedents should be inversely correlated with the value of the substantive law that creates those precedents for investors, that is the substantive law must favor managers' over shareholders' interest. Otherwise a network effect is of no import for an evaluation of the price effects of state competition, because its positive price effect would be reinforcing, not offsetting, the wealth effect of the regime.

It is not likely that over time the positive value of certainty offered by a stock of precedents would continue to outweigh the negative value of the legal rules that make up that stock. Rather, if the substantive law harmed investor interests, we should observe over time, negative returns for domicile changes and some investor reaction to the situation.[16] For example, we should observe an increasing number of proxy proposals by institutional investors—who are informed about legal rules and seek the removal of defensive tactics to takeovers through the proxy process—to reincorporate out of Delaware, but we do not. From 1987–1994, of 2,042 proposals only 10 were directed at reincorporation (and of those most were proposing moving out of a state other than Delaware) compared to over 1000 directed at eliminating defensive tactics (Romano, 2002, p. 72). Or we should observe fewer incorporations in Delaware. But the proportion of newly public corporations (whose insiders bear the cost of a domicile choice in the price received for the newly issued shares to the public) domiciled in Delaware has increased over time (Daines, 2002; Bebchuk and Cohen, 2003).

Another class of event studies that provide data for the state competition debate are event studies of changes in Delaware law. Such studies are less reliable tests than studies of domicile switches of the wealth effects of state competition, however, for several reasons. First, reincorporations are more difficult for investors to anticipate and therefore easier to date for statistical testing than legislative changes. Second, reincorporations are firm-specific events, and hence the endogeneity of the event's occurrence is automatically controlled for by the composition of the test portfolio—it includes only firms

[16] To the extent that firms can contract around, or out of, legal rules, and that is the value of a stock of precedents, self-help avoidance of wealth-decreasing laws would appear to be an alternative explanation consistent with Bebchuk et al.'s hypothesis that positive network effects offset negative effects from the substantive law. But such a practice would also undermine that hypothesis because it would mean that the wealth-decreasing rules were of no import (that is, because the legal rules can be contracted around, they have no force and therefore do not adversely affect investors' wealth).

experiencing the event. This is not true for the enactment of statutes, which are applicable to all domestic corporations but may actually have divergent effects on different corporations. When their impact is examined for a portfolio of domestic firms that does not control for the potential heterogeneity across firms of the effect of the legal rule change, the test may simply aggregate offsetting effects and therefore not be able to identify any wealth effect.

State takeover statutes have been examined intensively, a legislative context in which Delaware is a laggard rather than the leader of competitive activity (Romano, 1993a, p. 59). In addition, one Delaware statutory change has been closely studied, enactment of a statute permitting firms to limit outside directors' liability for negligence. Other states soon followed suit with similar limited liability statutes, but the wealth effects of those enactments have not been studied.

The wealth effects of takeover statutes are less uniform than the reincorporation studies—there are findings of negative, positive and insignificant price effects (see Romano, 1993a, pp. 60–68). But the most comprehensive study, by Karpoff and Malatesta (1989), which has the largest sample size because it includes 40 statutes enacted in 26 states, finds that the statutes have a significant, albeit small, negative price effect on domestic corporations (−0.4 percent), when the event date is the earliest newspaper report of the legislation. They find no significant price effect when days on which specific legislative events occurred, such as bill introduction, final passage and signing into law, are used as the event dates. Much of the differences in the event study findings can be explained by the type of statute: statutes more likely to raise the cost of a bid tend to produce negative price reactions compared to statutes less likely to affect a bid (compare the results for disgorgement, business combination and control share acquisition statutes in the studies by Szewczyk and Tsetsekos, 1992; and Karpoff and Malatesta, 1989; with those for fair price and other constituency statutes in the Karpoff and Malatesta, 1989; and Romano, 1993b, studies).[17] Differences also depend upon the event interval chosen (compare Ryngaert and Netter, 1988; with Margotta, MacWilliams, and MacWilliams, 1990), which cautions against drawing strong conclusions from any one study without adequate justification of the researchers' interval choice.

Finally, differences may reflect methodological choices (see Karpoff and Malatesta, 1995, who show how the small size of Pennsylvania portfolio firms biases results that use standard market portfolios consisting of large firms, given the statute's enactment during a time period in which in terms of overall market performance, there was a negative small firm effect).

[17] The price impact of a statute may be related to the absence of firm-level defenses. When Karpoff and Malatesta's (1989) sample is divided according to whether the firm has antitakeover charter amendments or a poison pill, only the portfolio without defenses experiences a significantly negative effect on the event date (p. 308). However, not all studies find the same abnormal return pattern controlling for firm-level defenses (e.g., Romano, 1993b; Jahera and Pugh, 1991). We thus are hesitant to conclude that characteristic differences across firms in sample portfolios explain the variation in the studies' results.

Most important for the state competition debate, Karpoff and Malatesta (1989) find that Delaware's takeover statute had an insignificant stock price effect. A study by Jahera and Pugh (1991) finds a significantly positive effect of the Delaware statute on legislative event dates for some but not all of the excess returns models that they investigate; they also find no significant price reaction on the newspaper announcement dates (Karpoff and Malatesta do not provide information on the price effect of the Delaware statute on legislative event dates).[18] The inference from these studies is that the Delaware takeover statute, in contrast to other states' enactments, did not adversely affect the wealth of investors. Table 3 summarizes the results of the above-mentioned studies, and additional event studies evaluating the same statutes as those discussed in the text; for a more complete tabulation of takeover statute event study results see Table 4-1 in Romano (1993a).

The nonnegative impact of the Delaware takeover statute is a fact of itself favorable to an assessment of state competition because Delaware is the leading incorporation state, and this result indicates that its legislature is more likely to consider the interest of investors than legislatures in other states. Indeed, the findings on takeover statutes are the strongest (and sole) empirical evidence against the efficacy of state competition for charters: they suggest that states other than Delaware enact laws that do not benefit shareholders. Ironically, then, for adherents of Cary's position on state competition, the data cast Delaware in a positive light for investors. A fair conclusion from the takeover statute event studies is that for at least some firms in some states, legislative initiatives making takeovers more difficult were bad news (wealth-decreasing events) for investors.

Delaware's limited liability statute, in contrast to other states' takeover statutes, but like its own takeover statute, did not have a significant stock price effect (Bradley and Schipani, 1989; Janjigian and Bolster, 1990; Romano, 1990). Delaware firms did experience significant negative returns on the effective date of the statute, which was two weeks after its enactment and two months after the first legislative event date, the day when the corporate law section council of the Delaware Bar Association approved the provision (for corporation code revisions, the Delaware legislature acts upon recommendations of the corporate bar).[19] But the statute's effective date is not a meaningful

[18] Contrary to Karpoff and Malatesta's finding on firm defenses that it is the firms without defenses that experience significant negative returns, Jahera and Pugh find that for Delaware firms, those with antitakeover charter defenses experienced a negative price effect, and those without them a positive price effect, on several event dates: the cumulated effect is insignificant for the former group and significant for the latter group only at 10 percent.

[19] Romano (1990) finds a significantly negative return on the day after a press report of the enactment of the bill and the day after the Senate passed the bill, whereas Janjigian and Bolster find a significantly negative return on the day of the press report and the day the Senate passed the bill, and a significantly positive return the day after the press report of the bar committee's approval (in Romano's study the return on that date is insignificantly positive). These suggest an adverse wealth effect from the statute. However, because the cumulated abnormal return over all of the event dates is insignificant, in both studies, we conclude, along with the study authors, that the statute did not decrease shareholder wealth. Note that Bradley and Schipani do not

Table 3
Event studies of takeover statute enactments

Study	Statute(s) studied	Sample size	Announcement window: (event days)	Announcement return (%)	Z-statistic
KM (1989)	40 statutes, 26 states, 1982–1987	1505	(−1, 0)	−0.29	−2.43**
KM (1989)	38 statutes, 26 states (no to)	1107	(−1, 0)	−0.39	−2.54**
KM (1989)	33 statutes, 23 states (to)	368	(−1, 0)	−0.13	−0.87
Mahla (1991)	49 statutes, 30 states, 1983–1989	678	(0)	−0.12	−1.85
PJ (1990)	5 statutes, 1 vetoed bill, 4 states	245	(0, +1)	−0.46	−1.62
KM (1989)	11 BC statutes	1030	(−1, 0)	−0.47	−2.70**
Mahla (1991)	BC statutes	248	(0)	−0.24	−2.35*
KM (1989)	Del BC statute, 1987	nr	(−1, 0)	−0.44	−1.10
JP (1991)	Del BC statute, 1987	920	(0, +1)	−0.09	−0.19
PJ (1990)	Ind BC statute, 1986	15	(0, +1)	−0.94	−1.12
Broner (1987)	NJ BC statute, 1986	51	(−1, +1)	−0.55	−1.13
PJ (1990)	NJ BC statute, 1986	26	(0, +1)	0.48	0.71
Schumann (1988)	NY BC statute, 1985	94	(−1, +1)	−0.96	−2.37*
KM (1989)	NY BC statute, 1985	nr	(−1, 0)	−0.22	−0.60
PJ (1990)	NY BC statute, 1985	72	(0, +1)	−0.72	−1.71
KM (1989)	12 CSA statutes	271	(−1, 0)	−0.01	−0.89
Mahla (1991)	CSA statutes	236	(0)	−0.017	0.18
KM (1989)	Ind CSA statute, 1986	nr	(−1, 0)	−2.14	−3.46**
PJ (1990)	Ind CSA statute, 1986	15	(0, +1)	−1.8	−2.15*
SW (1990)	Ind CSA statute, 1986	19	(0)	−5.91	1.97
Romano (1987)	Mo CSA statute, 1984	14	(−1, +1)	−0.01	−0.72
PJ (1990)	OH CSA statute, 1986	45	(0, +1)	−0.35	−0.67
KM (1995)	Penn DG, 1990	57	(0, +1)	−1.43	−2.89+
KM (1995)	Penn DG 1990	28 (no to)	(0, +1)	−2.33	−2.78+
KM (1995)	Penn DG 1990	29 (to)	(0, +1)	−0.56	−1.21
Margotta (1991)	Penn DG, 1990	55	(0)	−0.6	−2.47*
ST (1992)	Penn DG, 1990	56	(−1, +1)	−3.33	−4.80**
ST (1992)	Penn DG, 1990	44 (no to)	(−1, +1)	−3.94	−5.06**
ST (1992)	Penn DG, 1990	12 (to)	(−1, +1)	−1.11	−0.65
KM (1989)	11 FP statutes	329	(−1, 0)	−0.27	−1.30

event date because there was no new information released on it—the statute's enactment was well-publicized and there was no uncertainty regarding whether the statute would become effective on the stated date.

find significant abnormal returns on any legislative event dates; they find a significantly negative cumulated return for an interval measured a week around the effective date of the statute, which is after the statute was enacted.

Table 3
(*Continued*)

Study	Statute(s) studied	Sample size	Announcement window: (event days)	Announcement return (%)	Z-statistic
Mahla (1991)	FP statutes	74	(0)	0.06	−0.05
Romano (1993b)	25 OC statutes (pre-1991)	361	(0)	0.02	0.30
ASM (1997)	IN, NY, and OH OC statutes (pre-1993)	318	(0,1)	−0.33	v.nr
RN (1988)	Oh OC/PP statute, 1986	37	(−1, +1)	−2.08	−2.18*
MMM (1990)	Oh OC/PP Statute, 1986	53	(−1, +5)	1.43	1.69

*Significant at 0.05 level.
**Significant at 0.01 level.
+ significance level not specified. Event is first press report, unless otherwise indicated below.
v.nr: value not reported for Z-statistic; significance level is reported.
BC: Business Combination statute; CSA: Control Share Acquisition statute; FP: Fair Price statute; OC: Other Constitutency statute; PP: Poison Pill statute; Penn DG: Pennsylvania takeover statute including disgorgement, other constituency, control share acquisition and labor protection provisions.
KM (1989): Karpoff and Malatesta (1989); they report insignificant abnormal returns on legislative event dates; subsamples: to = firms with takeover defenses when statute adopted; no to = firms without takeover defenses when statute adopted.
Mahla (1991): event is introduction of bill.
PJ (1990): Pugh and Jahera (1990), event is introduction of bill.
JP (1991): Jahera and Pugh (1991); event is 8 legislative events; significant positive return reported using excess returns model.
Broner (1987): event is committee release of bill.
SW (1990): Sidak and Woodward (1990), 14 legislative events.
Romano (1987): event is introduction of bill.
KM (1995): Karpoff and Malatesta (1995); insignificant when cumulated over legislative events; subsamples: to = firms with takeover defenses when statute adopted; no to = firms without takeover defenses when statute adopted.
ST (1992): Szewczyk and Tsetsekos (1992), event is measured over 4 legislative events; subsamples: to = firms with takeover charter defenses when statute adopted; no to = firms with no takeover charter defenses when statute adopted.
Romano (1993b): 3 legislative events.
ASM (1997): Alexander, Spivey, and Marr (1997); subsample: to = firms with takeover defenses when statute adopted.
RN (1988): Ryngaert and Netter (1988), event is legislative action.
MMM (1990): Margotta, MacWilliams, and MacWilliams (1990), event is legislative action.

Because the coverage of the limited liability statute was optional, one explanation of the finding of insignificance, besides the issue raised by any event study of legislation, imprecision as to the dating of events, is that the effect of the statute would not be incorporated into stock prices until investors determined whether or not their firm would elect to be covered. The abnormal returns experienced by firms opting into the statute

vary, however, depending on the event window examined or the portfolio of firms. One study (Romano, 1990) finds significant positive returns over two, three and five day intervals, and insignificant returns over a seven day interval; another study (Bradley and Schipani, 1989) finds significant negative returns over a seven day interval; two studies (Janjigian and Bolster, 1990; Netter and Poulsen, 1989) find insignificant returns over a variety of time intervals; and a final study (Brook and Rao, 1994), which uses a four day interval, finds insignificant positive returns for its full sample of 120 firms and positive abnormal returns for poorly performing firms. Brook and Rao's explanation of their finding is that shareholders of poorly performing firms value limited liability provisions more highly than shareholders of other firms because it is more important for such firms to "attract and retain" the services of high quality outside directors.[20]

These findings suggest that the limited liability statute did not adversely affect shareholders. Providing further support for this interpretation of the data is the fact that shareholders vote overwhelmingly to opt into the limited liability statute. It is, of course, possible that shareholders vote for management-sponsored proposals that adversely affect firm value; see Bhagat and Jefferis (1991). But the Bhagat and Jefferis study investigated management-sponsored proposals related to takeover defenses, proposals that institutional investors have vigorously opposed in the same time period in which they have supported the limited liability provisions. Consistent with this distinction, institutional investors have also sponsored proposals to overturn takeover defenses implemented by management but not to overturn limited liability charter provisions. In the reverse legislative situation, a takeover statute that shareholders did not wish to have applied to their firm (the Pennsylvania disgorgement statute), institutional investors successfully pressured managers to opt out of the statute's coverage (Romano, 1993a, pp. 68–69). The event studies of that Pennsylvania statute report a negative wealth effect, in contrast to those of the limited liability statute.

Finally, in addition to the event studies of legislation, there have been event studies of judicial decisions (Bradley and Schipani, 1989; Kamma, Weintrop, and Wier, 1988; Ryngaert, 1988; Weiss and White, 1987). Because courts play an important role in Delaware's market position (e.g., Romano, 1993a, pp. 39–41), determining whether investors benefit from judicial decisions could proxy for determining whether they benefit from state competition. However, judicial decisions are not clearly "events," except for the litigants for whom a decision effects a wealth transfer. Decisions in corporate law cases may not effect firms other than the litigants because other firms and investors will often be able to contract around a rule and recalibrate costs and benefits. Judicial decisions are therefore only of limited value as subjects for the event study methodology—we can use the methodology to learn how a specific decision affects the parties, but it may not be useful for analyzing the decision's impact on nonlitigants.

Further complicating event studies of judicial decisions is the interaction between the court and state legislature in Delaware, which is a byproduct of the competition for

[20] Although these firms also had a smaller percentage of outside directors, there was no relation between the number of outside directors and the value of a limited liability provision (Brook and Rao, 1994, p. 495).

charters. A judicial decision with a significant adverse impact on firms stands a good chance of being overturned by the Delaware legislature: the limited liability statute, for instance, was a reaction to a judicial opinion holding outside directors liable for accepting too hastily a takeover premium (Romano, 1990). Since investors can anticipate the legislature's response to a judicial decision that is adverse to their interest, one would not expect to find a negative stock price effect for a portfolio of Delaware firms after a wealth-decreasing decision.

Not surprisingly, event studies of judicial decisions find insignificant price effects for portfolios of Delaware firms (Bradley and Schipani, 1989; Weiss and White, 1987). Judicial decisions produce significant abnormal returns to the litigants, however, and, when the decisions uphold (or invalidate) a specific takeover defense, to concurrent takeover targets as well (Kamma, Weintrop, and Wier, 1988; Ryngaert, 1988). The use of the methodology in these latter studies is equivalent to that of the litigation studies discussed in Section 4.1.

An alternative methodological approach to event studies for investigating the impact of state competition is to compare the performance of firms incorporated in Delaware to those in other states. Three studies have examined whether firms' performance improves after a change in domicile (Baysinger and Butler, 1985; Romano, 1996; Wang, 1995). These studies compare accounting measures of performance (return on equity and earnings before interest and taxes) for reincorporating firms before and after the domicile change or compared to non-reincorporating firms. They find no significant difference for any comparison, with the exception that Wang (1995) finds the change in earnings over the year after the domicile change was higher for firms reincorporating in Delaware than for firms reincorporating in other states. One interpretation of the absence of significant performance differences is that firms select the domicile that optimizes their future performance; the studies do not control for the fact that the domicile choice is endogenous, and therefore could affect performance, making cross-sectional comparisons difficult (the issue raised in Section 3). Wang's finding of higher performance of firms moving to Delaware is consistent with the event study data (the positive price effects indicate that investors anticipated increased earnings) and with the view that the law matters, but that interpretation can not be distinguished from a self-selection effect (higher quality firms chose to move to Delaware).

A fourth study by Daines (2001) compares the performance of firms incorporated in Delaware to those incorporated elsewhere, over the period 1981–1996. Daines uses as the performance measure, Tobin's Q, which is the ratio of a firm's market value to the replacement cost of its assets and conventionally interpreted to proxy for a firm's investment or growth opportunities. His insight is that opportunities added by corporate law rules can be considered a component of the value measured by Tobin's Q. Daines finds that Delaware firms have significantly higher Tobin's Q values, controlling for investment opportunities and other variables known to affect Tobin's Q, such as a firm's business diversification, as well as ownership, in the pooled sample and in 12 of 16 years when the model is estimated separately by year. In addition to the controls, a variety of robustness checks are performed (such as investigating, and finding unchanged, the

results for subsamples of mature firms, IPO firms of different quality, and excluding reincorporating firms), to ensure, as best as possible, that the identified effect is due to the legal regime, rather than selection (higher quality firms incorporate in Delaware).

Subramanian (2004) drops the first ten years from Daines' study (years 1981–1990) and adds four later years (1997–2000) and reports that the advantage of being incorporated in Delaware has decreased (a market value higher by 2.8 percent in his sample compared to the estimate of 5 percent in Daines' sample's last year) and the difference in Tobin's Q is not significant in the later years of Subramanian's sample.[21]

The absence of a difference in Tobin's Q in those years suggests that either other states had "caught up" to Delaware, by amending their codes to eliminate major differences, which would reduce the differential value of a Delaware incorporation (similar to how performance levels of low productivity nations converged to that of high productivity ones after World War II through their ability to learn from the leader through technology transfer, as detailed in Baumol, Blackman, and Wolff, 1989) or that the decline in the takeover market in the late 1990s with the economic downturn of that period was reflected in firms' Tobin's Q values.

It should be noted that there is a well-established practice of using Tobin's Q to measure performance (see e.g., Morck, Shleifer, and Vishny, 1988), but there are some distinct methodological issues regarding its use. Namely, there are problems regarding this performance measure especially when studying its relation to ownership. First, the denominator does not include the investments a firm may have made in intangible assets. If a firm has a higher fraction of its assets as intangibles, and if monitoring intangible assets is more difficult for the shareholders, then the shareholders are likely to require a higher level of managerial ownership to align the incentives. Since the firm has a higher fraction of its assets as intangibles it will have a higher Q since the numerator will impound the present value of the cashflows generated by the intangible assets, but the denominator, under current accounting conventions, will not include the replacement value of these intangible assets. These intangible assets will generate a positive correlation between ownership and performance, but this relation is spurious not causal. Second, a higher Q might be reflective of greater market power. Shareholders, cognizant of the fact that this market power shields the management to a greater degree from the discipline of the product market, will require managers of such a company to own more stock. Greater managerial ownership will tend to align managers' incentives better and offset the effect of the reduced discipline of the product market. In the above scenario we would again observe a spurious relation between performance as measured by Q and managerial ownership. Finally, as suggested by Fershtman and

[21] One important difference in sample construction should be noted that could account for the lower Tobin's Q values in the Subramanian study. Daines only includes firms that have 5 years of data over his 16 year sample period, whereas Subramanian includes all firms, without providing any distributional information on how many observations lack 5 years of data, such as, how many firms were in the sample for one or two years and then went bankrupt. Such firms are likely to have low Tobin's Q values, and given their short life span, comparisons containing such firms would not be informative on the effects of a Delaware domicile.

Judd (1987), shareholders may induce the managers (via greater share ownership) to engage in collusive behavior and generate market power. In this scenario we would also observe a spurious relation between performance as measured by Q and managerial ownership. Bhagat and Black (2002) have accordingly suggested using a variety of performance measures, especially accounting performance measures, in evaluating the relation among corporate governance, ownership and performance.

Finally, four recent studies have compared the domicile and physical location of firms to draw insights into state competition that may not be apparent when examining reincorporations (Bebchuk and Cohen, 2003; Daines, 2002; Kahan, 2006; Subramanian, 2002).[22] These studies have contributed new insight into the reincorporation debate, and we therefore discuss the analysis in greater detail than other studies reviewed in the chapter. Because the approach and results in the Bebchuk and Cohen and the Subramanian studies are virtually identical, comparing the state corporation codes of firms' current domiciles to that of their current physical location on one state-law dimension, whether the state code has a set of specific takeover statutes, we focus our discussion on only one of the two, the Bebchuk and Cohen study, which contains more controls in the regression model than the Subramanian study. Those two studies seek to test the hypothesis that takeover statutes dictate firms' choice of domicile; evidence consistent with the hypothesis would support the view that state competition for charters is a race to the bottom.

Bebchuk and Cohen consider a logit model of whether a firm is incorporated and physically located in the same state, which they term a state's "retention rate,"[23] as a function of the following: firm-level (such as sales and number of employees) and state-level variables (such as population, per capita income and number of firms located in-state), along with two state corporate law variables, a dummy variable for whether a state has a version of the Model Business Corporation Act (a model statute, drafted and periodically revised by the business law section of the American Bar Association, for the purpose of providing a template for states' corporation codes),[24] and a takeover

[22] We do not discuss a fifth paper, Ferris, Lawless, and Noronha (2004), because it investigates domicile choices using a poorly specified index of state corporate law regimes that combines takeover statutes with laws permitting flexibility in acquisitive transactions. It is not only theoretically incorrect to classify all of these laws as equivalently "pro-manager" or anti-shareholder in perspective, but also, the results in Kahan (2006) indicate that it is dubious methodologically to group the statutes in such a fashion without distinction.

[23] There is a methodological problem with the model specification: the estimation did not adjust for the correlation across observations arising from the fact that the error terms of firms located in the same states are correlated. The specification in Kahan (2006) controls for that problem, as do some models estimated in Daines (2002), and those studies report different statistical findings.

[24] Other means of modeling the significance of the model act for domicile choice, which are not employed by Bebchuk and Cohen, would be to adjust for the differences across states in the form of model act adoption, which varies considerably, such as, by which revision of the act has been adopted, whether important sections of a model act revision are adopted (since the act in some instances contains alternative provisions and is often not adopted in full), and, most importantly, the speed with which a model act or update has been enacted by the state.

statute index, constructed by adding up how many statutes a state has from among five specific types of takeover statutes. Two state-level independent variables are significant: the takeover statute index and an interaction term between a firm's sales and the population of the state in which it is physically located, as are dummy variables for geographical region. The authors conclude from these results that takeover statutes dictate the choice of domicile.

However, the choice of domicile is not likely to be unidimensional with respect to the content of a corporation code. In fact, other characteristics of state corporate law, besides takeover statutes, are known to be considered important by firms that have switched domicile, such as, the responsiveness with which a state updates its code (that is, the speed with which corporate law innovations are adopted compared to adoption by other states), the quality of its judiciary in corporate law cases and correspondingly, the degree of certainty provided by the legal regime (see Romano, 1985). It is highly probable that the presence (or absence) of those characteristics would similarly affect a state's retention rate (its attractiveness to local firms that might otherwise change domicile). California, for instance, has no takeover statutes, but it also has several other undesirable features from corporations' perspective, such as more unpredictable state courts with little expertise on corporate law (Romano, 1985) and an active plaintiff's bar that has used the state referendum process in an effort to increase the liability of directors and officers.

In addition, firm characteristics not included in the Bebchuk and Cohen model have been found by other researchers to affect domicile choice. For example, studies have identified firm characteristics associated with Delaware domiciles (e.g., Romano, 1985, finds that Delaware firms are more likely to undertake acquisitions or be the subject of shareholder litigation, and Baysinger and Butler, 1985, find that block ownership is inversely correlated with a Delaware domicile). These firm-level characteristics—ownership and propensity to make acquisitions—as well as propensity to be acquired, are likely to affect the relevance of takeover protection to the choice of domicile, as well as the domicile choice itself.

Second, the choice of a firm's physical location is as likely an endogenous choice for a publicly traded firm as is its statutory domicile. There are trends in firms' movement of physical location in relation to local economic and business conditions (the movement, for instance, from high tax and labor cost states in the North to the South); these could generate inferential problems from common factors affecting differences in domicile and headquarters from differences in economic conditions that could be spuriously correlated with differences in the number of takeover statutes across states. This issue would be further complicated by firms' changing location more than once.[25] Domicile

[25] For an example of the inference problem regarding the importance of takeover statutes raised by a firm's multiple changes in headquarters while retaining its original headquarters' state as domicile, see Romano (2002, p. 100). Because Bebchuk and Cohen's data are stock data from one year, they cannot distinguish physical moves from reincorporations in the sample.

choices may well be sticky over multiple physical moves because legal counsel's investment in human capital would be depreciated with frequent change (see Romano, 1985), a factor having nothing to do with the presence of a takeover statute.

Both Daines (2002) and Kahan (2006) examine the domicile choice, compared to the location, of firms that went public over 1978–1997, and 1990–2002, respectively. Daines controls for additional features of state corporate law (proxies for the quality of a state's law, including the presence of network effects, certain substantive provisions, and responsiveness) and firm-level characteristics (such as controlling shareholder ownership), expected to impact on the choice of domicile. He finds, contrary to Bebchuk and Cohen, that takeover statutes have no significant explanatory effect on the domicile choice. The vast majority of IPO firms choose Delaware or the state in which they are located, but they are not more likely to incorporate in the state of location rather than in Delaware when the home state has more takeover statutes. He further finds that some of the variables for the quality of the legal regime, such as being a Model Business Corporation Act state,[26] explain the other states' retention of firms.

Kahan also controls for key state corporate law features of interest to firms, the quality of the judiciary and the flexibility of code provisions unrelated to takeover defenses, in addition to the takeover statutes included in Bebchuk and Cohen's index and his own classification of takeover statutes. In contrast to the other studies, Kahan uses states' aggregate retention rates (the ratio of the number of firms both physically located and incorporated in a state to all of the firms physically located in the state) as the dependent variable, rather than the underlying firm observations. Kahan finds, in contrast to Bebchuk and Cohen, and paralleling Daines, that takeover statutes (whether using individual statutes, his own or Bebchuk and Cohen's index) are not related to states' retention rates, but rather, the states with more flexible statutes (unrelated to takeover defenses) and higher quality judicial systems have higher retention rates.

It is possible that the difference between the findings of the Daines and Kahan studies and Bebchuk and Cohen's findings is that the decision process of IPO firms differs from established firms. However, Kahan reports that the key result—statutory flexibility is significantly related to domicile retention while takeover statutes are not—holds up when the model is estimated using retention rates derived from the stock domicile data in the Bebchuk and Cohen study. This is not surprising because, as Daines (2002, pp. 1569–1570) notes, most firms do not change domicile after their IPO and there is considerable overlap in the data sets of the studies. This suggests that the difference in the results across the studies is, most likely, due to the difference in model

[26] Daines (2002, p. 1596) notes two explanations of the significance of the model act, which are not distinguishable in his model: IPO firms might value a state's adopting the act for its substantive content (the model act rules), or for a network effect (the model act provides a stock of precedents and commentary of value for firms' business planning). He notes that another variable measuring the quality of a legal regime, an index of a state's responsiveness to legal innovations (as measured by Romano, 1985) was significant in some model specifications.

specification—including state corporation code features of importance to firms in addition to takeover statutes—and not due to the difference in datasets. Daines further finds that a highly significant firm-level factor in the domicile decision of IPO firms is whether the firm is advised by a local or national (offices in numerous states) law firm: firms are more likely to incorporate in Delaware when advised by a national law firm. The explanatory significance of lawyers in the incorporation choice of the IPO firms in Daines' study is in keeping with the important role that lawyers play in the development of corporate law (Macey and Miller, 1987). It is also consistent with survey data in Romano (1985, p. 275) that indicated that outside counsel were more influential in firms' decisions to reincorporate in Delaware than in other states.

Daines interprets the finding that domicile choices differ according to whether the law firm is a local or national firm as suggesting that lawyers are advancing their own interest at the expense of their clients in the choice of domicile. Daines' hypothesis has two prongs: local lawyers are unable to advise a client on Delaware law (or better able to advise the client on local law) and, in light of his earlier finding that Delaware firms have higher Tobin's Q values, a Delaware domicile would be more wealth-enhancing than retaining the local domicile upon going public. Thus, he contends (p. 1595) that if a local lawyer does not advise such a move, it suggests the lawyer is benefiting himself—avoiding competition—at the client's expense (unless the lawyer does not understand the benefits of a Delaware domicile). That may well be true. An alternative possibility is that firms with local counsel are not likely to be firms that would benefit from a Delaware domicile, which, requiring higher franchise fees, is worth the expense only for firms anticipating future transactions that benefit from Delaware's code (see Romano, 1985), or that they are firms whose short-term profitability and long-term viability is at high risk, which would also make it desirable to avoid the greater expense of a Delaware domicile (which would include the use of non-local counsel).

Daines controls for the endogeneity of choice of counsel (a firm chooses a national law firm for some other reason that would also make it appropriate to reincorporate in Delaware) by estimating a two-stage regression for domicile choice in which the choice of law firm is modeled in the first stage. That specification includes a variable for subsequent acquisitions, which relates to the first possibility, the undertaking of future transactions that make a Delaware domicile attractive as well as a more experienced (hence national) law firm, but the specification does not control for characteristics of firms (such as future growth) related to the second possibility, lower expectations of profitability, that could affect the choice of a local domicile as well as the choice of local counsel (because the firm would not be willing or able to take on the additional expense of national counsel, as well as the additional fees for a Delaware domicile). Moreover, as local lawyers will have influence on the content of their states' corporation code, they can recommend provisions to mitigate disadvantages of an in-state domicile, including the adoption of rules to compensate for a less-experienced judiciary (see Romano, 2002, p. 87). The implication for an evaluation of the product of state competition from local lawyers' influence on the choice of domicile is thus ambiguous. The competitiveness of the legal profession, and the relation between inhouse and outside counsel with respect

to domicile choice, undoubtedly bear on the question. This is an area where further empirical work would be fruitful.

4.3. Empirical research on takeovers

4.3.1. The role of event studies in public policy toward takeovers

There was an intellectual revolution in corporate law scholarship in the 1980s with the introduction of financial economics and the economics of organization (e.g., Winter, 1993). Equally important, corporate law scholarship tends to follow deals, and there was a burst in new acquisitive activity at that time and consequently, corporate law became one of the more active and sophisticated fields of interdisciplinary legal scholarship.

Event studies became an important source of information with which to ground policy recommendations in the new context of hostile leveraged bids. The explosion in acquisitions, which occurred shortly after the development of modern finance theory, of which the event study technique is a spinoff, created a cottage industry of event studies. There was a plethora of studies of the price effects of acquisitions and review articles were repeatedly updated in order to keep up with the literature (e.g., Jensen and Ruback, 1980; Jarrell, Brickley, and Netter, 1988; Andrade, Mitchell, and Stafford, 2001). These studies highlighted that there were uniformly large and significant positive price effects for shareholders of targets. There is also consensus in the literature that, on average, bidding shareholders do not experience any significant wealth effect upon announcement of such transactions. Depending on the sample period and sample considered, studies document average bidder returns that cover the range from positive, economically small and statistically insignificant, to negative, economically small and statistically insignificant. Studies that have aggregated the wealth effects of both the target and bidder firms find, however, that despite the lower returns to the generally larger-sized bidders, the combined target and bidder return is positive (e.g., Andrade, Mitchell, and Stafford, 2001; Bhagat et al., 2005; Bradley, Desai, and Kim, 1988; Kaplan and Weisbach, 1992).

Concern has also been raised on the impact of takeovers on other stakeholders, notably, employees, customers and suppliers (see Bhagat, Shleifer, and Vishny, 1990; Kim and Singal, 1993; and Akhavein, Berger, and Humphrey, 1997). A policy-relevant question is whether the large positive returns to target shareholders are offset by negative (or non-positive) returns to employees, customers and suppliers. Several studies have attempted to measure the losses to these non-shareholder interests and the average effect is generally small and often statistically insignificant, in striking contrast to the significantly larger average target shareholder gain (see, e.g., Asquith and Wizman, 1990; Dennis and McConnell, 1986; Marais, Schipper, and Smith, 1989; Pontiff, Shleifer, and Weisbach, 1990; Rosett, 1990). We are not aware, however, of any study that has attempted to address the question with a consistent sample. A study that considers the impact of a sample of takeovers on target and bidder shareholders and bondholders, employees, customers and suppliers would be a valuable contribution to this literature.

Analogous to the shifting sentiment on state competition, the conclusion from the event study research regarding the benefits of takeovers for target shareholders led commentators and policymakers alike to conclude that takeovers should be encouraged rather than obstructed (e.g., Easterbrook and Fischel, 1991, pp. 175–205; Council of Economic Advisors, 1985, p. 215). The Delaware courts took note, tightening the fiduciary standard applicable to takeover defenses (*Unocal Corp. v. Mesa Petroleum Co.*, 1985).

The Delaware courts did not, however, go as far as the position advocated by some prominent commentators that all defenses should be banned (e.g., Easterbrook and Fischel, 1981). Indeed, they eventually adopted an approach that provided managers with substantial discretion to react to a takeover as long as the bid is not precluded (*Unitrin v. American General Corp.*, 1995). This restrained, fact-intensive judicial approach is, in fact, consistent with the inconclusive empirical evidence on the efficacy of defenses, despite legal commentators' support for more active judicial intervention. The event study literature does not uniformly find that the adoption of defenses produces negative price effects. Stock price reactions vary not only with the type of defense, but also with the type of firm. For example, adoption of golden parachutes produces positive price effects (Lambert and Larcker, 1985), elimination of cumulative voting produces negative ones (Bhagat and Brickley, 1984), and the effects of poison pills vary, being negative in the early to mid 1980s (e.g., Ryngaert, 1988) and insignificant in later years (e.g., Comment and Schwert, 1995).[27] (The difference in the findings on poison pills may be due to investors' having anticipated that firms would adopt poison pills by the later years of the Comment and Schwert study.) In addition, Brickley, Coles, and Terry (1994) found positive price effects for poison pill adoptions by firms with independent boards, and Datta and Iskandar-Datta (1996) find pill adoptions produce insignificant effects except for firms subject to a takeover bid, for which the price effect is negative. Bhagat and Jefferis (1991) argue that the earlier studies find conflicting or insignificant results since they do not control for the anticipation of the antitakeover proposal. After controlling for this anticipation effect they find a statistically negative 1 percent return for antitakeover charter amendments.

The Delaware courts moved to an increasingly restrained approach to managerial resistance over the time frame in which the results of empirical research were becoming

[27] Golden parachutes are extremely lucrative severance pay contacts for top management that are triggered by a change of control; cumulative voting permits the aggregation of the votes to be cast in the election of directors on individual candidates, facilitating the representation of minority blockholders on the board; and poison pills are shareholder rights plans that provide the holder with the right to purchase stock in the issuing company at a discount on either the announcement of a takeover bid, or the acquisition of a specified percentage of stock, if those transactions have not been approved by the board of directors. Poison pill plans typically include the right to flip over into shares of the acquirer at the same discount should the issuer of the rights plan not survive a merger and they permit the board to redeem the rights for a trivial price before the rights have been triggered by an actual stock purchase. Golden parachutes and poison pills therefore make an acquisition more expensive for an unwanted bidder; cumulative voting may be useful to a bidder with a toehold, making it easier to obtain representation on the board and thereby gain acceptance of its bid.

increasingly ambiguous (that is, as the range in the findings of event studies of defensive tactics increased). This may not have been a conscious reaction to the empirical literature (the judicial shift began before many of the discrepant findings were in, although the later studies presumably were demonstrating systematically the anecdotal and intuitional sense practitioners already had of defenses' effects).[28] Further evidence that the courts' shift was not simply a Cary-predicted tendency of the Delaware judiciary to favor managers over shareholders is data suggesting that the adoption of defenses (firm- or state-level) did not decrease the number of takeovers (e.g., Comment and Schwert, 1995; but for contrary data see Hackl and Testani, 1988; Pound, 1987), though the data cannot, of course, ever fully satisfactorily answer the counterfactual, what would the rate have been in the absence of defenses?

A more troubling issue for corporate law was presented by the event study results regarding the stock price of acquirers. Event studies indicated a change in acquiring firms' abnormal returns from positive or insignificant to negative from the 1970s into the 1990s, paralleling the increasing use of defensive tactics to encourage auctions (e.g., Jarrell, Brickley, and Netter, 1988). As corporate law is directed to the shareholders of targets rather than bidders, the owners least likely to benefit from an acquisition as the decade progressed—the shareholders of the acquirers—did not have the opportunity for legal recourse. Courts did not change their traditional response deferring to management on acquisitions compared to the defensive tactic setting where the conflict of interest is more clear-cut. But because even commentators concerned about this issue were divided on whether there ought to be a legal response (compare, e.g., Dent, 1986 with Coffee, 1984; and Black, 1989, pp. 651–652), legislatures' and courts' maintenance of the status quo is unexceptionable.

The uniformity in the empirical findings on takeovers for target shareholders also affected interpretation of the mandate of the securities laws. The Securities and Exchange Commission (SEC) issued rules to overturn defensive tactics (e.g., Securities Exchange Act Rules 13e-4(f)(8), prohibiting selective self-tenders, and 19c-4, requiring one share one vote), although the federal courts did not always find it had authority to do so (*Business Roundtable v. SEC*, 1990, overturning rule 19c-4). Over 20 years earlier, the agency had successfully lobbied to advantage incumbent managers over bidders in the enactment of the Williams Act. It would be fair to say that the transformation in perspective on hostile bids was not simply a function of a change in agency personnel, but was caused by a more diffuse shift in attitude toward bids that was, no doubt, in part influenced by the event study literature demonstrating the benefits to target shareholders of takeovers.

[28] Gordon (1991) contends that in making the doctrinal shift providing managers greater leeway to block a takeover, the Delaware Court ignored the empirical literature that takeovers benefited shareholders through higher premiums, and he ascribes the move to the Court's mirroring a shifted public opinion against hostile acquisitions and a "money culture," reflected in increasing negative press, including Hollywood films, expressing antagonism toward such transactions, at the time.

The event study findings of the positive impact of takeovers on targets also formed the backdrop for the Supreme Court's decision in *Basic v. Levinson* (1988), which held that merger negotiations were sufficiently material to investors that disclosure could be required prior to the firms' reaching an agreement in principle, a bright-line standard that several appeals courts had adopted.[29] The Court's drive to disclose such information as early as possible is an acknowledgment of the significance of the information, which was underscored by the salience of the value of bids as measured by event studies. In reaching this conclusion, the Court rejected the view of the importance of maintaining secrecy until a firm agreement was reached that had been adopted by some appeals courts, stating that the view that secrecy "maximize[d] shareholder wealth" was "at least disputed as a matter of theory and empirical research" (p. 235). Although the Court did not specifically cite the economic literature on takeovers, it is plausible that the event studies detailing the benefits to shareholders of takeovers had an impact on its decision-making as the opinion evidences an awareness of the finance literature.

4.3.2. The relation between takeovers, governance and performance

Several studies have investigated the relation between takeovers, firms' corporate governance mechanisms and performance. We report selected results from this literature, but caution the reader at the outset that because many of the papers discussed in the section do not estimate the effects of these variables simultaneously, the estimates may well be biased and the results may not hold up were the models reestimated to take the endogeneity of the variables into account.

Martin and McConnell (1991) study performance prior, to and managerial turnover subsequent to, successful tender offer-takeovers. They find that takeover targets are from industries that are performing well relative to the market, and targets of disciplinary takeovers are performing poorly within their industry. During the year subsequent to the takeover they document a rate of management turnover of 42 percent compared to an annual rate of about 10 percent in the five-year period prior to the tender offer. Furtado and Karan (1990) review the literature on management turnover and conclude that turnover increases after control contests and after poor performance.

[29] The Supreme Court decision in *Basic v. Levinson* had an even more profound impact on the conduct of securities litigation than it had on acquisition negotiations. It articulated a doctrine, known as the "fraud on the market" theory, that permits plaintiffs to establish reliance, a necessary component of securities fraud, by reference to the integrity of the market price rather than by evidence that they saw or heard misleading information from the defendant. This doctrine is an acceptance of the semi-strong form market efficiency hypothesis, and represents, undoubtedly, one of the high points in the impact of finance theory on public policy. Finance theory, and more particularly event studies, have numerous other uses in securities litigation, uses which expanded after the *Basic* decision. As discussed in Mitchell and Netter (1994), the SEC uses the event study methodology in its enforcement of insider trading to determine the materiality of information and to calculate the profits that an insider has to disgorge. Cornell and Morgan (1990) provide a comprehensive overview of the strengths and weaknesses of the use of event studies for private securities litigation.

DeAngelo and DeAngelo (1989) study management turnover subsequent to proxy contests. The cumulative survival rate for incumbent management in 60 firms one year after the proxy contest outcome (regardless of the outcome) is 28 percent; and three years after the outcome is 18 percent. Ikenberry and Lakonishok (1993) study the performance of firms subject to proxy contests before and subsequent to the contest. Both stock market and accounting based performance measures indicate poor performance five years prior to the proxy contest. Also, accounting based performance measures indicate poor performance five years subsequent to the proxy contest, especially if dissidents win. Taken together, these studies' findings paint a picture similar to that of the Martin and McConnell study and the review by Furtado and Karan, in relating changes in control to changes in top management and poor performance. However, Agrawal and Jaffe (2003) examine the prebid performance (using accounting and market measures) of a very large sample of takeover targets and subsamples of targets of hostile bids, and conclude that targets are not poor performers (or that poor performance occurs so many years prior to the bid that it is incorrect to consider takeovers as disciplinary devices for inefficient managers); they do not examine the turnover of management within their sample of target firms.

Denis, Denis, and Sarin (1997) study the impact of ownership structure on management turnover. They find that management turnover is more likely as the equity ownership of officers and directors decreases, and whether or not there is an outside blockholder. But they also document evidence suggesting that the impact of managerial ownership on turnover may be due, in part, to the impact of managerial ownership on corporate control activity; they observe a significantly higher occurrence of corporate control activity in the year prior to the management turnover, regardless of the level of management ownership.

Bhagat and Jefferis (1994) study the frequency of executive turnover in firms that paid greenmail. Greenmail or targeted repurchase refers to the purchase of a block of shares by the company at a premium from a single shareholder or group of shareholders; this offer is not made to all shareholders. The motivation for paying greenmail is alleged to be deterrence of a takeover on terms that would be unfavorable to incumbent management. They find management turnover is less frequent at repurchasing firms than control firms of similar size and industry. This is true unconditionally, and for a subsample of firms that do not experience a takeover. However, they argue that takeovers and managerial turnover are endogenous. Less frequent management turnover at repurchasing firms may suggest that managers of those firms are insulated from market discipline. Alternatively, it may be the case that managerial performance at repurchasing firms does not warrant discipline. They find that accounting based performance measures for firms that paid greenmail and the control sample are similar both prior to and subsequent to the repurchase.

Denis and Serrano (1996) study management turnover following unsuccessful control contests. Thirty-four percent of the firms experience management turnover from the initiation of the control contest through two years following resolution of the contest. This rate of management turnover is twice that of a random sample of firms during

the same period. Further, they find that turnover is concentrated in poorly performing firms in which investors unaffiliated with management purchase large blocks of shares during and subsequent to the control contest. In contrast, managers of firms with no unaffiliated block purchases appear to be able to extend their tenure despite an equally poor performance prior to the control contest.

Mikkelson and Partch (1997) study the impact of performance on management turnover during an active takeover market (1984–1988) compared to a less active takeover market (1989–1993) for a sample of unacquired firms. They find the frequency of managerial turnover is significantly higher during the active takeover market compared to the less active takeover market. Additionally, this decline in the frequency of managerial turnover is most conspicuous among poorly performing firms. This is an interesting finding in light of Agrawal and Jaffe's (2003) counterintuitive finding that takeover targets are not poor performers. Agrawal and Jaffe note, in attempting to reconcile that finding with the dominant view in the literature that takeovers discipline managers, that because their data relate only to actual takeovers, the threat of takeovers might discipline managers even though takeovers are only carried out when there are more compelling reasons than poor performance (i.e., managerial discipline); that is, "external control mechanisms (such as the threat of a takeover) may facilitate internal mechanisms (such as boards) in disciplining bad managers" (p. 744). This hypothesis would appear to be confirmed by Mikkelson and Partch's finding that turnover rates decrease as takeover activity (and hence the threat of a bid) decreases.

4.4. Research on corporate governance

Virtually all of the important mechanisms of corporate governance have been subjected to event study analysis. These include boards of directors, shareholder proposals, derivative lawsuits, and executive compensation. Although all of these devices have been posited to perform a critical function of reducing the agency costs of the separation of ownership and control in the U.S. public corporation, empirical studies do not provide strong support for this viewpoint. Neither shareholder proposals nor lawsuits have a significant positive price effect. A positive stock price effect is associated with appointment of an independent director to the board, but board composition has not been found to impact positively on performance. By contrast, the incentive-aligning device of stock-based executive compensation has been found to affect stock prices positively. These findings suggest that widely-shared beliefs concerning what are essential components for effective corporate governance may be mistaken, and that affirmative policies to foster such devices ought to be reconsidered.

4.4.1. Boards of directors

Directors are seen as performing a pivotal role in the corporation: ensuring that management acts in furtherance of the shareholders' interest. As the repository of the shareholders' agents to monitor the agents directly running the firm, the board structure

interposes an additional strata of agency problems on the more basic agency relation between managers and owners of the firm. Accordingly, commentators have emphasized the desirability of a board composed of independent or outside directors—directors without a financial or personal connection to management—to ensure that the board structure is not simply creating a further layer of agency problems that is one-step removed from operations (e.g., Eisenberg, 1976). This position has been incorporated into the legal system: stock exchanges require listed firms to have independent directors on their boards and on specified committees, such as the audit, nominating and compensation committees, and courts take the board's independence into account when assessing claims in shareholder lawsuits.

Consistent with the monitoring view of outside directors, the market views such directors favorably. An event study of the appointment of an outside director reports a significant positive price effect, even when a majority of the board was already independent (Rosenstein and Wyatt, 1990). This increase, while statistically significant, is economically small and could reflect signalling effects. Appointing an additional independent director could signal that a company plans to address its business problems, even if board composition doesn't affect the company's ability to address the problems.

Bhagat and Black (1999) surveyed the literature on how board composition affects firm performance or vice versa. Prior studies of the effect of board composition on firm performance generally adopt one of two approaches. The first approach involves studying how board composition affects the board's behavior on discrete tasks, such as replacing the CEO, awarding golden parachutes, or making or defending against a takeover bid. This approach can involve tractable data, which makes it easier for researchers to find statistically significant results. But it doesn't tell us how board composition affects overall firm performance. For example, there is evidence that firms with majority-independent boards perform better on particular tasks, such as replacing the CEO (Weisbach, 1988) and making takeover bids (Byrd and Hickman, 1992). But these firms could perform worse on other tasks that cannot readily be studied using this approach (such as appointing a new CEO or choosing a new strategic direction for the firm), leading to no net advantage in overall performance. Also, events such as CEO turnover and takeovers are rare occurrences for firms. The greater and more positive contribution of boards may be in the ongoing advice they give to senior management in private meetings; it would be difficult to study this via the traditional event-study method.

The second approach consists of examining directly the correlation between board composition and firm performance. This approach allows us to examine the "bottom line" of firm performance (unlike the first approach), but involves much less tractable data. Firm performance must be measured over a long period, which means that performance measures are noisy and perhaps misspecified as discussed in Section 2. As Bhagat and Black (1999) review, the bulk of the studies do not find a positive association between board independence and performance (see also Romano, 1996).

The inability to find a connection between performance and board composition in most empirical studies may be due to the endogenous relation between those vari-

ables, as discussed in Section 3. Several researchers have examined whether board composition is endogenously related to firm performance, with inconsistent results. Hermalin and Weisbach (1988) and Weisbach (1988, p. 454) report that the proportion of independent directors on large firm boards increases slightly when a company has performed poorly: firms in the bottom performance decile increase their proportion of independent directors by around 1 percent in the subsequent year, relative to other firms. Bhagat and Black (2002) address the possible endogeneity of board independence and firm performance by adopting a three-stage least squares approach (3SLS); this permits firm performance, board independence, and CEO stock ownership to be endogenously determined. Bhagat and Black find a reasonably strong correlation between poor performance and subsequent increase in board independence. The change in board independence seems to be driven by poor performance rather than by firm and industry growth opportunities. However, there is no evidence that greater board independence leads to improved firm performance. If anything, there are hints in the other direction.[30] The conventional wisdom that supports a very high degree of board independence, although it may explain why poorly performing firms increase the independence of their boards, appears to rest on a shaky empirical foundation.

It is possible that the failure to find that independent boards improve performance is due to the fact that not all outside directors are truly independent from management, and empirical researchers cannot distinguish between "effective" and "ineffective" independent boards. But a more compelling reason why increasing board independence does not result in improved performance is that having inside directors could add value in strategic planning[31] or evaluation of potential successors for the CEO (e.g., Vancil, 1987). From this perspective, independent boards at best could improve corporate decision-making in certain extraordinary situations, such as management-buyouts or poor performance (e.g., Weisbach, 1988; Lee et al., 1992), which are very low probability events for most firms.

These data suggest that it would be prudent for companies to consider experimenting with modest departures from the norm of a "supermajority independent" board with only one or two inside directors. The independent directors will still numerically dominate the board, and can take appropriate action in a crisis. In addition, effort should

[30] Agrawal and Knoeber (1996) estimate a system of simultaneous equations for firm performance, measured by Tobin's Q, and several mechanisms of corporate governance that can control the agency problem: insider ownership, board composition, debt policy, reliance on external labor markets for managers, and corporate control market activity. They find that the proportion of outsiders on the board has a negative effect on performance; the other governance devices are not significant. They conclude that, apart from board composition, control mechanisms are chosen optimally by firms (that is, use of the various devices are traded off so as to maximize firm value). This paper provides an excellent example of the endogeneity problem discussed in Section 3: significant associations that are found between variables when the relations are estimated in separate OLS regressions disappear in the simultaneous equations estimation.

[31] This is consistent with Klein's (1998) evidence that inside director representation on investment committees of the board correlates with improved performance.

be focused on devising mechanisms to enhance director independence or otherwise improve their incentives to monitor by encouraging greater equity ownership.[32] A final implication of these data is that the response of the stock exchanges to require a majority of independent directors in the aftermath of the corporate scandals involving such firms as Enron and WorldCom and ensuing federal legislation was misguided.[33]

4.4.2. Shareholder proposals and charter amendments

A mechanism of corporate governance used increasingly by certain institutional and individual investors is shareholder proposals, which are included in a firm's proxy materials under SEC Rule 14a-8 and voted on at the annual shareholders' meeting. The most active institutional users of this tool, public pension and union funds, sponsor a variety of proposals that they assert will improve performance, including proposals to enhance the independence of the board, reform executive compensation, remove takeover defensive tactics, and adopt confidential voting (see, e.g., Del Guercio and Hawkins, 1999). The institutions must notify management of their intent to submit a proposal in advance of the meeting under the SEC rule, a requirement that has the beneficial effect for the sponsor that management will frequently negotiate a compromise in order to avoid the proposal's submission (see id.).

Numerous event studies have been undertaken to determine whether the introduction of a shareholder proposal affects firm value. The uniform finding is that they do not (see Romano, 2001).[34] The absence of a significant effect is not likely to be due to imprecision in the event study methodology because in these studies the sample sizes are large and the event dates are precise (see Section 2). A plausible explanation of the absence of a price effect is that the objects of many shareholder proposals—independent boards, limits on executive compensation and in particular on incentive pay, and confidential voting—do not, when investigated by event studies, significantly affect firm value (see Romano, 2001). It is improbable that a proposal to undertake a governance strategy that does not itself significantly affect prices will produce a price effect.

[32] Hall and Liebman (1998) provide evidence of the sensitivity of management's financial wealth to firm performance. The hypothesis that director incentives affect firm performance is consistent with the evidence in Bhagat, Carey, and Elson (1999).

[33] The federal legislation, the Sarbanes-Oxley Act of 2002, required publicly traded firms to have audit committees composed of all independent directors, see section 301 of Pub. Law 107-204, codified as §10A(m) of the Securities Exchange Act of 1934, 15 U.S.C. §78j-1(m) (2003), a provision that is as misguided as the new stock exchange rules on the composition of the board. This is because there is a sizeable literature on audit committee composition that indicates that 100 percent independence does not improve performance or the quality of firms' accounting statements. See Romano (2005).

[34] Negotiations with management, by contrast, have been found to produce both significant positive and negative price effects. As discussed in Romano (2001), the difference may be explained either as evidencing that management selects the highest valued proposals for negotiation (or lowest for the negative effect studies) or that negotiation, by indicating management's responsiveness to certain investor concerns, provides a signal of management's quality (the price effect reflects market updating regarding management quality rather than the value to the firm of the omitted proposal).

It is troubling that institutional investors, who are, after all, in most cases, fiduciaries, would spend significant effort sponsoring proposals that are not likely to improve firm performance. A lack of information regarding the appropriate governance policy to adopt does not seem to be a plausible explanation for the behavior of at least the most prominent sponsors of proposals, who are sophisticated institutions. Public pension fund managers might well be informed about which proposals are useful and still champion fruitless proposals, however, if the managers obtain private benefits from submitting such a proposal, given the absence of strong incentives of boards of public funds to monitor their staff (or the presence of similar private benefits for board members).

The fact that, in contrast to public pension funds, private pension and mutual funds do not engage in comparable highly visible activism has been explained by the competitive nature of the industry, or as cost-conscious private funds' free-riding on the expenditures of activist public funds (e.g., Black, 1998, p. 460). This may be so. But there is a further, complementary explanation, that private institutions' managers are less likely to obtain private benefits from engaging in shareholder activism than public and union fund managers.[35] Both explanations are provided with support from survey data indicating that private fund managers perceive the costs and benefits of shareholder activism differently from public pension fund managers (Downes, Houminer, and Hubbard, 1999, pp. 32–34).

In short, financial economists have not been able to identify a positive performance effect of shareholder activism because much of that activism would appear to be misdirected. To the extent that this mismatch is due to problematic behavior on the part of fund managers sponsoring proposals involving private benefits, potential solutions are the adoption of better internal control mechanisms for fund boards, such as program audits, or a reduction in the current subsidy of the presentation of proposals by requiring sponsors of losing proposals to reimburse the corporation, in whole or in part, for the cost of the proposal (see Romano, 2001).

Most shareholder proposals receive low levels of support, although some subsets of institutional investors' proposals on corporate governance, such as those concerning the elimination of takeover defenses and the adoption of confidential voting, obtain high levels of support, and even majorities (see Gillan and Starks, 2000). Management proposals, in contrast, receive uniformly high levels of support and are virtually always adopted. This is not surprising, as managements consult with proxy solicitation firms and shy away from putting up proposals that the proxy firms suggest are not likely to be approved. The consequence has been that since institutional investors became active in opposing takeover defenses in the late 1980s, managements have discontinued, for the most part, proposing charter-level takeover defenses (indeed, most large public companies with defenses adopted them by the early 1980s, before the takeover market slowed and institutional shareholders began actively to oppose defenses).

[35] For examples of possible private benefits relevant to public pension or union fund officials in contrast to private fund managers, related to furthering political reputations or collective action goals see, e.g., Romano (1993c, p. 822); Thomas and Martin (1998, pp. 61–62).

Gompers, Ishii, and Metrick (2003) construct a governance index using the Investor Responsibility Research Center (IRRC) database. This governance index is constructed by considering various charter amendments that would increase the difficulty of a hostile takeover, state takeover legislation to which the firms are subject, charter provisions to indemnify officers and directors, management compensation arrangements following change in control, and shareholder voting rights. Gompers, Ishii, and Metrick find a positive correlation between stronger shareholder rights (basically, fewer takeover defenses) and firm value (as measured by Tobin's Q) and stock market performance. They provide evidence in support of a causal explanation, that weak governance caused the poor performance (weak governance firms appear to have higher agency costs in the form of inefficient investments), although they note that the "analysis of causality" cannot be "conclusive," as unobserved characteristics could be correlated both with their index and performance (p. 142). The competing evidence is that industry classification explains between one-sixth and one-third of the abnormal returns across the firms, a finding consistent with research by Gillan, Hartzell, and Starks (2002) that indicates that corporate governance characteristics of firms vary significantly across industries.

Core, Guay, and Rusticus (2006) seek to determine whether better governance causes higher returns or other unidentified, correlated factors are driving the Gompers, Ishii, and Metrick results. The causal explanation offered by Gompers, Ishii, and Metrick requires investors to have not anticipated the cost (the increased managerial agency costs) of weak governance devices; thus, when the agency costs result in lower profits, investors lower their earnings expectations and the stock price declines. The test Core, Guay, and Rusticus employ is, accordingly, twofold: first they examine whether there is a relation between governance and operating performance, and then, whether the market anticipated the performance. Core, Guay, and Rusticus measure operating performance by return on assets, adjusted for industry, and find a significant negative relation: firms with weak governance (as defined by the Gompers, Ishii, and Metrick measure) had poor operating returns. But they further find that the weak governance firms' underperformance was anticipated by the market. Analyst forecasts of the firms' earnings in relation to their governance characteristics were unbiased. Core, Guay, and Rusticus thus conclude that weak governance does not cause lower returns (the association found in the Gompers, Ishii, and Metrick study is not causal).

Core, Guay, and Rusticus thereupon investigate alternative explanations to the causal one. They look at whether the difference could be due to differences in risk (firms with weak governance happen to be firms with low risk). Although they find supporting evidence of this explanation, in that weak governance firms have the lowest cost of capital and lowest realized returns, the effect is too small to explain the differential in abnormal returns found by Gompers, Ishii, and Metrick. They then examine governance portfolio returns out-of-sample (2000–2002), to evaluate whether time-specific factors correlated with governance characteristics could explain the Gompers, Ishii, and Metrick results. They find that the abnormal returns trend reverses: the weak governance firms outperformed the strong governance firms in this period (although the reversal is not statistically significant when controlling for the Fama-French three-factor pric-

ing model and momentum, the returns model used in the Gompers, Ishii, and Metrick study). The weak governance firms did have lower operating performance in the out-of-sample period and analysts continued to forecast the difference. These data suggest that time-period specific factors may account for the finding of positive abnormal returns to strong governance firms in the Gompers, Ishii, and Metrick study.

It should also be noted that the accumulated evidence on the impact on shareholder wealth of the charter provisions in the Gompers, Ishii, and Metrick governance index is weak, with point estimates that range from slightly negative to slightly positive; see DeAngelo and Rice (1983) and Linn and McConnell (1983). Ownership data in firms that propose such amendments and voting patterns on these amendments suggest that the amendments are supported by corporate insiders and opposed by the typical institutional investor. Brickley, Lease, and Smith (1988) document voting patterns consistent with the hypothesis that institutional investors are more likely than nonblockholders to oppose antitakeover amendments, while corporate insiders support the adoption of amendments.[36] A plausible interpretation of these data is that antitakeover amendments protect managers from the discipline of the takeover market while potentially harming shareholders.

There are, however, reasonable arguments to support the view that management-sponsored antitakeover amendments do not actually injure shareholders. For instance, they may solve collective action problems on the part of dispersed shareholders that prevent them from negotiating with bidders to raise takeover premiums. The notion that antitakeover amendments increase managers' bargaining power is inconsistent with Pound's (1987) finding that antitakeover amendments do not increase bid premiums. Subramanian (2003) considers the impact of takeover defenses on the target management's bargaining power and its potential impact on takeover premiums. He finds no difference in takeover premiums for negotiated acquisitions (hostile and unsolicited takeovers are excluded from his sample) in states that have a differential encouragement of a specific takeover defense, poison pills (these defenses are not subject to shareholder

[36] Brickley, Lease, and Smith (1988) find higher levels of support for management proposals of takeover defenses when the voting pool contains institutions that they consider susceptible to management pressure because the institutions may have additional business relations with the firms (banks and insurance companies). However, Van Nuys (1993) finds, in a case study that tracks actual votes on a contested management proposal of defenses, that the institutions that had business relations with the firm were not more likely to vote in favor of management than other institutions. See also Davis and Kim (2005), who find that mutual funds' actual votes on shareholder proposals do not differ significantly across portfolio firms with whom they have pension business and those with whom they do not, but that funds' overall voting policies that appear to be more favorable to management are positively related to funds' overall volume of pension business. Institutions' conflicts in voting from business relations provided a basis for shareholder activists' advocacy of confidential proxy voting from the 1980s–1990s (the assumption being that institutions would be able to vote against management if it did not know how the institution voted), but in 2002 that policy position was ignored by SEC, which required mandated disclosure of the proxy votes by a subset of institutions (mutual funds). While that reform may have been misguided from a cost-benefit perspective, there is no evidence that confidential voting affects voting outcomes (see Romano, 2003).

approval, which limits extrapolating from his data to the above-mentioned charter-level defenses which shareholders must approve). Poison pills are considered to be the most powerful defense against hostile bids and thus the defense most likely to affect adversely shareholders' interest, although a recent study by Danielson and Karpoff (2002) finds, surprisingly, that firms' operating performance improves after a pill's adoption. Subramanian also interviews the heads of mergers and acquisitions of the major investment banks to learn about their perception of the impact of takeover defenses on takeover premiums. The perceptions of the investment bankers are consistent with his data on premiums; most think that defenses are irrelevant in negotiated transactions.

A variant on the bargaining thesis but related to the collective action problem is that takeover defenses facilitate an auction, by delaying the consummation of the initial bid in time for another bidder to appear, rather than by facilitating negotiation for a higher price with the first bidder. The argument is that dispersed shareholders would be unable to hold out for an auction, in contrast to a sole owner. Subramanian's (2003) analysis does not challenge this alternative view of defensive tactics, which motivates the opposition to defenses by critics such as Easterbrook and Fischel (1981), who objected to what they considered defenses' primary effect, that they result in auctions. The data on takeovers indicate that auctions do increase premiums (e.g., Bradley, Desai, and Kim, 1988), although other defenses do not appear to do so (e.g., Pound, 1987; Hackl and Testani, 1988). Firm-level defenses such as antitakeover charter amendments are not, however, unambiguously necessary to encourage an auction, as that is the likely effect of the Williams Act (the federal takeover regulation).

A second argument, that managers of firms adopting amendments are simply enjoying contractual protection against takeovers afforded them by shareholders, is consistent with the fact that shareholders vote to approve the overwhelming majority of proposals put forth by management. Jarrell, Brickley, and Netter (1988) attribute shareholder support for wealth-decreasing amendments to the free-rider problem. Bhagat and Jefferis (1991) argue that the transaction costs that give rise to the free-rider problem are, at least in part, an endogenous consequence of strategic behavior that might be eliminated through either changes in the charter or proxy reform. Despite the lack of such reform, in recent years managers have ceased to present such defenses to shareholders, presumably out of concern that the proposals would be defeated, as institutional investors have become better organized and more active in the proxy process.

4.5. Event studies and securities regulation

In addition to the application of event study methodology to litigate securities fraud cases, the methodology has been used to assist in policy analysis of securities regulation, as it has been used in the state competition and takeover debates, most recently in the evaluation of procedural reforms wrought by the Private Securities Litigation Reform Act of 1995. This congressional initiative was intended to render it more difficult to bring a civil action under the federal securities laws (H.R. Conference Report, 1995). Because the legislation was unexpectedly vetoed by President Clinton and the veto was overridden shortly thereafter (see Johnson, Kasznik, and Nelson, 2000, pp. 8–9), in

contrast to legislation that comes to fruition over a long period of time, fairly clean event dates for the Act can be identified.

Event studies have found that the Act had a significantly positive stock price effect (Spiess and Tkac, 1997; Johnson, Kasznik, and Nelson, 2000). This result is interpreted as validating the congressional impetus for the legislation, concern over the incidence of nonmeritorious lawsuits, because the market valued the legislation's benefits from curtailing frivolous suits as greater than its costs in restricting meritorious suits. Further supporting this conclusion, a court decision adopting the most stringent interpretation of the Act's pleading requirement, which furthered Congress' goal of making filing of a nonmeritorious suit more difficult, had a statistically significant positive effect on stock prices for a sample of high technology firms, which operate in an industry sector with a high probability of securities litigation (Johnson, Nelson, and Pritchard, 2000).[37]

There is, in fact, a long history of empirical research evaluating securities regulation, and in particular the mandatory disclosure regime, going back to the classic studies by Stigler (1964) and Benston (1973) of the original federal statutes enacted during the New Deal, the Securities Act of 1933 and the Securities Exchange Act of 1934. Those studies challenged the conventional legal wisdom that the federal legislation was of value to investors, as both Stigler and Benston found that the statutes did not improve the returns of affected firms. Not surprisingly, their work was quickly criticized (e.g., Friend and Herman, 1964; Friend and Westerfield, 1975). Recent work by Paul Mahoney (1999, 2001a) has raised a different set of concerns regarding the legislation but that lead to the same conclusion on its efficacy for investors, as his work shows how key provisions of the 1933 Act were enacted to benefit an established set of financial institutions over new entrants, and how crucial premises of the 1934 Act concerning market manipulation by stock pools were incorrect. We cannot begin to review the literature on securities regulation in this chapter. Instead, we refer the reader to Romano (2002), which reviews the literature and concludes that there is a paucity of evidence that the federal disclosure regime administered by the SEC has benefited investors.

4.6. Comparative corporate governance

In a series of influential papers, La Porta et al. (1997, 1998, 1999, 2000, 2002) analyze the role a country's legal system has in protecting investor rights. They argue (2000, p. 4): "Such diverse elements of countries' financial systems as the breadth and depth of their capital markets, the pace of new security issues, corporate ownership structures, dividend policies, and the efficiency of investment allocation appear to be explained both conceptually and empirically by how well the laws in these countries protect outside investors." La Porta et al. (1998) draw on the work of David and Brierley (1985)

[37] In contrast to state corporate law, the federal securities laws consist of mandatory rules. Firms cannot therefore contract around a court decision in this context as they can in the corporate law context, and decisions can therefore impose wealth effects on nonlitigants.

and Zweigert and Kotz (1987) to postulate that the commercial legal codes of most countries are based on four legal traditions: the English common law, the French civil law, the German civil law, and the Scandinavian law. They find that common law countries provide the most protection to investors (La Porta et al., 1998), and that they have the deepest stock markets and most dispersed corporate ownership structures (La Porta et al., 1997, 1999). They also document that countries develop substitute mechanisms for poor investor protection, such as mandatory dividends and greater ownership concentration. In a follow-up paper, La Porta et al. (2002) find that investor protection is positively correlated with valuation across countries.

In their most recent work, La Porta, Lopez-de-Silanes, and Shleifer (2006) construct two indices measuring the quality of securities regulation representing the strength of public and private enforcement mechanisms (the former consists of powers of the national securities regulator, the latter, private litigation regime features such as the burden of proof), to examine the effect of securities regulation on stock markets. As in the case of their investor protection measure, which they refer to as a shareholder rights or antidirector rights index, the public and private enforcement measures have higher values in nations with common law traditions. La Porta et al. find that the private enforcement measure is more significant than either the public enforcement measure or the shareholder rights index for the development of a stock market.

The overarching theme of the influential and extensive La Porta et al. corpus is that "law matters." The cluster of countries associated with the common-law legal tradition, which is identified with stronger investor protection and securities regulation, have deeper stock markets, less concentrated ownership of public firms, and in their view, given those nations' higher level of financial development, offer better opportunities for economic growth and prosperity. Their work has generated considerable discussion. Some scholars have disagreed with the construction of the investor protection measure (e.g., Vagts, 2002; Berglof and von Thadden, 1999). Others have sought to offer alternative explanations of why common law systems are associated with higher financial development.

For example, Mahoney (2001b, p. 523) contends that "legal origin affects growth through channels other than finance," that is, the source of the association between the common law and financial growth is that legal tradition's view of the role of the state, which emphasizes limited government and an independent judiciary, leading to secure property and contract rights, rather than its effect on financial markets through shareholder protection measures. Estimating generalized methods of moments (GMM) coefficients for endogeneous variables proxying for judicial independence, limited government (scope of civil liberties) and property and contract rights, using the legal tradition (common or civil law) as an instrument, Mahoney finds that the endogenous variables significantly explain growth in real per capita GDP (the null that legal origin affects growth solely through its affect on the endogeneous variables—that the instrument is uncorrelated with the error term—cannot be rejected). Roe (2000), by contrast, offers a politically-based explanation for La Porta et al.'s finding, emphasizing the political importance of stakeholders in civil law origin nations.

Notwithstanding disagreements over the significance of La Porta et al.'s findings, it cannot be denied that their work has had a major impact—international institutions such as the International Monetary Fund and World Bank focus on corporate governance as a key plank in their policy toward emerging market nations[38]—and that their corporate law index captures an important element driving cross-national differences in financial development, despite nuances of legal regime differences among nations that are grouped together in their legal categorization (see, e.g., Cheffins, 2001, distinguishing between the corporate law and institutions of the United States and United Kingdom, which are grouped together in La Porta et al.'s analysis). Another sign of the influence of La Porta et al.'s research agenda is the large body of literature that has developed using the La Porta et al. variables to investigate a variety of other cross-national differences. These studies also provide evidence that legal rules matter in important ways for national economies. We note three such examples; for a more extensive review see Denis and McConnell (2003).

Rossi and Volpin (2004) use the differential investor protection characterization across countries developed by La Porta et al. to study the volume and characteristics of cross-border acquisitions. They find that targets are typically in countries with poorer investor protection than acquirers. They conclude that cross-border acquisitions may be partially motivated by enhancement of investor protection in target firms. Wurgler (2000) studies the allocation of capital in financial markets across countries; he finds, among other results, that capital is more efficiently allocated (increased investments in growing industries and decreasing investments in declining sectors) in nations with higher investor protection as measured by La Porta et al. Hail and Leuz (2003) examine the differences in firms' cost of capital across countries. They find that higher levels of securities disclosure, greater public and private securities law enforcement (the measures in La Porta, Lopez-de-Silanes, and Shleifer, 2006), and, to a lesser extent, a commitment to the rule of law (a variable from La Porta et al., 1997 that is correlated with their shareholder rights index) reduce firms' cost of capital, controlling for other country factors and risk that predictably affect the cost of capital. The effect of securities regulation on the cost of capital is, however, smaller for globally integrated (more developed) markets.

Given the growth of transition and developing markets, with the expansion of the European Union and international trade agreements over the past decade, the research initiated by La Porta et al. will no doubt continue to influence the agenda for comparative research. This is because a better understanding of the connection between legal rules, particularly those related to the organization of economic activity in corporations, and nations' financial growth and development, which La Porta et al.'s painstaking empirical analysis has identified, is undoubtedly a key to improving social wealth and accordingly, individual welfare.

[38] For critiques of international financial institutions' application of an "Anglo-American" corporate governance paradigm to emerging nations, a policy supported by La Porta et al.'s research, see, e.g., Berglof and von Thadden (1999); and Singh, Singh, and Weisse (2002).

5. Conclusion

In this chapter we have attempted to provide the reader with a sense of the richness of the extensive body of empirical research in corporate law, and its usefulness for public policy analysis. With respect to the literature on corporate litigation, defendants experience economically-meaningful and statistically-significant wealth losses upon the filing of the suit (except for shareholder derivative suits), and significant wealth increases on the announcement of a settlement when the plaintiff is another firm. Plaintiff firms, however, experience no significant wealth effects upon filing a lawsuit, and the wealth implications of settlements appear to be trivial. These findings suggest that, at a minimum, lawsuits are not a value-enhancing way for corporations to settle their disagreements with other corporations.

Event studies in particular have been influential in the making of corporate law and in corporate law scholarship. They have informed the major policy debates over the production of corporate laws and takeovers, and the jurisprudence on securities law. The impact of empirical research on these issues can be overstated: the strength with which particular corporate law commentators hold priors concerning the appropriate policy will cause them to update those priors differentially. But over time empirical research does have an effect, and its effect has reached beyond the academy to corporate law decision-makers. This is precisely what has occurred in the state competition and takeover debates over the past two decades: academic consensus shifted to a more favorable assessment of state competition and of takeovers as empirical research accumulated that was probative on these issues, and the approach of the SEC and the Delaware courts to takeovers changed as well.

Empirical research on other mechanisms of corporate governance, such as boards of directors and shareholder proposals, is just beginning to enter the policy debates, as interest in these mechanisms is a relatively recent phenomenon, following the decrease in hostile takeovers in the 1990s. That research generally finds the wealth effects of these governance mechanisms are insignificant. As such findings are reinforced and cumulate, a process of reassessment of the conventional position on the efficacy of these corporate governance mechanisms as managerial monitoring devices may well occur, similar to what transpired in the state competition and takeover debates.

References

Acharya, S. (1988). "A generalized econometric model and tests of a signalling hypotehsis with two discrete signals". Journal of Finance 43, 413–429.

Agrawal, A., Jaffe, J.F. (2003). "Do takeover targets underperform? Evidence from operating and stock returns". Journal of Financial and Quantitative Analysis 38, 721–740.

Agrawal, A., Knoeber, C.R. (1996). "Firm performance and mechanisms to control agency problems between managers and shareholders". Journal of Financial and Quantitative Analysis 31, 377–397.

Agrawal, A., Knoeber, C.R. (2001). "Do some outside directors play a political role?" Journal of Law and Economics 44, 179–198.

Akhavein, J.D., Berger, A.N., Humphrey, D.S. (1997). "The effect of megamergers on efficiency and prices: evidence from a bank profit function". Review of Industrial Organization 12, 95–139.

Alexander, J.C. (1991). "Do the merits matter? A study of settlements in securities class actions". Stanford Law Review 43, 497–598.
Alexander, J., Spivey, M., Marr, M.W. (1997). "Nonshareholder constituency statutes and shareholder wealth: a note". Journal of Banking and Finance 21, 417–432.
Amemiya, T. (1985). Advanced Econometrics. Harvard University, Cambridge, MA.
Andrade, G., Mitchell, M., Stafford, E. (2001). "New evidence and perspectives on mergers". Journal of Economic Perspectives 15, 103–120.
Asquith, P., Wizman, T. (1990). "Event risk, covenants, and bondholder returns in leveraged buyouts". Journal of Financial Economics 27, 195–213.
Barber, B.M., Lyon, J.D. (1997). "Detecting long-run abnormal stock returns: the empirical power and specification of test statistics". Journal of Financial Economics 43, 341–372.
Barclay, M.J., Litzenberger, R.H. (1988). "Announcement effect of new equity issues and the use of intra-day price data". Journal of Financial Economics 21, 71–100.
Barnhart, S.W., Rosenstein, S. (1998). "Board composition, managerial ownership and firm performance: an empirical analysis". Financial Review 33, 1–16.
Baumol, W.J., Blackman, S.B., Wolff, E.N. (1989). Productivity and American Leadership: The Long View. MIT, Cambridge, MA.
Baysinger, B.D., Butler, H.N. (1985). "The role of corporate law in the theory of the firm". Journal of Law and Economics 28, 179–191.
Bebchuk, L.A. (1992). "Federalism and the corporation: the desirable limits on state competition in corporate law". Harvard Law Review 105, 1437–1510.
Bebchuk, L.A., Cohen, A. (2003). "Firms' decisions where to incorporate". Journal of Law and Economics 46, 383–425.
Bebchuk, L.A., Ferrell, A. (2001). "A new approach to regulatory competition and takeover law". Virginia Law Review 87, 111–164.
Bebchuk, L.A., Cohen, A., Ferrell, A. (2002). "Does the evidence favor state competition in corporate law?" California Law Review 90, 1777–1821.
Benston, G.J. (1973). "Required disclosure and the stock market: an evaluation of the Securities Exchange Act of 1964". American Economic Review 63, 132–155.
Berglof, E., von Thadden, E. (1999). "The changing corporate governance paradigm: implications for transition and developing countries". In: Annual World Bank Conference on Development Economics, Conference Paper, Washington, D.C.
Bhagat, S., Black, B. (1999). "The uncertain relationship between board composition and firm performance". Business Lawyer 54, 921–943.
Bhagat, S., Black, B. (2002). "The non-correlation between board independence and long-term firm performance". Journal of Corporation Law 27, 231–273.
Bhagat, S., Brickley, J.A. (1984). "Cumulative voting: the value of minority shareholder voting rights". Journal of Law and Economics 27, 339–365.
Bhagat, S., Jefferis, R.H. (1991). "Voting power in the proxy process: the case of antitakeover charter amendments". Journal of Financial Economics 30, 193–226.
Bhagat, S., Jefferis, R.H. Jr. (1994). "The causes and consequences of takeover defense: evidence from greenmail". Journal of Corporate Finance 1, 201–231.
Bhagat, S., Jefferis, R.H. (2002). The Econometrics of Corporate Governance Studies. MIT, Cambridge, MA.
Bhagat, S., Romano, R. (2002a). "Event studies and the law: part I—technique and corporate litigation". American Law and Economics Review 4, 141–167.
Bhagat, S., Romano, R. (2002b). "Event studies and the law: part II—empirical studies of corporate law". American Law and Economics Review 4, 380–423.
Bhagat, S., Bizjak, J., Coles, J.L. (1998). "The shareholder wealth implications of corporate lawsuits". Financial Management 27, 5–27.
Bhagat, S., Brickley, J.A., Coles, J.L. (1994). "The wealth effects of interfirm lawsuits: evidence from corporate lawsuits". Journal of Financial Economics 35, 221–247.

Bhagat, S., Carey, D., Elson, C. (1999). "Director ownership, corporate performance, and management turnover". Business Lawyer 54, 885–920.

Bhagat, S., Dong, M., Hirshleifer, D., Noah, R. (2005). "Do tender offers create value: new methods and evidence". Journal of Financial Economics 76, 3–60.

Bhagat, S., Shleifer, A., Vishny, R.W. (1990). "Hostile takeovers in the 1980s: the return to corporate specialization". In: Brookings Papers on Economic Activity: Microeconomics. Brookings Institution, Washington, D.C., pp. 1–84.

Bittlingmayer, G., Hazlett, T.W. (2000). "DOS kapital: has antitrust action aginst Microsoft created value in the computer industry?" Journal of Financial Economics 55, 329–359.

Bizjak, J.M., Coles, J.L. (1995). "The effect of private antitrust litigation on the stock-market valuation of the firm". American Economic Review 85, 436–459.

Black, B. (1989). "Bidder overpayment in takeovers". Stanford Law Review 41, 597–653.

Black, B. (1998). "Shareholder activism and corporate governance in the United States". In: Newman, P. (Ed.), The New Palgrave Dictionary of Economics and the Law, vol. 3. Stockton, New York, pp. 459–465.

Bradley, M., Schipani, C.A. (1989). "The relevance of the duty of care standard in corporate governance". Iowa Law Review 75, 1–74.

Bradley, M., Desai, A., Kim, E.H. (1988). "Synergistic gains from corporate acquisitions and their division between the stockholders of target and acquiring firms". Journal of Financial Economics 21, 3–40.

Brav, A., Gompers, P.A. (1997). "Myth or reality? The long-run underperformance of initial public offerings: evidence from venture and non-venture capital-backed companies". Journal of Finance 52, 1791–1821.

Brickley, J.A. (1986). "Interpreting common stock returns around proxy statement disclosures and annual shareholder meetings". Journal of Financial and Quantitative Analysis 21, 343–349.

Brickley, J.A., Coles, J.L., Terry, R.L. (1994). "Outside directors and the adoption of poison pills". Journal of Financial Economics 35, 371–390.

Brickley, J.A., Lease, R.C., Smith, C. (1988). "Ownership structure and voting on antitakeover amendments". Journal of Financial Economics 20, 267–291.

Broner, A. (1987). New Jersey Shareholders Protection Act: An Economic Evaluation. Office of Economic Policy, Trenton, NJ.

Brook, Y., Rao, R.K.S. (1994). "Shareholder wealth effects of directors' liability limitation provisions". Journal of Financial and Quantitative Analysis 29, 481–497.

Brown, S.J., Warner, J.B. (1985). "Using daily stock returns: the case of event studies". Journal of Financial Economics 14, 3–32.

Byrd, J.W., Hickman, K.A. (1992). "Do outside directors monitor managers? Evidence from tender offer bids". Journal of Financial Economics 32, 195–221.

Cary, W.L. (1974). "Federalism and corporate law: reflections upon Delaware". Yale Law Journal 88, 663–707.

Cheffins, B.R. (2001). "Does law matter? The separation of ownership and control in the United Kingdom". Journal of Legal Studies 30, 459–484.

Coffee, J.C. Jr. (1984). "Regulating the market for corporate control: a critical assessment of the tender offer's role in corporate governance". Columbia Law Review 84, 1145–1296.

Coffee, J.C. Jr. (1985). "The unfaithful champion: the plaintiff as monitor in shareholder litigation". Law and Contemporary Problems 48, 5–81.

Coles, J.L., Meschke, J.F., Lemmon, M.L. (2003). "Structural models and endogeneity in corporate finance". Arizona State University. Working Paper.

Comment, R., Schwert, G.W. (1995). "Poison or placebo? Evidence on the deterrence and wealth effects of modern antitakeover measures". Journal of Financial Economics 39, 3–43.

Core, J.E., Guay, W.R., Rusticus, T.O. (2006). "Does weak governance cause weak stock returns? An examination of firm operating performance and investors' expectations". Journal of Finance 61 (2), 655–687.

Cornell, B., Morgan, R.G. (1990). "Using finance theory to measure damages in fraud on the market cases". UCLA Law Review 37, 883–924.

Council of Economic Advisors (1985). Annual Report. U.S. Government Printing Office, Washington, D.C.

Cutler, D.M., Summers, L.H. (1988). "The costs of conflict resolution and financial distress: evidence from the Texaco-Pennzoil litigation". Rand Journal of Economics 19, 157–172.
Daines, R. (2001). "Does Delaware law improve firm value?" Journal of Financial Economics 62, 525–558.
Daines, R. (2002). "The incorporation choices of IPO firms". New York University Law Review 77, 1559–1611.
Danielson, M.G., Karpoff, J.M. (2002). "Do pills poison operating performance?". Manuscript. (Available in the Social Science Research Network Electronic Paper Collection at: http://papers.ssrn.com/abstract=304647.)
Datta, S., Iskandar-Datta, M. (1996). "Takeover defenses and wealth effects on securityholders: the case of poison pill adoptions". Journal of Banking and Finance 20, 1231–1250.
David, R., Brierley, J.E.C. (1985). Major Legal Systems in the World Today. Stevens and Sons, London.
Davis, G.F., Kim, E.H. (2005). "Would mutual funds bite the hand that feeds them? Business ties and proxy voting". University of Michigan Ross School of Business, manuscript.
DeAngelo, H., Rice, E.M. (1983). "Antitakeover charter amendments and stockholder wealth". Journal of Financial Economics 11, 329–360.
DeAngelo, H., DeAngelo, L. (1989). "Proxy contests and the governance of publicly held corporations". Journal of Financial Economics 23, 29–60.
Del Guercio, D., Hawkins, J. (1999). "The motivation and impact of pension fund activism". Journal of Financial Economics 52, 293–340.
Demsetz, H. (1983). "The structure of ownership and the theory of the firm". Journal of Law and Economics 26, 375–390.
Denis, D.J., Serrano, J.M. (1996). "Active investors and management turnover following unsuccessful control contests". Journal of Financial Economics 40, 239–266.
Denis, D.J., Denis, D.K., Sarin, A. (1997). "Ownership structure and top executive turnover". Journal of Financial Economics 45, 193–221.
Denis, D.K., McConnell, J.J. (2003). "International corporate governance". Journal of Financial and Quantitative Analysis 38, 1–36.
Dennis, D.K., McConnell, J.J. (1986). "Corporate mergers and security returns". Journal of Financial Economics 16, 143–187.
Dent, G.W. Jr. (1986). "Unprofitable mergers: toward a market-based legal response". Northwestern University Law Review 80, 777–806.
Desai, H., Jain, P.C. (1999). "Firm performance and focus: long-run stock market performance following spinoffs". Journal of Financial Economics 54, 75–101.
Dodd, P., Leftwich, R. (1980). "The market for corporate charters: 'unhealthy competition' vs. federal regulation". Journal of Business 53, 259–283.
Downes, G.R. Jr., Houminer, E., Hubbard, R.G. (1999). Institutional Investors and Corporate Behavior. AEI, Washington, D.C.
Dungworth, T., Pace, N. (1990). Statistical Overview of Civil Litigation in the Federal Courts. Rand Institute for Civil Justice, Santa Monica, CA.
Easterbrook, F.H., Fischel, D.R. (1981). "The proper role of a target's management in responding to a tender offer". Harvard Law Review 94, 1161–1204.
Easterbrook, F.H., Fischel, D.R. (1991). The Economic Structure of Corporate Law. Harvard University, Cambridge, MA.
Eisenberg, M.A. (1976). The Structure of the Corporation. Little, Brown, Boston.
Ellert, J.C. (1975). "Mergers, antitrust enforcement and stockholder returns". Journal of Finance 31, 715–732.
Engelmann, K., Cornell, B. (1988). "Measuring the costs of corporate litigation: five case studies". Journal of Legal Studies 17, 377–399.
Fama, E.F. (1990). "Efficient capital markets: II". Journal of Finance 46, 1575–1617.
Fama, E.F., Fisher, L., Jensen, M., Roll, R. (1969). "The adjustment of stock prices to new information". International Economic Review 10, 1–21.
Ferris, S.P., Lawless, R.M., Noronha, G. (2004). "The influence of state legal environments on firm incorporation decisions and values". Journal of Law, Economics and Policy, in press.

Fershtman, C., Judd, K. (1987). "Equilibrium incentives in oligopoly". American Economic Review 77, 927–940.
Fischel, D.R., Bradley, M. (1986). "The role of liability rules and the derivative suit in corporate law: a theoretical and empirical analysis". Cornell Law Review 71, 261–297.
Francis, J., Philbrick, D., Schipper, K. (1994). "Determinants and outcomes in class action securities litigation". University of Chicago Working Paper.
Friend, I., Herman, E.S. (1964). "The S.E.C. through a glass darkly". Journal of Business 37, 382–405.
Friend, I., Westerfield, R. (1975). "Required disclosure and the stock market". American Economic Review 65, 467–472.
Furtado, E.P.H., Karan, V. (1990). "Causes, consequences, and shareholder wealth effects of management turnover: a review of the empirical evidence". Financial Management 19, 60–75.
Gillan, S.L., Hartzell, J.C., Starks, L.T. (2002). "Industries, investment opportunities, and corporate governance structures". Center for Corporate Governance, University of Delaware College of Business and Economics Working Paper No. 2002-003.
Gillan, S.L., Starks, L.T. (2000). "Corporate governance proposals and shareholder activism: the role of institutional investors". Journal of Financial Economics 57, 275–305.
Gompers, P.A., Ishii, J.L., Metrick, A. (2003). "Corporate governance and equity prices". Quarterly Journal of Economics 118, 107–155.
Gordon, J.N. (1991). "Corporations, markets and courts". Columbia Law Review 91, 1931–1988.
Griffin, P.A., Grundfest, J.A., Perino, M.A. (2004). "Stock price response to news of securities fraud litigation: an analysis of sequential and conditional information". Abacus 4, 21–48.
Griliches, Z., Mairesse, J. (1999). "Production functions: the search for identification". Harvard University Working Paper.
Grossman, S.J., Hart, O.D. (1983). "An analysis of the principal-agent problem". Econometrica 51, 7–45.
Hackl, J.W., Testani, R. (1988). "Note, second generation state takeover statutes and shareholder wealth: an empirical study". Yale Law Journal 97, 1193–1231.
Hail, L., Leuz, C. (2003). "International differences in cost of capital: do legal institutions and securities regulation matter?". Manuscript.
Hall, B.J., Liebman, J.B. (1998). "Are CEOs really paid like bureaucrats?" Quarterly Journal of Economics 113, 653–691.
Haslem, B. (2003). "Managerial opportunism during corporate litigation". Indiana University Working Paper.
Hermalin, B.E., Weisbach, M.S. (1988). "The determinants of board composition". Rand Journal of Economics 19, 589–606.
Heron, R.A., Lewellen, W.G. (1998). "An empirical analysis of the reincorporation decision". Journal of Financial and Quantitative Analysis 33, 549–568.
Himmelberg, C.P., Hubbard, R.G., Palia, D. (1999). "Understanding the determinants of managerial ownership and the link between ownership and performance". Journal of Financial Economics 53, 353–384.
H.R. Conference Report (1995). No. 369, 104th Congress, 1st Session.
Hyman, A. (1979). "The Delaware controversy—the legal debate". Journal of Corporate Law 4, 368–398.
Ichimura, H., Lee, L. (1991). "Semiparametric least squares estimation of multiple index models: single equation estimation". In: Barnett, W., Powell, J., Tauchen, G. (Eds.), Nonparametric and Semiparametric Methods in Economics and Statistics. Cambridge University, Cambridge.
Ikenberry, D., Lakonishok, J. (1993). "Corporate governance through the proxy context: evidence and implications". Journal of Business 66, 405–435.
Ikenberry, D., Lakonishok, J., Vermaelen, T. (1995). "Market underreaction to open market share repurchases". Journal of Financial Economics 39, 181–208.
Jahera, J.S., Pugh, W.N. (1991). "State takeover legislation: the case of Delaware". Journal of Law, Economics and Organization 7, 410–428.
Janjigian, V., Bolster, P.J. (1990). "The elimination of director liability and stockholder returns: an empirical investigation". Journal of Financial Research 13, 53–60.
Jarrell, G.A., Poulsen, A.B. (1987). "Shark repellents and stock prices: the effects of antitakeover amendments since 1980". Journal of Financial Economics 19, 127–168.

Jarrell, G.A., Brickley, J.A., Netter, J.M. (1988). "The market for corporate control: the empirical evidence since 1980". Journal of Economic Perspectives 2, 49–68.

Jensen, M.C., Ruback, R.S. (1980). "The market for corporate control: the scientific evidence". Journal of Financial Economics 11, 5–50.

Jensen, M.C., Warner, J.B. (1988). "The distribution of power among corporate managers, shareholders, and directors". Journal of Financial Economics 20, 3–24.

Johnson, M.F., Kasznik, R., Nelson, K.K. (2000). "Shareholder wealth effects of the Private Securities Litigation Reform Act of 1995". Review of Accounting Studies 5, 217–233.

Johnson, M.F., Nelson, K.K., Pritchard, A.C. (2000). "In Re Silicon Graphics Inc.: shareholder wealth effects resulting from the interpretation of the Private Securities Litigation Reform Act's pleading standard". Southern California Law Review 73, 773–810.

Johnson, M.F., Nelson, K.K., Pritchard, A.C. (2007). "Do the merits matter more? The impact of the Private Securities Litigation Reform Act". Journal of Law, Economics and Organization 23, in press.

Kahan, M. (2006). "The demand for corporate law: statutory flexibility, judicial quality, or takeover protection?" Journal of Law, Economics and Organization 22, 340–365.

Kaplan, S.N., Weisbach, M.S. (1992). "The success of acquisitions: evidence from divestitures". Journal of Finance 47, 107–138.

Kamma, S., Weintrop, J., Wier, P. (1988). "Investors' perceptions of the Delaware supreme court decision in Unocal v. Mesa". Journal of Financial Economics 20, 419–430.

Karpoff, J.M., Lott, J.R. Jr. (1993). "The reputational penalty firms bear from committing criminal fraud". Journal of Law and Economics 36, 757–802.

Karpoff, J.M., Lott, J.R. Jr. (1999). "On the determinants and importance of punitive damage awards". Journal of Law and Economics 42, 527–573.

Karpoff, J.M., Malatesta, P.H. (1989). "The wealth effects of second generation takeover legislation". Journal of Financial Economics 25, 291–322.

Karpoff, J.M., Malatesta, P.H. (1995). "State takeover legislation and share values: the wealth effects of Pennsylvania's Act 36". Journal of Corporate Finance 1, 367–382.

Kim, E.H., Singal, V. (1993). "Mergers and market power: evidence from the airline industry". American Economic Review 83, 549–569.

Klausner, M. (1995). "Corporations, corporate law, and networks of contracts". Virginia Law Review 81, 757–852.

Klein, A. (1998). "Firm performance and board committee structure". Journal of Law and Economics 41, 275–303.

Kothari, S.P., Warner, J.B. (1997). "Measuring long-horizon security price performance". Journal of Financial Economics 43, 301–340.

Kothari, S.P., Warner, J.B. (2007). "Econometrics of event studies". In: Eckbo, B.E. (Ed.), Handbook of Corporate Finance, vol. 1. Elsevier/North-Holland, Amsterdam.

Lambert, R.A., Larcker, D.F. (1985). "Golden parachutes, executive decision-making and shareholder wealth". Journal of Accounting and Economics 7, 179–203.

Lamont, O.A., Thaler, R.H. (2003). "Can the market add and subtract? Mispricing in tech stock carve-outs". Journal of Political Economy 111, 227–268.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A. (2006). "What works in securities laws?" Journal of Finance 61, 1–32.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., Vishny, R. (1997). "Legal determinants of external finance". Journal of Finance 52, 1131–1150.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., Vishny, R. (1998). "Law and finance". Journal of Political Economy 106, 1113–1155.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., Vishny, R. (1999). "Corporate ownership around the world". Journal of Finance 54, 471–517.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., Vishny, R. (2000). "Investor protection and corporate governance". Journal of Financial Economics 58, 3–28.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., Vishny, R. (2002). "Investor protection and corporate valuation". Journal of Finance 57, 1147–1170.

Lee, C.I., Rosenstein, S., Rangan, N., Davidson, W.N. III (1992). "Board composition and shareholder wealth: the case of management buyouts". Financial Management 21, 58–72.

Linn, S.C., McConnell, J.J. (1983). "An empirical investigation of antitakeover amendments on common stock prices". Journal of Financial Economics 11, 361–399.

Loughran, T., Ritter, J.R. (1995). "The new issues puzzle". Journal of Finance 50, 23–52.

Lyon, J.D., Barber, B.M., Tsai, C. (1999). "Improved methods for tests of long-run abnormal stock returns". Journal of Finance 54, 165–201.

Macey, J.R., Miller, G.P. (1987). "Toward an interest-group theory of Delaware corporate law". Texas Law Review 65, 469–523.

MacKinlay, A.C. (1997). "Event studies in economics and finance". Journal of Economic Literature 35, 13–39.

Maddala, G.S. (1983). Limited-dependent and Qualitative Variables in Econometrics. Cambridge University, New York, NY.

Mahla, C.R. (1991). "State takeover statutes and shareholder wealth". University of North Carolina, Chapel Hill, NC, dissertation.

Mahoney, P.G. (1999). "The stock pools and the Securities Exchange Act". Journal of Financial Economics 51, 343–369.

Mahoney, P.G. (2001a). "The political economy of the Securities Act of 1933". Journal of Legal Studies 30, 1–31.

Mahoney, P.G. (2001b). "The common law and economic growth: Hayek might be right". Journal of Legal Studies 30, 503–525.

Marais, L., Schipper, K., Smith, A. (1989). "Wealth effects of going private for senior securities". Journal of Financial Economics 23, 155–191.

Margotta, D.G. (1991). "Stock price effects of Pennsylvania Act 36". Northeastern University, manuscript.

Margotta, D.G., MacWilliams, T.P., MacWilliams, V.B. (1990). "An analysis of the stock price effect of the 1986 Ohio takeover legislation". Journal of Law, Economics and Organization 6, 235–251.

Martin, K.J., McConnell, J.J. (1991). "Corporate performance, corporate takeovers, and management turnover". Journal of Finance 46, 671–688.

McConnell, J.J., Ozbilgin, M., Wahal, S. (2001). "Spinoffs, ex ante". Journal of Business 74, 245–280.

Mikkelson, W.H., Partch, M.M. (1997). "The decline of takeovers and disciplinary managerial turnover". Journal of Financial Economics 44, 205–228.

Mitchell, M.L., Netter, J.M. (1994). "The role of financial economics in securities fraud cases: applications at the SEC". Business Lawyer 49, 545–590.

Morck, R., Shleifer, A., Vishny, R. (1988). "Management ownership and market valuation: an empirical analysis". Journal of Financial Economics 20, 293–316.

Myerson, R.B. (1979). "Incentive compatibility and the bargaining problem". Econometrica 47, 61–74.

Netter, J., Poulsen, A. (1989). "State corporation laws and shareholders: the recent experience". Financial Management 18, 29–40.

Perino, M.A. (2003). "Did the Private Securities Litigation Reform Act work?" University of Illinois Law Review 2003, 913–976.

Peterson, P. (1988). "Reincorporation motives and shareholder wealth". Financial Review 23, 151–160.

Pontiff, J., Shleifer, A., Weisbach, M.S. (1990). "Reversions of excess pension assets after takeovers". Rand Journal of Economics 21, 600–613.

Pound, J. (1987). "The effect of antitakeover amendments on takeover activity: some direct evidence". Journal of Law and Economics 30, 353–368.

Prabhala, N.R. (1997). "Conditional methods in event studies and an equilibrium justification for standard event-study procedures". Review of Financial Studies 10, 1–38.

Prince, D.W., Rubin, P.H. (2002). "The effects of product liability litigation on the value of firms". American Law and Economics Review 4, 44–87.

Pugh, W.N., Jahera, J.S. (1990). "State antitakeover legislation and shareholder wealth". Journal of Financial Research 13, 221–231.
Roe, M.J. (2000). "Political preconditions to separating ownership from corporate control". Stanford Law Review 53, 539–606.
Romano, R. (1985). "Law as a product: some pieces of the incorporation puzzle". Journal of Law, Economics and Organization 1, 225–283.
Romano, R. (1987). "The political economy of takeover statutes". Virginia Law Review 73, 111–199.
Romano, R. (1990). "Corporate governance in the aftermath of the insurance crisis". Emory Law Review 39, 1155–1189.
Romano, R. (1991). "The shareholder suit: litigation without foundation?" Journal of Law, Economics and Organization 7, 55–87.
Romano, R. (1993a). The Genius of American Corporate Law. AEI, Washington, D.C.
Romano, R. (1993b). "What is the value of other constituency statutes to shareholders?" University of Toronto Law Journal 43, 533–542.
Romano, R. (1993c). "Public pension fund activism in corporate governance reconsidered". Columbia Law Review 93, 795–853.
Romano, R. (1996). "Corporate law and corporate governance". Industrial and Corporate Change 5, 277–339.
Romano, R. (2001). "Less is more: making institutional investor activism a valuable mechanism of corporate governance". Yale Journal on Regulation 18, 174–251.
Romano, R. (2002). The Advantage of Competitive Federalism for Securities Regulation. AEI, Washington, D.C.
Romano, R. (2003). "Does confidential proxy voting matter?" Journal of Legal Studies 32, 465–509.
Romano, R. (2005). "The Sarbanes-Oxley Act and the making of quack corporate governance" . Yale Law Journal 114, 1521–1611.
Rosenstein, S., Wyatt, J.G. (1990). "Outside directors, board independence, and shareholder wealth". Journal of Financial Economics 26, 175–191.
Rosett, J.G. (1990). "Do union wealth concessions explain takeover premiums? The evidence on contract wages". Journal of Financial Economics 27, 263–282.
Ross, S.A. (2005). "Things I don't know about finance". (Presentation available at http://islandia.law.yale.edu/ccl/index.html.)
Rossi, S., Volpin, P. (2004). "Cross-country determinants of mergers and acquisitions". Journal of Financial Economics 74, 277–304.
Ryngaert, M. (1988). "The effects of poison pill securities on shareholder wealth". Journal of Financial Economics 20, 377–417.
Ryngaert, M., Netter, J. (1988). "Shareholder wealth effects of the Ohio antitakeover law". Journal of Law, Economics and Organization 4, 373–383.
Schipper, K., Thompson, R. (1983). "Evidence on the capitalized value of mergers activity for acquiring firms". Journal of Financial Economics 11, 85–119.
Schumann, L. (1988). "State regulation of takeovers and shareholder wealth: the case of New York's 1985 takeover statutes". Rand Journal of Economics 19, 557–567.
Schwert, G.W. (2003). "Anomalies and market efficiency". In: Constantinides, G., Harris, M., Stulz, R. (Eds.), Handbook of the Economics of Finance. North-Holland, Amsterdam, pp. 937–972.
Seyhun, N. (1998). Investment Intelligence from Insider Trading. MIT, Cambridge, MA.
Shanley, M.G., Peterson, M.A. (1987). Posttrial Adjustments to Jury Awards. RAND Institute for Civil Justice, Santa Monica, CA.
Sidak, J.G., Woodward, S.E. (1990). "Corporate takeovers, the commerce clause, and the efficient anonymity of shareholders". Northwestern University Law Review 84, 1092–1118.
Singh, A., Singh, A., Weisse, B. (2002). "Corporate governance, competition, the new international financial architecture and large corporations in emerging markets". ESRC Centre for Business Research, University of Cambridge Working Paper No. 250.
Skinner, D. (1995). "Empirical evidence on the relation between earnings disclosures, firm characteristics, and shareholder lawsuits". Working Paper, University of Michigan.

Spencer, L. (1992). "The tort tax" Forbes February 17, 40.
Spiess, D.K., Tkac, P.A. (1997). "The Private Securities Litigation Reform Act of 1995: the stock market casts its vote". Managerial and Decision Economics 18, 545–561.
Stigler, G.J. (1964). "Public regulation of the securities market". Journal of Business 37, 117–142.
Stout, L. (2005). "Share price as a poor criterion for good corporate law" . Berkeley Business Law Journal 3, 43–58.
Stulz, R.M. (1988). "Managerial control of voting rights: financing policies and the market for corporate control". Journal of Financial Economics 20, 25–54.
Subramanian, G. (2002). "The influence of antitakeover statutes on incorporation choice: evidence on the 'race' debate and antitakeover overreaching". University of Pennsylvania Law Review 150, 1795–1873.
Subramanian, G. (2003). "Bargaining in the shadow of takeover defenses". Yale Law Journal 113, 621–686.
Subramanian, G. (2004). "The disappearing Delaware effect". Journal of Law, Economics and Organization 20, 32–59.
Szewczyk, S.H., Tsetsekos, G.P. (1992). "State intervention in the market for corporate control: the case of Pennsylvania Senate Bill 1320". Journal of Financial Economics 31, 3–23.
Thomas, R.S., Martin, K.J. (1998). "Should labor be allowed to make shareholder proposals?" Washington Law Review 73, 41–80.
Thompson, R.B., Thomas, R.S. (2003). "The new look of shareholder litigation: acquisition-oriented class actions". Vanderbilt Law Review 57, 133–209.
Vagts, D. (2002). "Comparative company law—the new wave". In: Schweizer, R., Burkert, H., Gasser, U. (Eds.), Festschrift für Jean Nicolas Druey zum 65. Geburtstag, Schulthess, Zurich, pp. 595–605.
Vancil, R.F. (1987). Passing the Baton: Managing the Process of CEO Selection. Harvard Business School Press, Boston.
Van Nuys, K. (1993). "Corporate governance through the proxy process: evidence from the 1989 Honeywell proxy solicitation". Journal of Financial Economics 34, 101–132.
Wang, J. (1995). "Performance of reincorporated firms". Yale School of Management, manuscript.
Weisbach, M.S. (1988). "Outside directors and CEO turnover". Journal of Financial Economics 20, 431–460.
Weiss, E.J., White, L.J. (1987). "Of econometrics and indeterminacy: a study of investors' reactions to 'changes' in corporate law". California Law Review 75, 551–607.
Weiss, E.J., White, L.J. (2004). "File early, then free ride: how Delaware law (mis)shapes shareholder class actions". Vanderbilt Law Review 57, 1797–1881.
Welch, I. (2000). "Views of financial economists on the equity premium and on professional controversies". Journal of Business 73, 501–538.
Winter, R.K. (1977). "State law, shareholder protection, and the theory of the corporation". Journal of Legal Studies 6, 251–292.
Winter, R.K. (1993). Forward to R. Romano, The Genius of American Corporate Law. AEI, Washington, D.C., pp. ix–xiii.
Wurgler, J. (2000). "Financial markets and the allocation of capital". Journal of Financial Economics 58, 187–214.
Zweigert, K., Kotz, H. (1987). An Introduction to Comparative Law. Clarendon, Oxford.

Case References

Basic v. Levinson 485 U.S. 224 (1988).
Business Roundtable v. SEC, 905 F.2d 406 (D.C. Cir. 1990).
Unitrin v. American General Corp., 651 A.2d 1361 (Del. 1995).
Unocal Corp. v. Mesa Petroleum Co., 493 A.2d 946 (Del. 1985).

Statutes and Regulations

Private Securities Litigation Reform Act of 1995, Pub. Law 104-67, 109 Stat. 737, modifying 15 U.S.C. §§77a et seq. and §§78a et seq.

Sarbanes-Oxley Act of 2002, Pub. Law 107-204, section 301, codified at 15 U.S.C. §78j-1(m).
Securities Act of 1933, 15 U.S.C. §§77a et seq.
Securities Exchange Act of 1934, 15 U.S.C. §§78a et seq.
Securities Exchange Act Rules 13e-4(f)(8), 17 C.F.R. 240.13e4-(f)(8).
Securities Exchange Act Rule 14a-8, 17 C.F.R. 240.14a-8.
Securities Exchange Act Rule 19c-4, 17 C.F.R. 240.19c-4.
Williams Act, Act of July 29, 1968, Pub. L. No. 90-439, 82 Stat. 454, amending 15 U.S.C. §§78a et seq.

Chapter 14

BANKRUPTCY LAW

MICHELLE J. WHITE*

Department of Economics, University of California, San Diego

Contents

* Professor of Economics, University of California, San Diego, and Research Associate, NBER. I am very grateful to Lucian Bebchuk for comments and to the National Science Foundation for research support under grant number 0212444. Portions of this chapter were presented at Harvard Law School, University of Southern California Law Center, and the 2005 ALEA Conference in New York and I benefited from comments by participants at these talks.

Handbook of Law and Economics, Volume 2
Edited by A. Mitchell Polinsky and Steven Shavell

DOI: 10.1016/S1574-0730(07)02014-2

Abstract

Bankruptcy is the legal process whereby financially distressed firms, individuals, and occasionally governments resolve their debts. The bankruptcy process for firms plays a central role in economics, because competition drives inefficient firms out of business, thereby raising the average efficiency level of those remaining. The main economic function of corporate bankruptcy is to reduce the cost of default by having a government-sponsored procedure that resolves all debts simultaneously. The main economic function of personal bankruptcy is to provide partial consumption insurance to individual debtors and therefore reduce the social cost of debt. This chapter surveys theoretical and empirical research on both types of bankruptcy.

Keywords

Corporate bankruptcy, personal bankruptcy, small business, financial distress, reorganization, liquidation, absolute priority rule (or APR), limited liability, cramdown, prepack (or prepackaged bankruptcy), human capital, Chapter 11, Chapter 7, Chapter 13, option

JEL classification: K2, G3, G33, H42

1. Introduction

Bankruptcy is the legal process by which financially distressed firms, individuals, and occasionally governments resolve their debts. The bankruptcy process for firms plays a central role in economics, because competition drives the most inefficient firms out of business, thereby raising the average efficiency level of those remaining. Consumers benefit because the remaining firms produce goods and services at lower costs and sell them at lower prices. The legal mechanism through which most financially distressed firms resolve their debts and exit the market is bankruptcy. Bankruptcy is also the process by which individuals and married couples in financial distress resolve their debts, although financially distressed individuals—unlike firms—do not shut down or exit. Governments sometimes also use bankruptcy to resolve their debts. Like individuals but unlike firms in financial distress, they do not shut down.

This chapter discusses the economics of bankruptcy law. Since the literatures on corporate and personal bankruptcy have developed in isolation of each other, a goal of this chapter is to draw out parallels between them. It is useful to start by defining terms. Corporate bankruptcy refers to the bankruptcy of large- and medium-sized businesses, which for convenience I assume to be organized as corporations. Personal bankruptcy refers to the bankruptcies of individual households and small businesses. Small business bankruptcy is treated as part of personal bankruptcy, since small businesses are owned by individuals or partners who are legally responsible for their businesses' debts. When their businesses fail, owners often file for bankruptcy so that their businesses' debts will be discharged. Even when small businesses are incorporated, owners often guarantee the debts of their businesses, so that personal bankruptcy law applies at least in part.

Regardless of whether the debtor is a business or an individual, bankruptcy law provides a collective framework for simultaneously resolving all debts when debtors' assets are less than their liabilities. This includes both rules for determining how much of the debtor's assets must be used to repay debt and rules for determining how those assets are divided among creditors. Thus bankruptcy is concerned with both the size of the pie—the total amount paid to creditors—and how the pie is divided.

For corporations in financial distress, both the size of the pie and its division depend on whether the corporation liquidates versus reorganizes in bankruptcy and corporate bankruptcy law includes rules for deciding whether reorganization or liquidation will occur. When corporations liquidate, the size of the pie is all of the firm's assets. The size of the pie reflects the doctrine of limited liability, which exempts corporate shareholders from liability for the corporation's debts beyond loss of their shares. The proceeds of liquidating the corporation's assets are used to repay creditors. The division of the pie follows the absolute priority rule (APR), which carries into bankruptcy the nonbankruptcy rule that all creditors must be paid in full before equityholders receive anything. The APR also determines the division of the pie among creditors and requires that higher-priority creditors be repaid in full before lower-ranking creditors receive anything. Thus under the APR, each class of creditors either receives full payment of

its claims or nothing at all (except that the lowest-ranking class of creditors to be repaid receives partial payment).

When corporations reorganize rather than liquidate in bankruptcy, the reorganized corporation retains most or all of its assets and continues to operate. The funds to repay creditors then come from the reorganized firm's future earnings rather than from sale of its assets. The rules for dividing the pie in reorganization also differ from those in liquidation. Instead of dividing the assets so that creditors receive either full payment or nothing, most creditors receive partial payment and pre-bankruptcy equityholders receive some of the reorganized firm's new shares. Bankruptcy law again provides a procedure for determining both the size and division of the pie, but the procedure involves a negotiation process rather than a formula.

For individuals in financial distress, bankruptcy also provides a framework for resolving all of the individual's debts. Again the procedure includes both rules for determining how much of the consumer's assets must be used to repay debt (the size of the pie) and rules for dividing the assets among creditors (the division of the pie). In determining the size of the pie, personal bankruptcy law plays a role analogous to that of limited liability for corporate shareholders, since it determines how much of their assets individual debtors must use to repay their debts. Unlike corporations, individual debtors in bankruptcy are not required to use all of their assets to repay their debts. Instead, personal bankruptcy specifies exemption levels, which are maximum amounts of both financial wealth and post-bankruptcy earnings that bankrupt individuals are allowed to keep. Amounts in excess of the exemption levels must be used to repay debt. To divide the pie, personal bankruptcy specifies a division rule. As in corporate bankruptcy, the division rule may either be the APR or a rule under which all creditors receive partial payment.

An important difference between personal and corporate bankruptcy procedures is that true liquidation never occurs in personal bankruptcy (even though the Chapter 7 personal bankruptcy procedure in the U.S. is called liquidation). Debtors' wealth consists of two components: financial wealth (including home equity) and human capital. The only way to liquidate the human capital portion of individual debtors' wealth would be to sell debtors into slavery—as the Romans did. Since slavery is no longer used as a penalty for bankruptcy, all personal bankruptcy procedures are forms of reorganization in which individual debtors keep their human capital and the right to use it (or not use it) after bankruptcy.[1]

The economic objectives are similar in corporate and personal bankruptcy. One objective of bankruptcy is to repay creditors enough that credit remains available on reasonable terms. Reduced access to credit makes debtors worse off because businesses

[1] Both Britain and the U.S. used debtors' prison as a punishment for bankruptcy during the nineteenth century and, in earlier periods, Britain occasionally used the death penalty against debtors who defrauded their creditors. While prison and the death penalty waste debtors' human capital, they presumably cause debtors to use their financial assets to repay debt even though the assets could otherwise be hidden from creditors. See Baird (1987).

need to borrow in order to start up and grow and individuals benefit from borrowing to smooth consumption. On the other hand, repaying more to creditors harms debtors by making it more difficult for financially distressed firms to survive and more onerous for financially distressed individuals to work. Both the optimal size and division of the pie in bankruptcy are affected by this tradeoff. Another way of expressing the same objective is to give both corporate and personal debtors an incentive to invest and consume efficiently before and after they become financial distressed. A second objective of both types of bankruptcy is to prevent creditors from harming debtors by racing to be first to collect. This is because aggressive collection efforts by creditors may force debtor firms to shut down even though the best use of their assets is to continue operating and may cause individual debtors to lose their jobs (if creditors repossess debtors' cars or garnish debtors' wages). Finally, personal bankruptcy law has an additional objective that has no counterpart in corporate bankruptcy—to provide individual debtors with partial consumption insurance by discharging debt when repayment would cause a substantial reduction in debtors' consumption levels. This is because if consumption falls substantially, long-term harm may occur, including debtors' children leaving school prematurely in order to work or debtors' medical conditions going untreated and becoming disabilities.[2]

In 1984, there were approximately 62,000 business bankruptcy filings and 286,000 filings by individuals and married couples. By twenty years later in 2004, the number of business bankruptcy filings had fallen in half to 34,000, while the number of filings by individuals and married couples had increased more than five-fold to 1,583,000.[3] Concern about the rising number of individual bankruptcies led Congress to adopt reforms of personal bankruptcy law in 2005.

Part A of this chapter deals with corporate bankruptcy and Part B with individual and small business bankruptcy. Each part contains separate sections that outline the law, discuss theoretical research, and present the empirical evidence. A third topic that is not discussed—because it has received little attention from economists—is governmental or sovereign bankruptcy.[4]

[2] Baird (1987) points out that discharge of debt in bankruptcy originally applied only to merchants and was intended to prevent them from being forced to close their businesses if an adverse event occurred for reasons beyond their control (such as a merchant ship sinking). Thus discharge provided a type of insurance to business owners. Over time, discharge expanded from covering only business debt to covering individual debt. But it gradually became less important for business debt as the corporate form and limited liability developed.

[3] See *Statistical Abstract of the United States*, 1988, table 837, and Administrative Office of the U.S. Courts (for recent years).

[4] Chapter 9 of the U.S. Bankruptcy Code provides a bankruptcy procedure for local governments. It does not apply to state or county governments and has been used only rarely. See McConnell and Picker (1993) for discussion. There is currently no bankruptcy procedure for countries that default, although the International Monetary Fund has considered establishing one. There are several important differences between sovereign bankruptcy and corporate/personal bankruptcy. One is that creditors have very limited collection options against sovereign debtors, so that the race to be first among creditors is less important. Another is that the

Part A: Corporate bankruptcy

2. Legal background—corporate bankruptcy law

The U.S. has two separate bankruptcy procedures for corporations in financial distress, Chapter 7 for liquidation and Chapter 11 for reorganization. In Section 2 I discuss the two Chapters separately and then discuss out-of-bankruptcy resolution of financial distress.

2.1. Chapter 7 liquidation

When a corporation firm files under Chapter 7, the bankruptcy court appoints a trustee who shuts the firm down, sells its assets, distributes the proceeds to the firm's creditors, and dissolves the corporation. Legal efforts by creditors to collect from the firm are terminated and all creditors' claims must be resolved in the bankruptcy proceeding, regardless of whether they come due in the present or the future. The APR is used to determine the division of the liquidated assets among creditors. The APR carries over to the bankruptcy context the non-bankruptcy rule that creditors must be paid in full before equityholders receive anything, thus preserving creditors' non-bankruptcy rights vis-à-vis equityholders. But the APR also advances other claims so that they take priority over debt claims in bankruptcy. The highest priority under the APR goes to the administrative expenses of the bankruptcy process itself (including filing fees, lawyers' fees and the trustee's fee); followed by claims taking statutory priority (including tax claims, rent claims, and some unpaid wage and benefit claims); followed by unsecured creditors' claims (including trade creditors, bondholders, and those holding tort judgments against the firm). Equity has the lowest priority. Claims in each class are paid in full until funds are exhausted.

Within the class of unsecured claims, various rankings are consistent with the APR. If there are subordination agreements that place certain unsecured claims above others, then these are followed in bankruptcy. In the literature, the best-known ranking is the "me-first" rule of Fama and Miller (1972), under which unsecured claims take priority in chronological order based on when creditors made their loans. The opposite of the "me-first" rule is the "last-lender-first" rule, under which priority is in reverse chronological order. If there are no subordination agreements, then all unsecured claims have equal priority.

Secured creditors are outside the priority ordering. They have bargained with the firm for the right to seize a particular asset if the firm defaults and/or files for bankruptcy. Thus only assets that are not subject to secured creditors' liens are included in the pool

cost of default is very high, since default usually leads to a severe recession in the country's economy. Unlike bankrupt corporations but like bankrupt individuals, countries can only be reorganized ("restructured"), not liquidated. A final difference is that when countries default, the IMF plays an important role in restructuring negotiations. See White (2002) for discussion.

of assets used to pay other creditors. When firms liquidate in bankruptcy, often all or nearly all of their assets are subject to secured creditors' liens, so that other creditors receive nothing.

When creditors realize that a debtor firm might be insolvent, they have an incentive to race against each other to be first to collect. This is because, as in a bank run, the earliest creditors to collect will be paid in full, but later creditors will receive nothing. The race to be first is inefficient, since the first creditor to collect may seize assets that the firm needs for its operations and, as a result, may force the firm to shut down. Early shutdown wastes resources because the piecemeal value of the firm's assets may be less than their value if the assets are kept together and the firm sold as a going concern. However the existence of bankruptcy mutes creditors' incentive to race to be first. This is because when one creditor wins the race and tries to collect by seizing assets, the firm's managers are likely to file for bankruptcy. And because bankruptcy is a collective procedure that settles all claims at once according to the APR, a bankruptcy filing deprives creditors of their reward for winning the race. Muting creditors' incentive to race to be first by imposing a collective procedure for resolving all of the firm's debts is the traditional economic justification for bankruptcy (Jackson, 1986).

But bankruptcy does not abolish creditors' incentive to compete with each other. Instead, it replaces the race to be first to collect with a competition among creditors to leapfrog over each other in the priority ordering. The most common method by which creditors raise their priority is to shift from unsecured to secured status. They do this by negotiating with managers to renew their loans in return for obtaining a lien on a particular asset owned by the firm or, if the creditor is a bank, by requiring that the firm keep funds in an account at the bank (since these funds act as collateral for the bank's loan). If the firm is planning to file under Chapter 11 rather than Chapter 7, then another leapfrogging method is for creditors to raise their priority by renewing their loans after the firm files for bankruptcy, since doing so makes the loan an administrative expense of bankruptcy that takes highest priority. But when creditors compete to raise their priority in bankruptcy, the result is often that firms delay filing for bankruptcy because creditors renew their loans in return for higher priority. This delay is inefficient if the best use of the firm's assets is something other than their current use.

Bankruptcy liquidation procedures in other countries are similar to the U.S. procedure. But in the United Kingdom, one type of creditor, called a "floating charge" creditor, has the right to prevent managers from filing for bankruptcy. If the firm defaults, the floating charge creditor may liquidate any assets of the firm that are not subject to secured creditors' claims. Only after the floating charge creditor is repaid in full does the bankruptcy trustee begin to liquidate the firm's remaining assets for the benefit of other creditors. The partial liquidation by the floating charge creditor may cause firms to shut down even though their assets are more valuable if they continue to operate.[5]

[5] Webb (1991) analyzes U.K. bankruptcy procedures as a prisoner's dilemma and argues that, as a result, too much liquidation occurs. See also Franks and Sussman (2005).

2.2. Chapter 11 reorganization

In the U.S., managers of corporations in financial distress have the right to choose between filing for bankruptcy liquidation under Chapter 7 versus for bankruptcy reorganization under Chapter 11. Under Chapter 11, the firm continues to operate and pre-bankruptcy managers usually remain in control as "debtors-in-possession." A reorganization plan must eventually be adopted that resolves all of the firm's debts. Under the plan, firms repay part or all of their debt from future earnings, rather than from selling their assets.

Chapter 11 includes a number of provisions that are intended to aid financially distressed firms and increase the likelihood that they will continue operating. Creditors' efforts to collect from the firm are stayed and debtor firms cease making interest and principle payments to creditors until a reorganization plan goes into effect (although the firm must continue paying interest on secured loans). Also with the bankruptcy court's approval, firms in Chapter 11 may obtain new loans and give post-bankruptcy lenders highest priority, even though much of the payoff to post-bankruptcy creditors is likely to come at the expense of pre-bankruptcy creditors. This gives firms in Chapter 11 a new source of working capital. Also, firms in Chapter 11 are allowed to reject their unprofitable contracts and their traditional pension plans. Penalties for breach of contract become unsecured debts, so that they receive only a fractional payoff; while responsibility for meeting the obligations of under-funded pension plans goes to the Pension Benefit Guaranty Corporation—a U.S. government agency. Firms that reorganize successfully also escape the obligation to pay taxes on debt forgiveness until they become profitable. These provisions greatly improve the cash flow of firms in Chapter 11.

Firms in Chapter 11 must adopt reorganization plans that resolve all of their debts. Because the reorganized firm retains some or all of its pre-bankruptcy assets and pays creditors from its future earnings, the reorganization plan determines both the size of the pie and its division among creditors. Bankruptcy law affects the size and division of the pie by setting procedures both for bargaining over the terms of reorganization plans and for adopting them. For at least the first four months after the bankruptcy filing, managers have the exclusive right to propose a reorganization plan and creditors have only a take-it-or-leave-it choice. Managers' exclusive right to propose the plan reduces the size of the pie, because managers have an incentive to propose the smallest pie that creditors will accept. Furthermore, bankruptcy judges often extend managers' exclusivity period and this also reduces the size of the pie, since additional delay makes creditors willing to accept less. The most commonly-used procedure for adopting a reorganization plan is a voting procedure. Under it, each class of creditors must vote in favor of the plan by a margin of at least two-thirds in amount and one-half in number of claims and, in addition, two-thirds of all pre-bankruptcy equityholders must vote in favor. The less-than-100% voting requirement also reduces the size of the pie, because the plan does not have to satisfy the demands of holdout creditors in each class. Also the requirement that all classes of creditors and pre-bankruptcy equityholders vote in favor of the plan

means that even low-priority creditors and equityholders receive positive payoffs in reorganization.[6]

The rules of Chapter 11 also provide some protection for creditors. Reorganization plans that have met the voting requirements for adoption must also be confirmed by the bankruptcy judge. For a plan to be confirmed, the judge must decide that it meets the "best interest of creditors" test, which requires that each class of creditors receive at least what it would have received if the firm liquidated under Chapter 7. If the reorganization plan was rejected by one or more classes of creditors, then the judge can use "cramdown" to confirm the plan. Cramdown requires that classes of creditors that have rejected the plan receive either full payment of their claims over the period of the plan (usually 6 years) or else that all lower-ranking classes of creditors receive nothing. Alternately, the judge may allow creditors to offer their own reorganization plans, may replace managers, or may order that the firm be sold as a going concern under Chapter 11 or liquidated under Chapter 7. If the firm is sold under either Chapter, then the proceeds are distributed according to the APR. Thus, regardless of how firms emerge from Chapter 11, creditors must either receive as much or more than they would receive if the firm liquidated under Chapter 7.

Chapter 11 thus substitutes a bargaining process and a voting procedure for the actual sale of firms' assets that occurs in Chapter 7. In theory, the overall size of the pie and each creditor's individual slice must be at least as large in reorganization as in liquidation, since the "best interest of creditors" test requires that each class of creditors receive as much or more in reorganization as in liquidation. But in practice the size of the pie in reorganization could be smaller than in liquidation. This is because managers of large corporations rarely choose Chapter 7 when they file for bankruptcy, so that when large corporations liquidate, it is generally only after they have operated for prolonged periods in Chapter 11. While in Chapter 11, managers have little incentive to operate their firms efficiently and often bankruptcy court supervision fails to prevent waste and asset-stripping. When these firms eventually liquidate, the value of their assets tends to be very low. This means that even a low payoff to creditors in reorganization exceeds what they expect to receive in liquidation.[7] In addition, the division of the pie differs sharply in reorganization versus liquidation. In liquidation, high-priority creditors receive full payment and lower-priority creditors and equity receiving nothing;

[6] See Bebchuk and Chang (1992) for a common knowledge model of the bargaining process in Chapter 11 that uses the Rubinstein alternating offer bargaining game. They show how rules that favor managers/equity, such as giving managers the exclusive right to propose the first reorganization plan and requiring that the class of equityholders consents to the plan, reduce the amount that creditors receive. Other models of bargaining in Chapter 11 include Brown (1989), Baird and Picker (1991), and Aivazian and Callen (1983).

[7] The best-known example is Eastern Airlines, which filed for bankruptcy under Chapter 11 in 1989 and continued to operate for nearly two years. While in bankruptcy, its value fell by $2 billion. Many of its assets were sold to fund continued operating losses. When it finally shut down, secured creditors received 82% of their claims, unsecured creditors received 11%, and equity received nothing. See Weiss and Wruck (1998) for a detailed analysis.

while in reorganization, each class of creditors receives partial payment and equity receives some of the shares of the reorganized firm. Unsecured creditors and equity must receive something in order to obtain their votes for the reorganization plan, so that they get more in reorganization than in liquidation. But secured creditors usually receive less, because Chapter 11 delays or prevents them from seizing their collateral and the interest they receive is often insufficient to compensate them for the delay. Transfers from higher-priority to lower-priority creditors and/or from creditors to equityholders under Chapter 11 are referred to in the literature as "deviations from the APR." As will be discussed below, many economists have argued that the negotiation process in reorganization is itself economically inefficient and should be replaced.

The United Kingdom, France and Germany have all adopted new bankruptcy procedures recently that were intended to encourage reorganization of firms in financial distress. These procedures differ substantially from Chapter 11 and also differ substantially among themselves. In all three countries, pre-bankruptcy managers are given much less power over the reorganization process than they have in Chapter 11. Instead, the bankruptcy judge or an official appointed by the judge decides whether the firm will shut down or reorganize and, if reorganization is chosen, formulates the reorganization plan. In France, bankruptcy officials appointed to decide whether firms in bankruptcy will be liquidated or reorganized have "safeguarding the business" and saving jobs as their primary objectives. However in the United Kingdom and Germany, bankruptcy procedures are more pro-creditor than in the U.S. or France and reorganization is less likely to occur.[8]

2.3. *Non-bankruptcy workouts*

Because bankruptcy involves high transactions costs, managers of corporations in financial distress often attempt to avoid it by renegotiating the firm's debts outside of bankruptcy. These renegotiations, called workouts, are common in the U.S. (see below for evidence).

Workout negotiations usually involve managers proposing a plan for creditors to forgive part of the firm's debt and creditors deciding whether to accept or reject. Economists have pointed out two reasons why workouts tend to fail. One is the problem of strategic default, meaning that if creditors accept workout proposals, then managers have an incentive to offer them even when their firms are not in financial distress. Creditors can only discourage strategic default by rejecting workouts. The second is that individual creditors have an incentive to reject workout proposals and act as holdouts. This is because if most creditors accept the workout, then the debtor firm will repay the holdouts in full or at least strike a better deal with them. But if all creditors choose to be

[8] For comparisons between corporate bankruptcy reorganization procedures in the U.S. and other countries, see Franks, Nybourg, and Torous (1996), White (1996), Berkovitch and Israel (1999), and Franks and Sussman (2005).

holdouts, then workout proposals will fail. Managers in turn have two ways to increase the probability that workout proposals succeed. One is that if the workout proposal is supported by at least two-thirds of creditors in each class (by value), then managers can file for bankruptcy under Chapter 11 and use the workout proposal as the firm's reorganization plan. This is because, in bankruptcy, only a two-thirds majority of each class of creditors is needed for adoption of the plan. Using a workout proposal as a Chapter 11 reorganization plan is referred to as a prepackaged bankruptcy, or "prepack." Even though prepacks involve a bankruptcy filing, they are much quicker and less costly than normal bankruptcies. Managers' other method of increasing the probability that workouts are accepted is to make "coercive offers." Under the Trust Indenture Act of 1939, the financial terms of a bond issue cannot be changed outside of bankruptcy without the unanimous consent of bondholders, but non-financial terms can be changed by majority vote. Therefore managers offer a workout that involves a reduced payment to bondholders combined with changes in the non-financial terms that make the bond issue less valuable—such as ending public trading. If a majority of bondholders accepts the offer, then the changes in the non-financial terms go into effect and the holdouts are made worse off. Coercive offers give bondholders an incentive to accept workouts.[9]

As discussed above, individual creditors also have an incentive to improve their position in the priority ordering by negotiating individually with managers before managers propose a workout or file for bankruptcy. Banks and other short-term creditors have frequent opportunities to initiate negotiate with managers, since their loans come due frequently and are generally renegotiated and renewed. Long-term debts come due less frequently, but debt contracts contain clauses that allow creditors to declare the loan in default whenever any pre-specified event occurs, such as the firm's working capital falling below a certain level. Default accelerates the due date of the loan from the future to the present and therefore presents creditors with an opportunity to renegotiate. Long-term debt contracts often contain thousands of such clauses.[10] Creditors are generally better off when they negotiate individually with managers than when they participate in a collective negotiation such as a workout or a bankruptcy reorganization.

3. Research on corporate bankruptcy—theory

3.1. Effects of priority rules on the bankruptcy decision, managerial effort, and the choice between safe versus risky investments

Priority rules in bankruptcy affect the efficiency of managers' decisions both to invest in safe versus risky investment projects and to file for bankruptcy versus remain out of

[9] See Roe (1987), Gertner and Scharfstein (1991), and Schwartz (1993) for discussion and Kahan and Tuckman (1993) for a theoretical model which shows that coercive offers may succeed. Kahan and Tuckman also present empirical evidence that coercive offers do not make bondholders worse off, but their sample excludes firms in financial distress. Coercive offers are also used in renegotiation of sovereign debt. See White (2002).

[10] See Smith and Warner (1979) for discussion.

bankruptcy. If managers invest in risky projects when safe projects have higher expected returns, then the additional return from the safe project is lost, and vice versa. If managers choose to avoid bankruptcy and continue the firm's operations, but its assets are more valuable in some alternate use, then resources are wasted. Conversely when managers choose liquidation but continuation has a higher expected return, the cost is that the firm's assets are shifted to alternative uses when they would be worth more if they remained together in their current use.[11] When managers invest inefficiently or make inefficient bankruptcy decisions, creditors' return is likely to be lower and they respond by raising interest rates and/or reducing credit availability.

It should be noted that models of the economic effects of priority rules include their effects on both the size and division of the pie. When "deviations from the APR" occur, the firm's pre-bankruptcy equityholders receive a positive payoff (rather than zero) and its creditors receive less. Thus deviations from the APR imply that the size of the pie falls. When one group of creditors leapfrogs over another, the division of the pie changes. But the size of the pie may also change if the firm's investment behavior is affected.

In this section, I first discuss basic models that illustrate these points and then turn to extensions, including models with asymmetric or incomplete information.

3.1.1. Models with complete information

Turn first to models of the bankruptcy decision.[12] Suppose a firm is in financial distress and managers—representing equity—are considering whether to file for bankruptcy. Assume initially that the only bankruptcy procedure is liquidation, so that managers' bankruptcy decision is a choice between liquidating the firm in bankruptcy versus continuing to operate the firm outside of bankruptcy. Managers make economically efficient choices if they file for bankruptcy whenever the firm's assets are more valuable in alternate uses and continue to operate whenever the firm's assets are more valuable in their current use. Assume that managers and creditors are fully informed about the value of the firm's assets in both their current and alternate uses.

Suppose the firm has total debt of D, divided between D_1 due in period 1 and D_2 due in period 2, where $D = D_1 + D_2$. The firm has no cash on hand. The liquidation value of the firm's assets in period 1 is L and, since $L < D$, it is insolvent. Managers can either file for bankruptcy in period 1 or continue the firm's operations outside of bankruptcy until period 2. In order for continuation to occur, managers must obtain a new loan that allows the firm to repay D_1 in period 1. The new lender, if one exists, is referred to as the bank and it must lend an amount $B_2 = D_1$. If the firm continues

[11] Railroads are an important example of firms whose assets are worth more if they remain together. Reorganization in the U.S. began as a procedure to prevent secured creditors from seizing and selling the track of financially distressed railroads, since track is worth little if it is dispersed. See Baird (1987) and Warren (1935).

[12] See Bulow and Shoven (1978), White (1980), (1983) and (1989), and Gertner and Scharfstein (1991).

to operate, it earns P_2 with certainty in period 2, but the liquidation value of its assets falls to zero. Ignoring the time value of money, continuation in period 1 is economically efficient if $P_2 > L$ and liquidation is economically efficient otherwise. At the end of period 2, assume that the firm is liquidated and the amount P_2 is distributed according to the APR. Priority among creditors in liquidation is according to "me-first," i.e., debts are paid in chronological order based on when the loans were made.

The bank and managers—representing equity—are assumed to act as a coalition in making the bankruptcy decision in period 1, so that the bank makes the loan if continuation benefits the bank and equity taken together. If the firm liquidates in period 1, equity receives nothing since $D > L$. If the bank lends and the firm continues to operate, the coalition receives $\max[P_2 - D_2, 0]$ in period 2, so that its net return is $\max[P_2 - D_2, 0] - B_2$. (This is because the debt D_2 has priority over the bank loan.) In order for the coalition to form and continuation to occur, this expression must be positive, which implies that $P_2 > B_2 + D_2 = D$. Since $D > L$, this means that $P_2 > L$. Thus the coalition chooses continuation only when it is economically efficient. However this efficiency result is one-sided, since the coalition sometimes chooses liquidation even when continuation is more efficient. Suppose $L < P_2 < D$. Then the coalition chooses liquidation, but continuation is more efficient.

Thus the result under the APR and the "me-first" rule is that too much liquidation occurs. This is because continuation increases the value of the debt D_2, but managers and the bank ignore this gain because they do not share it. This result is an example of Myers' (1977) "debt overhang" problem, since inefficient liquidation is more likely to occur when the firm's debt is high.

Now suppose the APR continues to hold, but priority among creditors is according to "last-lender-first." Then if the bank lends, its loan takes priority over the debt D_2 in period 2. In this situation, the coalition receives the first B_2 dollars of the firm's earnings in period 2, none of the next D_2 dollars, and all of the firm's earnings above $B_2 + D_2$. The condition for the coalition to form and the firm to continue operating therefore becomes $P_2 \geq B_2$. Therefore continuation is more likely to occur when "last-lender-first" priority is used than when "me-first" priority is used. Using the insolvency condition, the condition for continuation to occur can be expressed as $P_2 \geq B_2 \geq L - D_2$, while the condition for continuation to be efficient is $P_2 \geq L$. Thus under the "last-lender-first" rule, less inefficient liquidation and more inefficient continuation occur, because continuing the firm increases the value of the coalition at the expense of the debt D_2. The additional continuation is an example of how leapfrogging by creditors may reduce economic efficiency—here the increase in the bank's priority relative to the debt D_2 increases the probability of continuation even though liquidation may be more efficient.[13]

[13] See Bebchuk and Fried (1996) for an article questioning whether secured creditors should receive priority in bankruptcy. The model discussed here, in which last-lender-first priority is substituted for me-first priority, can alternately be interpreted as an illustration of the effect of a creditor shifting from unsecured to secured status. As the discussion shows, the shift increases the probability of inefficient continuation. See also Stulz and Johnson (1985).

Now suppose the firm's period 2 earnings are uncertain rather than certain. To keep the model simple, assume that period 2 earnings under continuation are either $P_2 + G$ or $P_2 - G$, each with 0.5 probability. Also assume that $P_2 + G \geq D_2 \geq P_2 - G$. Suppose again that the "me-first" rule applies, so that the debt D_2 has priority over the bank's continuation loan. Under these assumptions, the coalition's expected return if continuation is chosen is $0.5(P_2 + G - D_2) - B_2$ (since the coalition gets nothing if the firm is unsuccessful in period 2). This implies that the coalition chooses continuation if $P_2 \geq 2B_2 + D_2 - G$, but continuation is only efficient if $P_2 \geq L$. Thus if $2B_2 + D_2 - G < P_2 < L$, then continuation occurs but liquidation is more efficient, and if $L < P_2 < 2B_2 + D_2 - G$, then liquidation occurs but continuation is more efficient. As the firm's earnings become more uncertain (G rises), inefficient continuation is more likely to occur. This is because the coalition gains when the firm's return is risky, since it keeps the additional return in the good outcome, but shares the loss with the other creditor in the bad outcome. These results illustrate the moral hazard problem pointed out by Stiglitz (1972) and Jensen and Meckling (1976) that, in the presence of debt, managers favor risky projects over safe ones, even if risky projects offer lower expected returns, because equity gains disproportionately from risky projects if they succeed. This effect applies to the firm's bankruptcy decision as well as to investment decisions more generally.[14]

Now suppose Chapter 11 reorganization is introduced into the analysis. Suppose in period 1 the coalition chooses among liquidation under Chapter 7, reorganization under Chapter 11, or continuation outside of bankruptcy. Under Chapter 11, the firm does not have to repay the debt D_1 in period 1, but it must obtain a loan of T in period 1 to cover the transactions costs of the reorganization process. Assume that at the beginning of period 2, the firm adopts a reorganization plan that requires it to repay a fraction r of the debts D_1 and D_2. These payments are made in period 2.[15] Therefore the amount that the bank must lend the firm in order for the coalition to form is T rather than D_1. Assuming that $T < D_1$, the difference $D_1 - T$ represents the improvement in the firm's immediate cash flow that occurs when it files under Chapter 11. Assume also that the bank's loan takes post-petition priority over the firm's other debts as an expense of reorganization. Finally, assume that $P_2 + G > r(D_1 + D_2) + T$ and $P_2 - G > T$. Then if the firm reorganizes, the coalition's expected return net of the cost of the loan is $0.5(P_2 + G - r(D_1 + D_2)) + 0.5T - T$. Here the coalition receives $P_2 + G - r(D_1 + D_2)$ if the firm is successful in period 2 and T if the firm is unsuccessful. The coalition therefore prefers reorganization to both liquidation and continuation outside of bankruptcy if $0.5(P_2 + G - r(D_1 + D_2) - T) > \max[0.5(P_2 + G - D_2) - B_2, 0]$. Reorganization is more likely to be preferred to liquidation as G increases and reorganization is more

[14] The bias toward too much continuation becomes stronger when the bank is also the lender that is owed D_1. In this case the bank's opportunity cost of joining the coalition falls since it does not have to provide new funds.

[15] Alternately if the two debts had different priority, they might receive different repayment rates under the reorganization plan.

likely to be preferred to both liquidation and continuation as T and r fall. Thus the introduction of reorganization as an alternative bankruptcy option makes it more likely that the firm will continue operating rather than liquidate, although it may operate in Chapter 11 rather than outside of bankruptcy. Relative to continuation, reorganization benefits the coalition by reducing the cost of the loan that the bank must provide in period 1 and by forgiving a proportion $(1 - r)$ of the firm's debt. But these benefits have little to do with whether it is economically efficient for the firm to continue operating. Since reorganization is economically efficient only when $P_2 > L$, the increase in the probability of failing firms continuing to operate is likely to be inefficient.

Now turn to the effect of priority rules on the efficiency of investment decisions that managers make ex ante, when the firm is not in financial distress. Bebchuk (2002) examines a model in which each firm has only one creditor, so that the only priority rules considered are the APR versus deviations from the APR. Bebchuk characterizes both as a proportional sharing rule under which equity gets a fraction α of the value of the firm's assets in bankruptcy. In Chapter 7 bankruptcy liquidation, there are no deviations from the APR, so that $\alpha = 0$. In Chapter 11 bankruptcy reorganization, deviations from the APR occur, so that $\alpha > 0$. Bebchuk assumes that creditors lend only if they expect to make zero profits. If the value of α changes, creditors adjust the interest rate so that expected profits remain equal to zero, i.e., they cannot be cheated by priority rule changes.[16]

Bebchuk compares the efficiency of ex ante investment incentives under the APR versus deviations from the APR. He shows, first, that at a given interest rate, equityholders are more likely to choose risky over safe investment projects when deviations from the APR occur. When there are no deviations from the APR, equityholders have an incentive to favor risky over safe projects because they receive all of the return net of interest payments when the project succeeds, but creditors bear most of the loss when the project fails. Deviations from the APR further increase the attractiveness of risky relative to safe projects, since equity's return remains the same when the project succeeds, but rises when the project fails. Second, Bebchuk shows that creditors raise the interest rate when α rises, both because equityholders are more likely to choose risky projects and because creditors gets less when failure occurs. Finally, higher interest rates further increase the likelihood that equityholders choose risky projects, since when interest rates are high, only investments that have very high upside returns allow managers to repay costly debt and still have something left over for equity if the investment succeeds. Thus introducing Chapter 11 as an alternative to Chapter 7 distorts the efficiency of investment incentives and causes equity to favor inefficiently risky projects even more strongly. The larger is α, the worse the distortion.

Bebchuk also uses his model to examine how priority rules affect the efficiency of investment incentives ex post, when firms are already in financial distress. He shows

[16] See below for empirical evidence concerning the size of α. Cornelli and Felli (1997) also model the effect of priority rules on ex ante efficiency.

that in this situation, the results are reversed and deviations from the APR reduce rather than increase equityholders' bias toward risky investment projects. This is because when the project is likely to fail and the firm to file for bankruptcy, equityholders' main return comes from their share α of the firm's value in bankruptcy. Therefore the safer the project, the more equity receives. As a result, if Chapter 11 reorganization is substituted for Chapter 7 liquidation as the bankruptcy procedure, there is an ambiguous overall effect on the efficiency of managers' investment decisions: they become less efficient ex ante but more efficient ex post.[17]

Overall, these models suggest that none of the commonly-used priority rules in bankruptcy always give managers/equityholders incentives to make efficient bankruptcy decisions or efficient investment choices. When firms are financially distressed and their future earnings are certain, the me-first and last-lender-first versions of the APR may result in either too much liquidation or too much continuation. As firms' future earnings become more uncertain, inefficient continuation is more likely to occur. When reorganization is introduced as a third bankruptcy option, the bias toward inefficient continuation becomes yet stronger. When the alternatives are no deviations from the APR versus deviations from the APR, then deviations from the APR worsen managers' bias toward choosing inefficiently risky investment projects ex ante, but have the opposite effect ex post. Although other priority rules might theoretically result in efficient bankruptcy and investment decisions, no general rule has been proposed.[18]

3.1.2. Models with asymmetric or incomplete information

Turn now to "filtering failure." Suppose there are two types of financially distressed firms: type 1 firms that are economically efficient and should reorganize versus type 2 firms that are economically inefficient and should liquidate. In the first-best bankruptcy outcome, all type 1 firms would reorganize and all type 2 firms would liquidate. "Filtering failure" occurs in bankruptcy whenever type 1 firms liquidate and/or type 2 firms reorganize. White (1994) examined an asymmetric information model of filtering failure under which managers of failing firms are assumed to know their firms' type, but creditors do not. The structure of the model incorporates features of U.S. bankruptcy law, including managers' right to choose between Chapter 7 versus Chapter 11, managers' right to offer the first reorganization plan under Chapter 11, and creditors' right to accept or reject managers' proposed plan. But the model ignores conflicts of interest among creditors.

[17] In the context of the model discussed above, equityholders receive $\alpha(P_2 - G)$ when the project fails, where failure is assumed to occur with high probability. Assuming that α is positive (Chapter 11 is in effect), equity's return rises as G falls, i.e., as the project becomes safer.

[18] See the discussion of contracting about bankruptcy below for discussion of alternate priority rules that achieve efficiency in particular models. These generally involve creditors promising to bribe managers to liquidate rather than reorganize in bankruptcy.

Managers of type 1 firms always file for bankruptcy under Chapter 11, but they choose between offering reorganization plans with high versus low payoff rates to creditors. Managers of type 2 firms choose between filing under Chapter 7 versus Chapter 11. If they file under Chapter 11, then they offer the same low-payoff reorganization plans as type 1 firms. Creditors must decide whether to accept or reject managers' reorganization plans without knowing individual firms' types. Creditors always accept high-payoff reorganization plans, but they may either accept or reject low-payoff plans. If creditors accept low-payoff plans, then the plans go into effect and the game ends. If creditors reject low-payoff plans, then they are assumed to learn individual firms' types (because the bankruptcy judge replaces managers and gives creditors more control). If the firm turns out to be type 1, then creditors receive a higher payoff than if they had accepted managers' plan; but if the firm turns out to be type 2, then it liquidates and creditors receive less than if they had accepted. Thus rejecting a low-payoff reorganization plan is a gamble for creditors. Managers of both types of firms also gamble when they offer low-payoff plans rather than choosing their alternative strategy, since they are better off if creditors accept these plans but worse off if creditors reject.

I show that either efficient filtering or filtering failure may occur in equilibrium, depending on the proportion of firms in financial distress that are type 1 versus type 2. If most distressed firms are type 1, then creditors always reject low-payoff reorganization plans since their expected return when they reject these plans is higher. Therefore all type 1 firms offer high payment reorganization plans under Chapter 11 and all type 2 firms liquidate under Chapter 7. A separating equilibrium occurs in which there is no filtering failure. But if most distressed firms are type 2, then creditors always accept low-payoff plans and, as a result, managers of both types of firms always offer them. A pooling equilibrium therefore occurs in which there is filtering failure, since all type 2 firms reorganize when they should liquidate. There also may be mixed strategy equilibria in which some type 2 firms reorganize and others liquidate. The model thus suggests that filtering failure may occur in bankruptcy and that it takes the form of too much reorganization.

Now turn to strategic default and its interaction with bankruptcy costs. Suppose firms are either solvent or insolvent, and again only managers know their firms' types. Because the bankruptcy process is costly, it is efficient for firms that are in financial distress to avoid filing for bankruptcy by negotiating non-bankruptcy workouts. Suppose managers of both types of firms choose whether to propose a workout that will reduce payments to creditors. If managers propose a workout, then creditors must either accept or reject without knowing their firms' types. Creditors have an incentive to accept workout proposals, since accepting allows the firm to avoid filing for bankruptcy. But if creditors accept all workout proposals, then managers have an incentive to default strategically by proposing workouts even when their firms are solvent. In order to discourage strategic behavior, creditors must therefore reject some or all of managers' workout proposals. But if creditors reject workouts, then at least some firms in financial distress must end up in bankruptcy. The model thus implies that, when information is

asymmetric, either some strategic default or some costly bankruptcy (or a combination of both) must occur.[19]

A similar tradeoff occurs in financial contracting models.[20] The financial contracting literature considers the optimal method of financing investment projects when entrepreneurs/managers have projects but no cash and investor have cash but no projects. Suppose an investor lends D dollars to an entrepreneur in period 0. In period 1, the project either succeeds or fails. If it succeeds, then it generates a return of $R_2 > D$ in period 2 and an additional return of $R_3 > D$ in period 3. If it fails, then it earns zero in period 2, but it still earns R_3 in period 3. Also assume that the project's assets have a positive liquidation value of L in period 2, but zero in period 3. Since $R_3 > L$, it is efficient for the project to continue until period 3 regardless of whether it succeeds or fails.

Information is assumed to be incomplete in the sense that, while all parties can observe the firm's returns each period, investors and entrepreneurs cannot make a contract based on the firm's returns because they are not verifiable in court. But they can contract for entrepreneurs to make a fixed dollar payment to investors at a particular time and for investors to have the right to liquidate the project if the entrepreneur defaults. Suppose the parties to agree that the entrepreneur will pay investors D in period 2 and that investors will otherwise have the right to liquidate the firm in period 2 and collect L. Under this contract, entrepreneurs never default strategically: they repay D in period 2 if the project succeeds and they default only if it fails. Entrepreneurs prefer to repay in period 2 whenever they can, since they gain from retaining control and collecting R_3 in period 3. The contract does not call for the entrepreneur to pay anything to investors in period 3, since no obligation to pay is enforceable when the firm's liquidation value is zero.

While the contract eliminates strategic default, it results in costly bankruptcy. This is because investors liquidate all projects that default in period 2, but liquidation is always inefficient since it results in a loss of $R_3 - L$. If instead investors allowed entrepreneurs to remain in control following default, then entrepreneurs would default even when their firms were successful. Other possible contracts, such as investors playing mixed strategies, result in less bankruptcy but more strategic default (see Bolton and Scharfstein, 1996a). But because of incomplete information, no contract can eliminate both bankruptcy and strategic default.

Several papers in the financial contracting literature consider alternative ways of reducing strategic default. Bolton and Scharfstein (1996a) extend their model to consider the optimal number of creditors and find that, when entrepreneurs borrow from multiple creditors, they are less likely to strategically default. This is because strategic

[19] Other models of default and workouts include Schwartz (1993) and Gertner and Scharfstein (1991).

[20] This discussion draws on Hart and Moore (1998). The financial contracting literature is concerned with the more general problem of determining the most efficient method of financing investment projects. Debt contracts are shown to be efficient under fairly general assumptions, since they induce entrepreneurs to pay out some of their projects' returns to investors, rather than always defaulting.

default only succeeds if none of the creditors liquidates the project and this outcome becomes less likely as the number of creditors increases. Berglof and von Thadden (1994) consider a similar model in which the project has both short-term and long-term debt. Short-term and long-term debtholders have differing stakes in the project, since the latter benefit from its future earnings, while the former do not. As a result, short-term debtholders are more likely to liquidate the project following default. Berglof and von Thadden show that entrepreneurs are less likely to default strategically if the investors who hold the project's short-term debt do not hold any of its long-term debt as well. Bester (1994) considers whether it is efficient for investors to lend on a secured rather than unsecured basis, where secured claims have the advantage that they reduce strategic default, but have the drawback of higher transactions costs. Bolton and Scharfstein (1996b) consider how debt contracts affect the competitive structure of the industry. Hart and Moore (1998) consider non-debt contracts.[21]

Another issue that is important for corporate (as well as personal) bankruptcy is how bankruptcy law affects entrepreneurs' effort levels. Povel (1999) uses a financial contracting model to analyze the tradeoff between entrepreneurs' effort levels and delay in filing for bankruptcy. Suppose entrepreneurs borrow in period 0 to invest in a project and choose their effort levels in period 1. Projects may turn out to be good, intermediate, or bad, where returns are highest for good projects, next highest for intermediate projects, and lowest for bad projects. Higher effort by entrepreneurs raises the probability that projects turn out to be good or intermediate, rather than bad. Higher effort is economically efficient, but it lowers entrepreneurs' utility. Investors are assumed unable to observe managers' effort levels. In period 2, the entrepreneur receives a signal concerning the project's type, which investors do not observe. If the signal is that the project's type is bad, then it is efficient to liquidate it immediately. If the signal is intermediate, then it is efficient for investors to rescue it by investing additional funds, where rescues convert projects with intermediate signals into projects equivalent to those that receive good signals. After receiving the signal, entrepreneurs must choose between filing for bankruptcy versus continuing to operate the firm outside of bankruptcy. Filing for bankruptcy reveals the signal to investors, while continuing outside of bankruptcy conceals it. If entrepreneurs file for bankruptcy, then investors rescue projects that have intermediate signals and liquidate projects that have bad signals. (Entrepreneurs do not file if their projects receive good signals.) In period 3, if the project is still in existence, its true type is revealed and it earns a final return. Entrepreneurs have an incentive to avoid filing for bankruptcy when their projects receive intermediate or bad signals, both because they benefit from remaining in control for longer and, since returns in period 3 are uncertain, delay may solve the firm's financial problems without investors' intervention. But delay is costly since rescues are only possible if they take place early.

[21] See also Webb (1987). An earlier literature, not discussed here, argued that amount of debt in firms' capital structures is determined by a tradeoff between the tax advantage of using additional debt rather than equity versus the increase in expected bankruptcy costs as debt increases. See, for example, Gordon and Malkiel (1981) and Bergman and Callen (1991).

Povel shows that the first best outcome is for entrepreneurs use high effort and to reveal information by filing for bankruptcy in period 2 whenever the signal is intermediate or bad. But this outcome does not occur in equilibrium. Povel analyzes the model under two different bankruptcy laws, which he refers to as "soft" versus "tough." "Tough" bankruptcy law corresponds to Chapter 7 liquidation and, under it, entrepreneurs are fired whenever they file for bankruptcy in period 2. "Soft" bankruptcy law corresponds to Chapter 11 reorganization. Under it, if entrepreneurs file for bankruptcy in period 2, they remain in control when the project has an intermediate signal and creditors rescue it, while they receive a payoff when the project has a bad signal and creditors liquidate it. Povel shows that, when bankruptcy law is soft, managers file for bankruptcy in period 2 whenever they receive intermediate or bad signals, since they are treated well. But because they have a soft landing in bankruptcy, they use less effort. In contrast when bankruptcy law is tough, managers never file for bankruptcy in period 2, since doing so costs them their jobs. But then they have an incentive to use high effort in order to increase the probability that the project's type will be good. Thus neither "soft" versus "tough" bankruptcy law results in both efficient effort levels and early bankruptcy filings. Depending on whether high managerial effort or early bankruptcy filings is more important, either type of bankruptcy law could be more economically efficient.[22]

Berkovitch, Israel, and Zender (1998) also analyze a model in which entrepreneurs make an effort-level decision that investors cannot observe and in which there is an early signal that the project's quality is good, intermediate or bad. But in their model, the signal is observed by both entrepreneurs and investors, so that there is no strategic default or delay in filing for bankruptcy. If the signal is bad, then investors liquidate the project, which is efficient. If the signal is intermediate, then the best outcome is for the project to continue operating without any additional investment. However the loan contract must be renegotiated, since the entrepreneur would abandon the project if investors had to be repaid in full. Berkovitch et al. show that entrepreneurs choose an efficient level of effort if, when the signal is intermediate, investors receive the project's liquidation value L if it liquidated immediately and the entrepreneur receives all of the project's final period earnings net of its liquidation value. This solution is efficient because it allows entrepreneurs to keep all of the marginal product of their extra effort. The efficient outcome can be implemented by either of two bankruptcy reorganization procedures: in the first, entrepreneurs and investors renegotiate their contracts and entrepreneurs are allowed to make take-it-or-leave it offers to investors; while in the second, the project is auctioned, but the original investors are not allowed to bid.[23] Then in equilibrium, entrepreneurs either make an offer of L to investors in the renegotiation and investors accept or entrepreneurs win the auction by bidding L. Thus the model suggests that in bankruptcy, either a renegotiation process (similar to the actual Chapter 11 procedure) or an auction

[22] Povel (1999) also considers which bankruptcy law the parties would prefer if they were allowed to choose when they write their contracts.

[23] The original investors are restricted from bidding because, unlike new investors, they have an incentive to bid more than L.

process (similar to several bankruptcy reform proposals discussed below) can result in efficient outcomes. But the authors do not consider whether the same result would occur if only the entrepreneur received the signal.[24]

To summarize this section, theoretical models show that bankruptcy law affects managers' incentive to use effort, to default strategically when the firm is not in financial distress, to conceal the firm's financial distress from creditors, to file for bankruptcy too early or too late, and to choose inefficiently safe or risky investment projects. The models consider both the effects on economic efficiency of changing the priority rules in bankruptcy and changing bankruptcy law in other ways—including making either Chapter 7 or Chapter 11 the only bankruptcy procedure, substituting an auction process for the current negotiation process in Chapter 11, and compensating managers for liquidating projects that turn out badly. But the models suggest that, except in special cases, no one bankruptcy procedure results in economically efficient outcomes along all the dimensions considered. In the past, it was generally thought that using the APR to divide the assets of firms in bankruptcy led to economically efficient results. However the models discussed here suggest that use of the APR does not prevent managers from behaving inefficiently by choosing excessively risky investment projects, delaying too long before filing for bankruptcy, and/or concealing information about the firm's financial distress.

In the next section, I discuss the more law-oriented literature on bankruptcy reform.

3.2. Proposed reforms of Chapter 11—auctions, options, and bankruptcy by contract

A number of authors have argued for reforms of bankruptcy law. Many of the proposed reforms are based on the assumption that using the APR to divide the assets of firms in bankruptcy is optimal and that the current Chapter 11 negotiation procedure—which usually results in deviations from the APR—is sub-optimal. The reform proposals advocate substituting various market-based methods of valuing the assets of firms in reorganization for the negotiation procedure of Chapter 11. The justification for these proposals is that use of the market would result in more accurate valuations of bankrupt firms' assets and, if valuations were more accurate, then the APR (without deviations) could be used to divide firms' assets and efficiency would increase. As an example of how inaccurate valuations lead to deviations from the APR, suppose the true value of a firm's assets is $8 million and it has $8 million in high priority claims and $4 million in low priority claims. If the firm is valued at $8 million or less, then high priority creditors receive 100% of the claims against the reorganized firm, while low priority creditors and old equityholders receive nothing. But if the firm's valuation

[24] Other issues that have been explored in the literature include how bankruptcy law affects managers' incentives to invest in firm-specific human capital (see Berkovitch, Israel, and Zender, 1997), whether it is efficient for creditors or debtors to have the right to initiate bankruptcy (see Berkovitch and Israel, 1999), and how bankruptcy law affects the efficiency of buyers' and sellers' incentives to breach contracts and to make reliance investments (see Triantis, 1993).

instead is set at an inflated level of \$14 million, then high priority creditors receive only \$8 million/\$14 million = 57% of the claims against the reorganized firm, low priority creditors receive 29%, and equityholders receive 14%. Thus accurate valuations allow the firm's value to be divided according to the APR, while inflated valuations result in deviations from the APR. Negotiations over reorganization plans in Chapter 11 frequently result in inflated valuations, because adoption of a reorganization plan by the voting procedure requires that low priority creditors and equityholders vote in favor, and they only do so if they receive some of the claims on the reorganized firm. The reform proposals also abolish the voting procedure for adoption of reorganization plans in Chapter 11. This would have the effect of separating the decision concerning how to divide the value of the firm's assets from the decision concerning how to use the firm's assets. Some of the proposals also include new ways of determining how the reorganized firm's assets would be used, while others assume that the market will decide.

But it should be noted that the theoretical models discussed above paint a more nuanced picture of the efficiency of deviations from the APR. They cast some doubt on the idea that strict application of the APR in reorganization would increase efficiency.

3.2.1. Auctions

One proposal is to auction all firms in bankruptcy. If firms in Chapter 11 are operating, then they would be auctioned as going concerns and, if they have shut down, then their assets would be auctioned piecemeal. The proceeds of the auction would be distributed to creditors and equity according to the APR. This proposal would eliminate the distinction between reorganization and liquidation in bankruptcy. Under it, the winner of the auction—rather than the firm's old managers—would make the choice between shutting down the firm versus reorganizing it. This would increase efficiency since, while managers invariably favor reorganization over liquidation, buyers have their own money at stake and have an incentive to make value-maximizing decisions. Under the auction proposal, it is likely that fewer financially distressed firms would be saved and more would liquidate, i.e., there would be less filtering failure. An advantage of the auction proposal, along with similar market-based proposals, is that the reorganization process would be much quicker, since there would be no need to negotiate reorganization plans and have them approved. [25]

Roe (1983) proposed a variant on the auction idea for firms in Chapter 11 that are large enough to have publicly-traded equity. Under his proposal, reorganized firms would have all-equity capital structures and a small fraction of the reorganized firm's shares would be sold on the market during the reorganization process. The sale price of

[25] See Baird (1986), (1987) and (1993) and Jackson (1986) for discussion. Note that all of the reform proposals discussed here would require new bankruptcy legislation to be passed. For example, under current law it is difficult to auction firms that have filed under Chapter 11, since equityholders generally receive nothing in an auction and they can stop it from occurring by registering objections with the bankruptcy court.

these shares would provide an objective basis for valuing the entire firm and this valuation would be used to divide the reorganized firm's value according to the APR. The same procedure could be used if the reorganized firm has debt in its capital structure, as long as the value of the debt is clear and the total amount of debt is low enough that the reorganized firm's shares would trade at a positive price. But Roe argues that debt should be limited in order to ensure the reorganized firm's financial viability. Roe does not specify a method for determining how the firm's assets would be used after reorganization. Presumably a buyer would eventually take control of the reorganized firm by purchasing a controlling interest in its shares.

Roe notes another problem with his procedure, which is that old equity and/or junior creditors may have an incentive to artificially bid up the price of the new shares, since a higher valuation increases their payoff. Suppose the reorganized firm has 10,000 shares, of which 1,000 are sold during reorganization for $100 each, so that the firm's total value is set at $1 million. Also suppose senior and junior debt have face values of $1.5 million and $500,000, respectively. Then junior creditors have an incentive to bid up the price of the new shares, since they receive nothing in reorganization unless the reorganized firm's value exceeds $1.5 million. Suppose they bid up the price of the new shares to $200 each. Then the reorganized firm's value would be set at $2 million and junior creditors would receive $500,000/2,000,000 = 25% of the shares. Since the firm's true value is $1 million, these shares would actually be worth $250,000. Temporarily bidding up the value of the new shares from $100 to $200 would be worthwhile to junior creditors if it cost less than this amount. Given the small number of shares sold during reorganization, manipulating the market might be relatively inexpensive and therefore worthwhile.

Other potential problems with bankruptcy auctions have also been noted. One problem is that, if few bankrupt firms are auctioned, then buyers may assume that they are lemons and respond with low bids. This problem would disappear if all firms in bankruptcy were auctioned. Another problem is that initial public offerings are expensive and risky, so that they may not be worthwhile for many firms in bankruptcy. A third problem is that bidders for a bankrupt firm are likely to be other firms in the same industry. But the financial condition of firms in particular industries tends to be positively correlated. This means that if one firm in an industry is bankrupt, then other firms in the industry are likely to be in financial difficulties as well and, therefore, their bids will be low. The result may be that the winning bidder is a firm in another industry, even though the buyer that can make the best use of the firm's assets is another firm in the same industry. Or it may mean that the best use of the firm's assets is for the old manager and creditors to remain in control, i.e., for the firm to be reorganized.[26] Finally, quick auctions of bankrupt firms may force bidders to make their bids when they are very uncertain about the firm's value. Thus while quick auctions save on bankruptcy costs, they may result in lower bids. An alternative would be to delay holding auctions while

[26] See also Baird (1993), Shleifer and Vishny (1992), and Berkovitch, Israel and Zender (1997) and (1998).

the bankruptcy trustee or an interim manager generates additional information about the bankrupt firm's true financial situation.

3.2.2. Options

Bebchuk (1988) and (2000) proposed using options rather than auctions to value the assets of firms in bankruptcy. His proposal allows creditors and equityholders to be compensated according to the APR even though the value of the reorganized firm's assets is uncertain. To illustrate, suppose a bankrupt firm has 100 senior creditors who are each owed \$1, 100 junior creditors who are each owed \$1, and 100 shares of equity. Also suppose the reorganized firm will have 100 shares of equity. Under the options approach, each junior creditor is given an option to purchase the interests of a senior creditor for \$1 and each equityholder is given an option to purchase the interests of a junior creditor for \$2. All options must be exercised at a particular date. One possibility is that neither the junior creditors nor the equityholders exercise their options, which means that shares are worth less than \$1. Then each senior creditor ends up with 1 share of the reorganized firm worth less than \$1 and junior creditors and equity receive nothing. Another possibility is that junior creditors exercise their options, but equityholders do not. This means that shares are worth between \$1 and \$2 each. Each senior creditor then ends up with \$1, each junior creditor ends up with 1 share of the reorganized firm minus \$1, for a net value of less than \$1, and equityholders receive nothing. The final possibility is that both junior creditors and equityholders exercise their options, so that shares are worth more than \$2 each. Then each senior and junior creditor ends up with \$1 and each equityholder ends up with one share of the reorganized firm minus \$2. Regardless of whether the options are exercised, the APR is always followed, since each creditor either ends up with full payment (\$1) or else ends up owning a share of the reorganized firm worth less than \$1 and lowering ranking claims receive nothing. Similarly, equityholders either pay \$2 for a share of the reorganized firm worth more than \$2 or else they receive nothing. A market for the options would operate before the exercise date, so that junior creditors and equityholders would have a choice between exercising their options if they think that doing so is worthwhile or selling their options if they are liquidity-constrained or do not think that exercising them is worthwhile. An important difference between the options proposal and other market-based proposals is that the reorganized firm ends up with debt in its capital structure, although some of the old debt is converted to equity.

In Bebchuk's proposal, there is no explicit method for determining whether the old managers will be replaced and how the reorganized firm's assets will be used. After the options are exercised, the new equityholders would elect a board of directors that would hire a manager—the same procedure as is followed by non-bankrupt firms. Aghion, Hart, and Moore (1992) extended Bebchuk's options scheme to include a vote by the new equityholders on how the reorganized firm's assets will be used. Under their proposal, the bankruptcy judge solicits bids that could involve either cash or non-cash offers for the reorganized firm's new equity or simply offers to manage the firm with the new

equityholders retaining their shares. The bids would be announced at the same time that the options are issued, so that the parties could use the information contained in the bids when they decide whether to exercise their options. After the options are exercised, the new equityholders would vote on the bids and the one receiving the most votes would be selected. Both Bebchuk (2000) and Aghion, Hart, and Moore (1992) argue that an advantage of the options process is its speed—firms would exit bankruptcy within a few months after filing.[27]

3.2.3. Contracting about bankruptcy

Bankruptcy is a mandatory procedure in the sense that, when firms become insolvent, the state-supplied bankruptcy procedure must be used to resolve creditors' claims. Debtors and creditors are not allowed to contract for any alternative dispute-resolution procedure or for any limits on debtors' right to file for bankruptcy and to choose between Chapter 7 versus Chapter 11. They also cannot contract out of use of the APR in Chapter 7. In this sense, bankruptcy differs from other aspects of commercial law, where the law provides a set of default rules, but the parties are generally allowed to contract out of the default rules by agreeing on alternative arrangements. Schwartz (1997) argued that efficiency would be enhanced if creditors and debtors could choose some of the characteristics of their bankruptcy procedure when they negotiate their debt contracts.[28] The argument that allowing parties to choose their own bankruptcy procedure could enhance efficiency makes sense in light of the models of Povel (1999) and Berkovitch, Israel, and Zender (1998), discussed above, which show that the optimal bankruptcy procedure varies depending on exogenous characteristics of the parties or the legal environment. This suggests that allowing debtors and creditors to contract over the bankruptcy procedure could potentially improve efficiency.

Schwartz first examines a model in which the bankruptcy procedure is mandatory. As under current bankruptcy law, he assumes that there are separate liquidation and reorganization procedures and debtors have the right to choose between them. Firms in financial distress are divided into two types: type 1's that have higher value if they reorganize and type 2's that have higher value if they liquidate. Schwartz assumes that debtors prefer reorganization over liquidation even when their firms are type 2, because reorganization allows them to remain in control and take perks for longer. Therefore under the mandatory bankruptcy regime, some or all type 2 firms reorganize when it would be more efficient for them to liquidate, i.e., filtering failure occurs. Filtering failure in bankruptcy reduces creditors' return, thereby raising interest rates and reducing the level of investment.

[27] However disputes over the priority of particular creditors' claims could delay the process. See also Hart et al. (1997) for a proposal that combines options and auctions. See Bebchuk (1998) for discussion of auctions versus options.

[28] See Rasmussen (1992) and Adler (1994) for a similar argument that the parties should be allowed to choose their bankruptcy procedure at the time they adopt a corporate charter.

Schwartz then examines whether filtering failure might be reduced if debtors and creditors were allowed to contract over certain aspects of bankruptcy. In the contracting regime, he assumes that separate liquidation and reorganization procedures still remain in effect and debtors still have the right to choose between them (the same as under mandatory bankruptcy). But now creditors and debtors are allowed to contract in advance for creditors to pay the debtor a pre-determined fraction of the firm's liquidation value if the debtor chooses liquidation rather than reorganization in bankruptcy. Thus while the mandatory bankruptcy regime uses the APR when liquidation occurs, debtors and creditors are allowed to contract for deviations from the APR when liquidation occurs. Schwartz shows that a bribe of this type can result in efficient bankruptcy filtering, i.e., managers of type 2 firms always choose liquidation and managers of type 1 firms always choose reorganization. This is because when managers of type 2 firms are rewarded rather than penalized for choosing liquidation, they are more likely to do so. (But the reward cannot be too high, or else managers of type 1 firms would also choose liquidation.) Schwartz also considers contracts that involve debtors and creditors agreeing to renegotiate when the firm is in financial distress and shows that these contracts can also lead to efficient bankruptcy filtering. Thus a variety of possible bankruptcy contracts leads to more efficient outcomes than the current mandatory bankruptcy regime.

Schwartz' results suggest that allowing debtors and creditors to contract about the bankruptcy process in theory could improve economic efficiency. However his model only begins to probe the issue, since it ignores important issues such as asymmetric information, strategic default, and conflicts of interest among creditors. In addition, bankruptcy contracting may harm certain types of creditors—such as tort and tax claimants and trade creditors—that do not have contracts with the firm. This is because debtors and contracting creditors have an incentive to agree on a bankruptcy process that diverts value from non-contracting creditors. This topic seems ripe for further research.[29]

3.2.4. Contracts as substitutes for bankruptcy

Adler (1993) suggested an approach to contracting about bankruptcy that involves completely abolishing bankruptcy. Under his approach, called "chameleon equity," insolvent firms would not file for bankruptcy. Instead some of their debts would be converted to equity, starting with the lowest priority claims. The new equity would replace old equity—thus preserving the APR. Enough debt would be converted to equity to restore the firm to solvency. Debt contracts would no longer give creditors the right to sue firms for repayment following default or to force defaulting firms into bankruptcy. Instead, they would contain procedures for converting debt into equity in the event of insolvency. As an example, suppose a firm's assets are worth $1,000,000, but it is insolvent because it has $1,000,000 in senior debt and $500,000 in junior debt. Then the junior debt would be converted to equity and the firm's old equity would be eliminated. These changes would restore the firm to solvency.

[29] The articles by Povel (1999) and Berkovitch, Israel, and Zender (1998) consider some of these issues.

The proposal has a number of problems. An important one is that Adler assumes complete information, so that creditors and equity always agree on the firm's value. If the parties disagreed on the firm's value or the firm's value were unknown, then it would not be clear whether the firm is insolvent and if the debt conversion procedure should go into effect. Another problem is that if information were asymmetric, then managers would have a strong incentive to default strategically, i.e., to claim insolvency even when the firm's financial condition is good, since doing so allows them to avoid repaying the firm's debt. The lack of a penalty for default would undermine credit markets and greatly reduce credit availability. In addition, there would be a high level of filtering failure, since failing firms would continue to operate as long as their revenues covered variable costs, even if their assets were more valuable in some other use.

4. Research on corporate bankruptcy—empirical work

For reasons of data availability, most empirical research on corporate bankruptcy in the U.S. focuses on large corporations that have publicly traded debt or equity. This means that the studies all have small samples, since relative few large corporations file for bankruptcy. Also large corporations generally file for bankruptcy under Chapter 11, so that the available information about corporate bankruptcy is mainly for firms in Chapter 11. When large corporations liquidate in bankruptcy under Chapter 7, it is generally after a prolonged period of operating in Chapter 11 and failing to adopt a reorganization plan. This means that we know little about what would happen if large corporations filed under Chapter 7 and liquidated without first spending time in Chapter 11. It also means that comparisons of payoff rates to creditors of large corporations under Chapter 11 versus Chapter 7 are biased upward.[30]

Empirical research has concentrated on measuring the costs of bankruptcy and the size and frequency of deviations from the APR. More recent papers also examine how out-of-bankruptcy workouts and prepacks differ from normal Chapter 11 filings. In both workouts and prepacks, negotiations over a plan to restructure debt occur outside of bankruptcy. Depending on the outcome of the negotiations, the firm may file under Chapter 11 with a reorganization plan already agreed on or a restructuring plan might go into effect without a bankruptcy filing.[31]

4.1. Bankruptcy costs

An ideal measure of the costs of bankruptcy would cover both direct and indirect costs. Direct costs include the legal and administrative costs of bankruptcy, while indirect costs include all the costs of bankruptcy-induced disruptions, including asset disappearance,

[30] For an empirical study of small firms in bankruptcy, see LoPucki (1983).
[31] See the discussion of workouts and prepacks in Section 2.3 above.

loss of key employees, and investment opportunities foregone because managers' time is spent on the bankruptcy. Most studies measure only the direct costs of bankruptcy, because bankrupt corporations must report these costs to the bankruptcy court. Weiss' (1990) study of 37 corporate bankruptcies during the early 1980's found that the direct costs of bankruptcy averaged 3.1% of the combined value of debt plus equity. Other studies by have found similar results (see Ang, Chua, and McConnell, 1982).

Indirect bankruptcy costs are more difficult to measure, but are likely to be much greater than direct bankruptcy costs. White (1983) solved for upper bound expressions on indirect bankruptcy costs, using a coalition model of the bankruptcy decision. Her results suggest that the indirect costs of bankruptcy may be as high as twenty times the direct costs of bankruptcy.

Other studies provide indirect evidence suggesting that bankruptcy is very disruptive. Gilson (1990) and Gilson and Vetsuypens (1994) found that the turnover rates of top executives and directors were much higher for large corporations in Chapter 11 than for those not in bankruptcy. Carapeto (2000) found that when a large corporation in Chapter 11 offers multiple reorganization plans to creditors, the total amount offered declines by 14% between the first and the last plan. This implies that the marginal costs of remaining in bankruptcy longer increase quickly. Hotchkiss (1995) found that filing for bankruptcy under Chapter 11 and adopting a reorganization plan does not necessarily solve the financial problems of distressed corporations, since one-third of her sample of firms that successfully reorganized required further restructuring within a few years. Her results are consistent with a model in which some inefficient firms reorganize even though they should liquidate, but are also consistent with models in which reorganized firms fail simply because they have too much debt in their capital structures.

4.2. Deviations from the absolute priority rule

A number of studies have estimated the frequency and size of deviations from the APR. Following Franks and Torous (1989), these studies classify reorganization plans as involving deviations from the APR if equity receives more than it would under the APR and they measure the size of deviations from the APR by the amount paid to equity in violation of the APR divided by the total amount distributed under the reorganization plan. For example if a firm owes $1,000,000 to creditors, then deviations from the APR occur if equity receives anything when creditors receive less than $1,000,000. Assuming that the reorganization plan calls for creditors to receive $500,000 and equity to receive shares in the reorganized firm having a value of $50,000, then deviations from the APR amount to $50,000/500,000 or 10%.[32]

[32] This ignores the fact that payments to creditors under the plan are usually made over six years, so that additional deviations from the APR occur because payments are delayed and because the reorganized firm may later default. It also ignores deviations from the APR that involve payments to lower-priority creditors when higher-priority creditors are not repaid in full.

Weiss (1990) examined a sample of 38 corporations that filed for bankruptcy. Of these, 31 adopted reorganization plans, of which 28 involved deviations from the APR. (The remaining seven corporations in his sample liquidated, including one that liquidated in Chapter 11.) Eberhart, Moore, and Roenfeldt (1990) found deviations from the APR in 23 of 30 reorganization plans they studied and Betker (1995) found deviations in 54 of 75 reorganization plans.[33] Carapeto (2000) found similar results using a more recent sample of firms in Chapter 11. Thus about three-quarters of Chapter 11 reorganization plans involve deviations from the APR. Turning to the size of deviations from the APR, Eberhart, Moore, and Roenfeldt (1990) found that the average deviation from the APR in their sample was 7.5%, with a range from 0 to 36%; while Betker (1995) found an average deviation of 2.9%.

How do deviations from the APR relate to the financial condition of corporations in Chapter 11? This relationship can be estimated by regressing the amount paid to equity as a fraction of unsecured creditors' claims on the amount paid to unsecured creditors as a fraction of their claims (i.e., the payoff rate to unsecured creditors). If the APR were always followed, the estimated coefficient of the payoff rate to unsecured creditors would be zero whenever creditors' payoff rate is less than 100%, but would become infinite whenever creditors' payoff rate exceeds 100%. Deviations from the APR are predicted to make this relationship positive even when creditors' payoff rate is low. But the coefficient of the payoff rate to unsecured creditors is predicted to rise as creditors' payoff rate approaches 100%.

White (1989) estimated this relationship, using data from the studies by LoPucki and Whitford (1990) and Eberhart, Moore, and Roenfeldt (1990). The results showed a smooth relationship with a gradually increasing slope. In particular equity receives a minimum payoff of about 5 percent of creditors' claims, regardless of how little creditors receive. When unsecured creditors' payoff rate is around 50%—a common figure—equity receives about 15% of creditors' claims and, when unsecured creditors' payoff rate reaches 90%, equity receives about 40% of creditors' claims. These results are consistent with a bargaining model of Chapter 11 such as Bebchuk and Chang (1992), in which equity gets a low payoff in return for giving up its right to delay adoption of the reorganization plan and gets more as equity's option on the firm comes closer to being in the money. Betker (1995) finds similar results. He also finds that deviations from the APR are smaller when a higher proportion of the firm's debt is secured.

Finally, several studies examine the frequency of out-of-bankruptcy workouts and compare them to Chapter 11 reorganization plans. Gilson, John, and Lang (1990) examined 169 large corporations that defaulted on their debt during the 1980s and found that 47% negotiated restructuring agreements that allowed them to avoid bankruptcy, while of the remainder, at least 70% attempted to restructure outside of bankruptcy, but failed and filed under Chapter 11. Thus about 85% of firms in their sample attempted

[33] See also LoPucki and Whitford (1990). These studies all involve samples of corporations that filed under Chapter 11 during the 1980's and there is considerable overlap.

to negotiate workouts, suggesting that workouts are the preferred procedure for corporations dealing with financial distress. However the percent of firms that succeeded in negotiating workouts outside of bankruptcy—47%—is much smaller than the percent of firms that succeeded in negotiating reorganization plans in bankruptcy—29/38 or 76% in Weiss' (1990) study. This suggests that strategic default is an important problem in workouts, i.e., creditors reject workouts because they believe that many firms are not truly in financial distress. Tashjian, Lease, and McConnell (1996) compared deviations from the APR in workouts versus Chapter 11 bankruptcies and found that workouts were associated with smaller deviations from the APR, i.e., creditors did better in workouts than in Chapter 11. This result also suggests that shareholders are in a weaker bargaining position in workout negotiations than in Chapter 11 negotiations.[34]

Part B: Personal bankruptcy

Like corporate bankruptcy procedures, personal bankruptcy procedures determine both the total amount that debtors must repay their creditors—the size of the pie—and how repayment is shared among individual creditors—the division of the pie. A larger pie benefits all individuals who borrow, because higher repayment causes creditors to lend more at lower interest rates. But a larger pie requires that debtors use more of their post-bankruptcy earnings to repay pre-bankruptcy debt, which reduces their incentive to work. A larger pie also affects whether debtors consume versus invest their wealth and whether they choose safe or risky investments. The division of the pie also has efficiency implications, because it affects whether creditors race against each other to be first to collect and how aggressively they pursue collection efforts. We discussed above how the race to be first to collect from corporate debtors has been replaced by a race to leapfrog over other creditors in the priority ordering. But in the consumer debt context, debts do not tend to be individually negotiated, so that creditors have a stronger incentive to race to be first. The race to be first can harm debtors, since they may stop working or lose their jobs if creditors repossess their cars or institute wage garnishment.

Despite these similarities, there are important differences between personal and corporate bankruptcy. One difference is that, while corporations in bankruptcy may either shut down/liquidate or continue to operate/reorganize, individual debtors in bankruptcy always reorganize. This is because an important part of individual debtors' assets is their human capital, which can only be liquidated by selling debtors into slavery. Since slavery is no longer used as a penalty for bankruptcy, all personal bankruptcy procedures are forms of reorganization.[35] Individual debtors keep their human capital and the right

[34] However Gilson, John, and Lang (1990) found somewhat contradictory results. See also Franks and Torous (1994) and Asquith, Gertner, and Scharfstein (1994).

[35] Both the U.S. and Britain also used debtors' prison in the past as a penalty for bankruptcy. But debtors' prison is inefficient as a punishment for bankruptcy because debtors cannot work (use their human capital) while in prison.

to use it and they keep some or all of their financial assets. Depending on the bankruptcy procedure, they may be obliged to use some of their wealth and/or some of their future earnings to repay debt. These features also characterize corporate reorganization under Chapter 11. Because there is no liquidation in personal bankruptcy, there is no "filtering failure," i.e., no deadweight costs occur as a result of individual debtors reorganizing in bankruptcy when they should liquidate or vice versa.[36]

Another difference between personal versus corporate bankruptcy is the insurance objective of personal bankruptcy. Individual debtors may suffer long-term harm if their consumption falls so much that they become homeless or their illnesses become disabilities for lack of medical care. Also, individual debtors' financial distress can have negative external effects on their family members, since sharp falls in consumption may cause debtors' children to drop out of school prematurely in order to work or may result in family members' illnesses going untreated. Personal bankruptcy reduces the probability of financial distress causing long-term harm to debtors or their family members by providing partial consumption insurance. It does this by discharging debt when debtors' wealth or earnings turn out to be low and they file for bankruptcy. The insurance objective of personal bankruptcy has no counterpart in corporate bankruptcy.[37]

As a result of these fundamental differences between personal and corporate bankruptcy, personal bankruptcy has exemptions that allow individual debtors to keep some of both their financial assets and their future earnings in bankruptcy, regardless of how much they owe. Higher exemptions for financial assets and future earnings benefit debtors and their family members by increasing their consumption when it would otherwise be very low. Higher exemptions for future earnings also increase efficiency by giving debtors stronger incentives to work/use their human capital after bankruptcy. But higher exemptions reduce the size of the pie, which makes borrowing less attractive to debtors. In contrast, there are no exemptions for corporations that liquidate in bankruptcy. However when corporations reorganize in bankruptcy, they keep their assets and repay creditors from their future earnings. "Deviations from the APR" are the corporate equivalent of personal bankruptcy exemptions, since they reduce the amount that debtors repay to creditors—i.e., they reduce the size of the pie.

This part of the chapter contains separate sections that discuss personal bankruptcy law, statistics on personal bankruptcy filings, theoretical research on personal bankruptcy, and empirical evidence concerning personal bankruptcy.

[36] Nonetheless, one of the two U.S. personal bankruptcy procedures is called liquidation. See the discussion below.

[37] Rea (1984) was the first to point out the insurance aspect of personal bankruptcy. Jackson (1986) argued that post-bankruptcy wages should be more fully exempt than financial wealth in personal bankruptcy, because of debtors' inability to diversify their human capital. See also Dye (1986) and Hynes (2002).

5. Legal background—personal bankruptcy law

The U.S. has two main personal bankruptcy procedures: Chapter 7—called "liquidation"—and Chapter 13—formally called "adjustment of debts of consumers with regular income."[38] I first discuss creditors' legal remedies outside of bankruptcy, then discuss Chapters 7 and 13, and finally discuss the main provisions of the recent (2005) bankruptcy reform.

5.1. Creditors' legal remedies outside of bankruptcy

When individual debtors default on their debt obligations but do not file for bankruptcy, creditors usually send letters and telephone, reminding debtors of the overdue debt and threatening to harm their credit ratings if they fail to repay. Creditors also add late charges and interest. Creditors' next step is to sue the debtor. On winning (usually by default), they can obtain a court order to garnish debtors' wages. Under the Federal Consumer Credit Protection Act, 75% of wages or 30 times the federal minimum wage per week, whichever is higher, is exempt from garnishment. A few states restrict garnishment more tightly, or ban it completely. Because the total amount that can be garnished is limited, creditors have an incentive to race to be first to garnish debtors' wages. However debtors often file for bankruptcy when their wages are garnished, since a bankruptcy filing terminates garnishment.[39]

Creditors can also seize debtors' bank accounts and/or foreclose on their houses, but they rarely do so. This is because each state has a set of exemptions for particular types of financial assets and the debtor receives up to the value of the exemption before the creditor receives anything. For example, suppose a debtor owes $10,000 on a credit card. The debtor also owns a house worth $100,000 that has a mortgage of $75,000 and the "homestead" exemption in the debtor's state covers home equity of $25,000 or more. Then foreclosing is not worthwhile for the credit card lender, since the mortgage lender receives the first $75,000 of the sale proceeds and the exemption covers the rest.

5.2. Chapter 7 "liquidation"

Although I argued above that all personal bankruptcy procedures are forms of reorganization, nonetheless one of the two U.S. personal bankruptcy procedures is called liquidation. When an individual or married couple files for bankruptcy under Chapter 7, the formal procedure is very similar to the corporate Chapter 7 bankruptcy procedure. Wage garnishment and other collection efforts by creditors terminate. Most unsecured

[38] A few individual debtors also file under Chapter 11 or Chapter 12 (intended for farmers).

[39] See White (1998a) for discussion and a state-by-state list of exemptions and limits on garnishment. The Consumer Credit Protection Act also restricts collection practices in other ways, such as limiting the hours during which creditors can call and preventing employers from firing workers the first time a creditor garnishes their wages.

debts—including credit card debt, installment loans, medical debt, unpaid rent and utility bills, tort judgments, and business debt if the debtor owns an unincorporated business—are discharged. (Other types of debt, including secured loans, student loans, child support obligations, and debts incurred by fraud, cannot be discharged in Chapter 7.) All of the debtor's future earnings and some of the debtor's financial assets are exempt from the obligation to repay—the 100% exemption for future earnings is referred to as the "fresh start." The bankruptcy court appoints a trustee to find and liquidate all of the debtor's non-exempt financial assets and the absolute priority rule (APR)—discussed above—is used to divide the proceeds among creditors. Highest priority under the APR goes to the administrative expenses of the bankruptcy process itself; followed by priority claims (mainly taxes); followed by unsecured creditors' claims. Claims in each class are paid in full until funds are exhausted.

Secured creditors—mainly mortgage creditors who have liens on debtors' houses and automobile creditors who have liens on debtors' cars—are outside the priority ordering. In Chapter 7, the debtor has a choice between continuing payments on secured loans and retaining the collateral versus defaulting and giving up the collateral. If the debtor gives up the collateral and the bankruptcy trustee sells it, then the difference between the sale proceeds and the face value of the loan becomes an unsecured debt.

Thus under Chapter 7, the size of the pie—the pool of assets that debtors must use to repay creditors—is smaller for individual debtors than for corporations. This is because individual debtors benefit from the "fresh start" and the exemptions for financial assets, while exemptions for corporations in Chapter 7 are zero. Higher exemptions reduce individual debtors' obligation to repay and increase their minimum consumption levels, since they allow debtors to keep more of their financial assets (although higher exemptions have no effect on debtors' consumption if their assets are below the exemption levels). The responsibility to set exemption levels is split between the Federal government and the states. Federal law mandates the "fresh start" in Chapter 7, so that it applies all over the U.S.[40] There is also a set of Federal bankruptcy exemptions for various types of wealth. However in 1978, Congress gave the states the right to opt out of the Federal wealth exemptions by adopting their own, so that wealth exemptions vary across states. States' wealth exemptions apply both in and outside of bankruptcy, while the Federal wealth exemptions apply only in bankruptcy. States generally have separate exemptions for equity in owner-occupied homes ("homestead" exemptions), clothing and furniture, "tools of the trade," automobiles, retirement accounts, and other assets. Homestead exemptions in particular vary widely, from zero in the Delaware to unlimited in Texas, Florida and five other states. Because debtors can easily convert non-exempt

[40] Other countries do not generally apply the fresh start in bankruptcy. For example, in Germany, individual debtors are not allowed to file for bankruptcy voluntarily and their debts are not discharged in bankruptcy, although creditors' efforts to collect are stayed. Debtors are required to repay from future earnings. See Domowitz and Alexopoulos (1998) for discussion. Note that in the U.S., not all debt is discharged in bankruptcy, so that in practice debtors receive only a partial fresh start.

assets such as bank accounts into home equity before filing for bankruptcy, high homestead exemptions protect all types of wealth for debtors who are homeowners.[41]

Debtors can file for bankruptcy under Chapter 7 no more than once every six years. This means that the right to file for bankruptcy under Chapter 7 has an option value, since filing in the future may be more valuable than filing immediately.

5.3. Chapter 13 "adjustment of debts of consumers with regular income"

Individual debtors have the right to choose between Chapter 7 versus Chapter 13 when they file for bankruptcy. Under Chapter 13, they keep all of their financial assets, but they must propose a plan to repay part of their debt from future earnings over three to five years. The debtor proposes the schedule of payments—called a repayment plan. The plan must give creditors as much as they would have received under Chapter 7, but no more. (This is called the "best interest of creditors" test.)[42] If and when the debtor completes most or all of the payments under the plan, then the remaining debt is discharged. Unlike Chapter 11 for corporations, only the bankruptcy judge must approve repayment plans; creditors do not have the right to vote on repayment plans.

The "best interest of creditors" test implies that the size of the pie must be at least as large in Chapter 13 as in Chapter 7. Also because the test applies individually to all creditors, each slice of the pie must be at least as large in Chapter 13 as in Chapter 7. But because debtors are generally obliged to repay little or nothing in Chapter 7, repayment in Chapter 13 is also low, because most debtors would prefer to file under Chapter 7 if they had to repay more in Chapter 13. As a result, debtors in Chapter 13 often propose token repayment plans in which they promise to repay only 1% of their debts, and bankruptcy judges accept these plans since debtors would otherwise shift to Chapter 7.[43]

Chapter 13 has various special features that make it attractive to debtors in particular circumstances. Some types of debts—such as those incurred by fraud–can be discharged

[41] About one-third of the states allow their debtors to choose between their states' wealth exemptions and the Federal exemptions when they file for bankruptcy. See Lin and White (2001) for a list of wealth exemptions by state.

[42] An additional requirement for discharge of debt in Chapter 7, adopted by Congress in 1984, is that the bankruptcy petition not constitute "substantial abuse" of the Bankruptcy Code. In theory this requirement could force debtors with relatively high wealth or earnings to file under Chapter 13 and to repay more than they would under Chapter 7, because they would fail the "substantial abuse" test if they filed under Chapter 7. But courts have generally held that ability to repay debt does not by itself constitute "substantial abuse" of Chapter 7. Another requirement for approving a Chapter 13 repayment plan, also adopted in 1984, is that if creditors object to the proposed repayment plan, then debtors must use all of their "projected disposable income" for three years to repay. This requirement has also been ineffective, in part because it is difficult for judges to determine what income is or should be disposable, since high-earning debtors normally have high expenses. See White (1998b) and Hynes (2002) for discussion.

[43] Note that administration of Chapter 13 varies across bankruptcy judges. Some judges require debtors to repay more than would be required in Chapter 7 and others force many debtors to file under Chapter 13 even if they would benefit more under Chapter 7. Debtors who file under Chapter 13 often fail to complete their repayment plans. See Braucher (1993) for discussion and references.

only in Chapter 13. Also debtors often file under Chapter 13 if they have fallen behind on their mortgage or car payments and wish to delay foreclosure while they make up the arrears. If the secured debt is a car loan, then filing under Chapter 13 is beneficial for debtors because the principle amount of the loan is reduced to the current market value of the car. Finally, debtors sometimes file under Chapter 13 because they have filed under Chapter 7 within the past six years and are therefore ineligible to file again. Debtors can file under Chapter 13 as frequently as every six months.

Overall, the bankruptcy exemptions and the relationship between Chapters 7 and 13 imply that there is a basic mismatch in U.S. personal bankruptcy law between individual debtors' ability to repay and their obligation to repay once they file for bankruptcy. Creditors lend to individual debtors based on their ability to repay, which increases with both financial assets and future earnings, and, outside of bankruptcy, debtors are obliged to use both assets and future earnings to repay. But once debtors file for bankruptcy under Chapter 7, their future earnings are completely exempt and some or all of their financial assets are also exempt. Even if debtors have appreciable financial wealth, they can often protect it in bankruptcy by converting it from a non-exempt form to an exempt form before filing. As a result, most individual debtors repay little in bankruptcy even when their ability to repay is high.

The Chapter 11 corporate reorganization procedure is similar to Chapter 13 in that corporate managers have the right to choose which Chapter they file under and corporate reorganization plans must only repay creditors in reorganization the amount that they would receive in liquidation. But the degree of the mismatch is greatly reduced for corporations, because corporations have no exemptions in Chapter 7 bankruptcy and no "fresh start." Corporate creditors also have the right to approve the firm's reorganization plan. As a result, corporations in Chapter 11 generally repay a much higher fraction of their debts than do individual in Chapter 13.

5.4. The new bankruptcy law

A new bankruptcy law was adopted in 2005, of which the main changes are in the area of personal bankruptcy.[44] Individual debtors must take a financial counseling course before filing for bankruptcy. Also, they must pass a series of means tests in order to file for bankruptcy under Chapter 7. If debtors' household income is greater than the median level in their state and if their disposable income over a five year period exceeds either $10,000 or 25% of their unsecured debt, then they must file for bankruptcy under Chapter 13 rather than Chapter 7. In addition, the homestead exemption is limited to $125,000 unless debtors have owned their homes for 3.3 years at the time they file for bankruptcy. Debtors' costs of filing for bankruptcy have sharply increased.

These changes are expected to reduce the number of personal bankruptcy filings by debtors who have relatively high earnings and they will also prevent millionaire debtors

[44] The new law is the Bankruptcy Abuse Prevention and Consumer Protection Act of 2005. See White (2007) for discussion.

from moving to high exemption states such as Texas and Florida to shelter their millions from creditors. The reform also seems likely to reduce the number of filings by debtors with low earnings, since many of them will be unable to afford the new high costs of filing.

6. Trends in personal bankruptcy filings

The number of personal (non-business) bankruptcy filings increased from 241,000 in 1980 to more than 1.6 million in 2003—more than six-fold. During the 6-year period from 1980 to 1985, a total of 1.8 million personal bankruptcy filings occurred; while during the 6-year period from 1998 to 2003, there were 8.6 million filings. Since the same individual cannot file for bankruptcy under Chapter 7 more often than once every six years, this means that the proportion of households that filed for bankruptcy rose from 2.2% in 1980–1985 to 8.2% in 1998–2003. One of the important issues in personal bankruptcy is to explain the large increase in the number of filings.

Because Chapter 7 is so favorable to debtors, 70% of personal bankruptcy filing occur under Chapter 7. 95% of debtors who file under Chapter 7 have no non-exempt assets and repay nothing to creditors.[45]

7. Research on personal bankruptcy—theory

7.1. Optimal personal bankruptcy policy—consumption insurance and work effort

In this section I discuss a model of optimal personal bankruptcy exemptions that takes account of both the tradeoff between loan availability and work incentives after bankruptcy and the objective of insuring debtors against very low consumption levels.[46] However the model ignores conflicts of interest among creditors by assuming that each debtor has only a single creditor and it assumes that there are no alternate forms of consumption insurance, such as unemployment compensation, welfare, or income taxes. The model also assumes that there is only one personal bankruptcy procedure that combines Chapters 7 and 13. Under it, debtors may be obliged to repay from both financial wealth *and* post-bankruptcy earnings. This differs from current U.S. bankruptcy law,

[45] See Executive Office for U.S. Trustees (2001) for data on payoff rates. For bankruptcy filing data, see Statistical Abstract of the United States, 1988, table 837, and Administrative Office of the U.S. Courts (for recent years).

[46] The objective of minimizing negative externalities that harm debtors' family members, discussed above, is assumed to be part of the insurance objective. This section draws on White (2005), Fan and White (2003), Wang and White (2000), and Adler, Polak, and Schwartz (2000). Other theoretical papers on the economic effects of personal bankruptcy law include Domowitz and Alexopoulos (1998) and Athreya (2002) (exploring the macroeconomic effects of bankruptcy law).

which allows debtors to choose between two bankruptcy procedures and exempts either financial wealth or future earnings completely. In particular, the model examines whether and when the "fresh start" policy of exempting all post-bankruptcy wages is economically efficient. The fresh start has traditionally been justified based on the argument that it causes debtors to work more after bankruptcy, since they keep all of their earnings rather than paying them to creditors. But this argument has never been carefully analyzed.[47]

Suppose in period 1, a representative individual borrows a fixed amount B at interest rate r, to be repaid in period 2. The interest rate is determined by lenders' zero profit constraint. The loan is assumed to be the individual's only loan. In period 2, wealth is uncertain. The debtor first learns her period 2 wealth, then decides whether to file for bankruptcy, and, finally, chooses her period 2 labor supply. Period 2 labor supply depends on whether the debtor files for bankruptcy.

There is a wealth exemption X in bankruptcy that combines states' exemptions for home equity and other assets. It can take any non-negative dollar value. There is also an exemption for a fixed fraction m of post-bankruptcy earnings, where $0 < m \leq 1$.[48] Bankruptcy costs are assumed to be a fixed dollar amount, S. In bankruptcy, the debt is discharged, but the debtor must use all her non-exempt wealth and earnings (up to the amount owed) to repay.

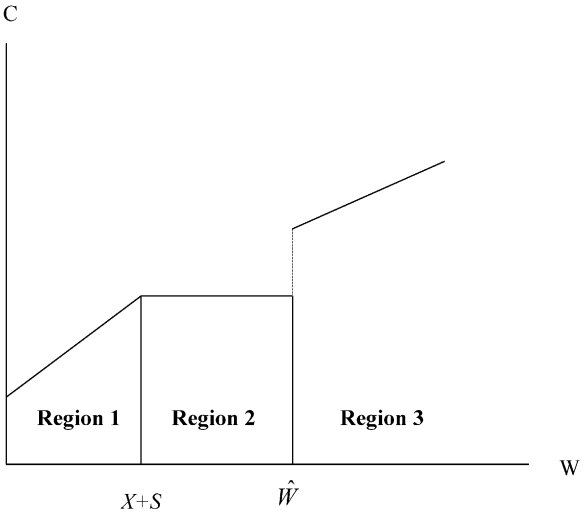
The representative individual's utility function is assumed to depend positively on consumption and negatively on labor supply in each period. Individuals are assumed to be risk averse. Period 2 work hours are denoted N_b in bankruptcy and N_n outside of bankruptcy, where N_b and N_n are both variables. When debtors file for bankruptcy, there is a negative substitution effect that causes their labor supply to fall, since debtors keep only the exempt fraction of their marginal earnings rather than 100% (assuming that $m < 1$). Filing for bankruptcy also causes a wealth effect on labor supply. If the substitution effect exceeds the wealth effect, then in the neighborhood of $\hat{W}$, $N_b < N_n$.[49]

Individual debtors decide whether to file for bankruptcy depending on which alternative maximizes their utility. (Note that debtors do not default without filing for bankruptcy—see below for discussion of the default decision.) Debtors file for bankruptcy in period 2 if their wealth turns out to be below a threshold level $\hat{W}$ and repay in full otherwise. Figure 1 shows debtors' period 2 consumption as a function of their period 2 wealth. Consumption is divided in three regions: region 3 where $W > \hat{W}$ and the

[47] The U.S. Supreme Court provided this justification for the fresh start: "from the viewpoint of the wage earner, there is little difference between not earning at all and earning wholly for a creditor." Local Loan Co. v. Hunt, 202 U.S. 234 (1934).

[48] Note that even a wealth exemption of zero provides some insurance to debtors, since their wealth cannot become negative as a result of debt repayment. The earnings exemption is assumed to be a fraction of earnings since the non-bankruptcy wage garnishment exemption takes this form. The latter covers 75% of earnings as long as weekly earnings exceed 30 times the Federal minimum wage rate. See Hynes (2002) for discussion of alternate ways of taxing debtors' post-bankruptcy earnings.

[49] See the empirical section below for evidence on the labor supply response to bankruptcy.



Note: The diagram shows period 2 consumption as a function of period 2 wealth, assuming that labor supply is fixed at N_b in bankruptcy and N_n outside of bankruptcy, where $N_n > N_b$. Debtors file for bankruptcy in regions 1 and 2 and do not file in region 3.

Figure 1. The insurance effect of bankruptcy.

debtor avoids bankruptcy and repays in full; region 2 where $X \leq W \leq \hat{W}$, the debtor files for bankruptcy and repays part of her debt from both wealth and future earnings; and region 1 where $W < X$, the debtor files for bankruptcy and repays only from future earnings. There is a discontinuous jump in consumption at $\hat{W}$ that reflects the effect of the discontinuous change in labor supply from N_b to N_n at $\hat{W}$. Assuming that labor supply falls when debtors file for bankruptcy ($N_b < N_n$), consumption must rise in order for debtors to be indifferent between filing versus not filing.

While increasing either of the two exemptions in bankruptcy provides debtors with additional consumption insurance in period 2, there are important differences between them. Raising the wealth exemption X transfers consumption from region 3 to region 2 of Figure 1, or from the highest to the middle consumption region. Consumption increases in region 2 since more of debtors' wealth is exempt; but it falls in region 3 since lenders raise interest rates. However raising the earnings exemption m transfers consumption from region 3 to regions 1 and 2 of Figure 1, or from the highest to the middle and lowest consumption regions. Consumption increases in both regions 1 and 2 since debtors keep a higher fraction of their earnings in bankruptcy. This means that the consumption insurance provided by a higher earnings exemption is more valuable at the margin than that provided by a higher wealth exemption, since only a higher earnings exemption raises consumption in the region where it is most valuable. This suggests a

new justification for the "fresh start"—that it provides particularly valuable consumption insurance.

Assume that there are many representative individuals and they all apply to borrow in period 1. Lenders' zero profit condition determines the market-clearing interest rate, r. When either of the exemption levels change, the interest rate also changes. At very high exemption levels, lenders may cease lending because no interest rate is high enough to satisfy the zero profit constraint.[50]

Because all individuals are identical in period 1, the representative individual's expected utility function is the same as the social welfare function. The optimal wealth and earnings exemption levels are therefore determined by maximizing the social welfare function with respect to m and X, subject to lenders' zero profit constraint.

The first order conditions determining the optimal wealth and earnings exemption levels have an intuitive explanation if debtors' period 2 work effort is assumed to be fixed rather than variable. In this situation, higher values of either m or X benefit debtors by providing additional consumption insurance. But debtors pay twice for the additional insurance: first in the form of higher interest rates and, second, in the form of higher expected bankruptcy costs, since debtors file for bankruptcy and pay the bankruptcy costs of S more often when exemption levels rise. Because creditors are constrained to break even, the first cost represents the fair price for the additional consumption insurance. But the second cost implies that debtors pay more than the fair price. This means that if debtors were risk neutral, they would prefer to forego consumption insurance completely and the optimal wealth and earnings exemption levels would both be zero. But if debtors are risk averse, then they prefer to buy some consumption insurance even though it costs more than the fair price. In the risk aversion case, the optimal earnings and wealth exemption levels occur where the declining marginal utility of additional consumption insurance is just offset by the marginal cost of insurance. As debtors become more risk averse, the optimal wealth and earnings exemptions rise.

Now consider how the optimal exemption levels are affected if debtors' period 2 labor supply varies in response to changes in the exemption levels. Introducing variable labor supply in bankruptcy adds two additional terms to the first order condition for the optimal earnings exemption. The first is the effect on debt repayment. Within the bankruptcy region, labor supply N_b now increases as m rises, so that debtors repay more in bankruptcy and creditors reduce interest rates. As a result, the consumption insurance provided by a higher earnings exemption becomes cheaper, debtors wish to buy more, and the optimal earnings exemption rises. The second of these terms involves the covariance of labor supply in bankruptcy with the marginal utility of consumption in bankruptcy. Since this covariance is positive,[51] variable labor supply causes period 2 consumption to become riskier, which makes consumption insurance more valuable. Variable labor supply thus causes the optimal earnings exemption to increase.

[50] See White (2005) and Longhofer (1997) for discussion.

[51] The covariance is positive because, within the bankruptcy region, higher wealth causes both labor supply and the marginal utility of consumption to fall.

Now consider how the optimal wealth exemption changes when period 2 labor supply is assumed to vary. Only one additional term is added to the first order condition for the optimal wealth exemption. Within the bankruptcy region, the larger exemption causes debtors' wealth to rise and their labor supply to fall, so that the wealth effect on labor supply is negative. Since there is no substitution effect on labor supply, the overall effect is that labor supply falls, debtors repay less in bankruptcy and creditors therefore raise interest rates. This makes the consumption insurance provided by the wealth exemption more expensive, so that debtors wish to buy less, and the optimal wealth exemption falls.

These results suggest that the first order condition for the optimal earnings exemption is likely to have a corner solution and the first order condition for the optimal wealth exemption to have an interior solution. Thus the optimal exemption policy is likely to be the "fresh start"—the 100% earnings exemption—combined with a less-than-unlimited wealth exemption.

Wang and White (2000) used simulation techniques to explore an extended version of the model in which there are two types of debtors—opportunists and non-opportunists. Non-opportunists behave as discussed above, but opportunists hide a fraction of their wealth when they file for bankruptcy. Since hiding wealth increases the gain from filing for bankruptcy, opportunists file more often than non-opportunists. (Opportunists do not hide any of their post-bankruptcy earnings in the model—perhaps because the bankruptcy trustee can check on debtors' earnings but not their wealth.) In Wang and White's model, debtors choose whether to behave opportunistically based on an individual taste for cheating. The more debtors behave opportunistically, the higher are interest rates and the worse off are non-opportunists.

Wang and White first show that when all individuals are non-opportunists, the optimal bankruptcy policy is always the fresh start combined with an intermediate wealth exemption. But when individuals are allowed to choose whether to be opportunists or not, then it is sometimes efficient to abolish the fresh start and set the earnings exemption below 100%. This is because the fresh start makes opportunistic behavior particularly attractive, since opportunists gain from hiding wealth in bankruptcy and also keep all of their post-bankruptcy earnings. But when the fresh start is abolished, opportunists' gain from hiding wealth comes at the cost of lower net earnings, since they pay the "bankruptcy tax" on earnings more often. Thus abolishing the fresh start is particularly effective in discouraging opportunism. Wang and White also find that, when the optimal bankruptcy policy is to abolish the fresh start by setting the earnings exemption below 100%, it is simultaneously efficient to raise the wealth exemption. This is because, since the two exemptions are partial substitutes in providing consumption insurance, it is efficient to offset a reduction in one exemption with an increase in the other.[52]

[52] Wang and White (2000) also found that as opportunists hide a larger fraction of their wealth when they file for bankruptcy, eventually the fresh start again becomes the optimal bankruptcy policy.

The theoretical model of bankruptcy yields several testable hypotheses. Most involve hypotheses concerning how variable wealth exemption affect debtors' and creditors' behavior, since these predictions can be tested using the variation in wealth exemptions across U.S. states. First, in jurisdictions that have higher wealth exemptions in bankruptcy, consumption is more fully insured and therefore is predicted to vary less. Second, in jurisdictions with higher wealth exemptions, interest rates are predicted to be higher and the supply of credit is predicted to be lower. Third, if debtors are risk averse, then their demand for credit will be higher in jurisdictions with higher wealth exemptions, since they prefer to borrow more when the downside risk is lower. Fourth, if potential entrepreneurs are risk averse, then jurisdictions with higher wealth exemptions are predicted to have more entrepreneurs. This is because potential entrepreneurs are more willing to take the risk of going into business if a generous bankruptcy exemption reduces the cost of business failure.

I survey the empirical literature in Section 8 below.

7.2. Additional theoretical issues

Now turn to other theoretical issues.

7.2.1. Default versus bankruptcy

In the previous section, we assumed that debtors who default on repaying their debt always file for bankruptcy. But in reality, debtors may default without filing for bankruptcy or default first and file for bankruptcy later. When debtors default but do not file for bankruptcy, creditors may garnish a fraction—usually 25%—of debtors' wages. However, pursuing garnishment is a risky strategy for creditors, because debtors may turn out to be unemployed, may quit their jobs or be fired, or may file for bankruptcy in response to garnishment.

White (1998b) used an asymmetric information model to examine whether, in equilibrium, debtors might default but not file for bankruptcy. The model has two types of debtors, type A's and type B's. Both types decide whether to default, and, following default, creditors decide whether to pursue garnishment. The two types of debtors differ in how they respond to garnishment: type A's respond by repaying in full, while type B's file for bankruptcy. Creditors are assumed unable to identify individual debtors' types when they default. I show that, in equilibrium, all type B's default, type A's play mixed strategies (they either default or repay in full) and creditors play mixed strategies (they either pursue garnishment or not). This means that in equilibrium, some debtors default and obtain the benefit of debt forgiveness without bearing the cost of filing for bankruptcy or losing wages to garnishment. The model suggests that the U.S. personal bankruptcy system encourages some debtors to default even when they could repay their debts.

7.2.2. Waiving the right to file for personal bankruptcy

In the corporate bankruptcy context, several researchers have argued that debtors should be allowed to waive their right to file for bankruptcy or to contract with creditors about bankruptcy procedures (see Schwartz, 1997, and the discussion above). But under current U.S. bankruptcy law, waivers are unenforceable and the rules of bankruptcy cannot be changed by contract. In this section I discuss whether debtors should be allowed to waive their right to file for personal bankruptcy.[53]

What does it mean for individual debtors to waive their right to file for bankruptcy? Debtors who issue waivers cannot obtain a discharge of their debts by filing for bankruptcy. However they can still default and, if so, they are protected by their states' wealth exemptions, which also apply outside of bankruptcy, and by the Federal or state limits on wage garnishment, which restrict garnishment to 25% of debtors' wages or less in a few states. Individuals who borrow and waive their right to bankruptcy make a default decision that is similar to the bankruptcy decision analyzed above. Applying the bankruptcy decision model discussed above to debtors' decision to default, debtors determine a threshold level of wealth such that they are indifferent between defaulting versus repaying in full. They default if wealth turns out to be less than this threshold.[54]

Would individual debtors ever choose to issue waivers? Formally, this amounts to a choice by debtors between facing the bankruptcy decision described in Section 7.1 versus facing a default decision with no option of filing for bankruptcy. Debtors would make this decision by comparing their ex ante expected utility in the two situations, with the expected utility expression for the bankruptcy decision evaluated at the relevant wealth and earnings exemptions in bankruptcy and for the default decision evaluated at the relevant garnishment exemptions and non-bankruptcy wealth exemptions in default. Interest rates would also differ in the two situations. Suppose creditors are allowed to garnish 25% of debtors' wages following default, while the fresh start prevails in bankruptcy. Then debtors who issued waivers would face more risk in their period 2 consumption, because their consumption in high wealth states would rise as a result of lower interest rates, but their consumption in low wealth states would fall because of wage garnishment following default. Debtors who issued waivers would probably

[53] See Rea (1984), Jackson (1986), and Adler, Polak, and Schwartz (2000) for discussion of waivers in the personal bankruptcy context. Jackson (1986) points out that not allowing waivers has the benefit of encouraging lenders to monitor to whom they lend. Rea (1984) considers the possibility of debtors agreeing to bear some pain, such as the pain of a broken arm, if they default. Adler, Polak, and Schwartz (2000) point out that giving a creditor security is equivalent to issuing a waiver for a particular debt, so that waivers are permitted if they take this form. Adler et al. also discuss reaffirmations, which involve debtors in bankruptcy agreeing to forego discharge of particular debts. These agreements are allowed because they occur after debtors file for bankruptcy.

[54] See Hynes (2004) for an argument that the system for protecting debtors outside of bankruptcy could substitute for the personal bankruptcy system. The main difference between the bankruptcy versus non-bankruptcy systems of protecting debtors is that debt is discharged only in bankruptcy. Hynes argues that debt could be discharged outside of bankruptcy by adopting short statutes of limitations for debt collection.

increase their work effort as a means of reducing risk. This suggests that debtors who are risk averse would not issue waivers. But now suppose there are both risk averse and risk neutral debtors, where the majority of debtors is risk averse and the minority is risk neutral. Then if the fresh start and a high wealth exemption in bankruptcy were adopted to accommodate the preferences of the risk averse majority, the risk neutral minority may prefer to issue waivers.

However there are a number of externality arguments that support the current policy of prohibiting waivers. One is that waivers may make individual debtors' families worse off, since spouses and children bear most of the cost of reduced consumption if the debtor has a bad draw on wealth, but debtors may not take this into account in deciding whether to issue waivers. Also, debtors may underestimate the probability of having a bad draw on wealth, so that they may issue waivers even when it is against their self-interest. Third, prohibiting waivers benefits the government itself, since its expenses for social safety net programs are lower when debtors can file for bankruptcy and avoid repaying their debts. Fourth, allowing waivers might have adverse macroeconomic effects. This is because debtors who issue waivers are more likely to repay than debtors who retain the right to file for bankruptcy. As a result, debtors who issue waivers reduce their consumption more in response to a bad draw on wealth. But if many debtors simultaneously reduce consumption, the economy could go into a recession.[55]

Finally, there is an information asymmetry argument in favor of prohibiting waivers. Suppose there are two types of debtors who differ not because they are risk averse versus risk neutral, but because they have high versus low variance of period 2 wealth. Also suppose creditors cannot observe individual debtors' types. If waivers are prohibited, then suppose a pooling equilibrium occurs in the credit market and all debtors borrow at an intermediate interest rate that reflects the average probability of default. But if waivers were permitted, then low variance debtors might prefer to issue them as a means of signaling their type. Lenders would then respond by lowering the interest rates they charge debtors who issue waivers (since they default less often) and raising the interest rates they charge debtors who do not issue waivers, i.e., the pooling equilibrium would be replaced by a separating equilibrium. In this situation, allowing waivers would be economically inefficient if the low variance debtors' gain is less than the high variance debtors' loss.[56]

7.2.3. The option value of bankruptcy

In the first section of this chapter, I discussed how the positions of corporate creditors and equityholders can be expressed as options. Similarly, the position of consumer

[55] Olson (1999) argues that the Great Depression resulted from many debtors' sharply reducing consumption in order to avoid defaulting on their debts (mainly car and furniture loans) after the stock market crash of 1929. At that time, most consumer debt was secured by the goods that the loans were used to buy. Debtors who defaulted lost the entire value of the collateral even if the remaining amount owed on the loan was small.

[56] See Aghion and Hermalin (1990) for a model in which the two types of debtors are entrepreneurs who have good versus bad projects.

debtors can be expressed as put options. If debtors' future wealth turns out to be high, then they repay their debts in full. But if debtors' future wealth turns out to be low, then they can exercise their option to "sell" the debt to creditors by filing for bankruptcy. The price of exercising the put option is the amount that debtors are obliged to repay in bankruptcy, which equals the minimum of debtors' non-exempt wealth or zero.

White (1998a) calculated the value of the option to file for bankruptcy for households in the Panel Survey of Income Dynamics (PSID), a representative sample of U.S. households. The PSID asks questions concerning respondents' wealth at five-year intervals and, for many households in the panel, there are multiple observations on wealth. This allows a household-specific variance of wealth and a household-specific value of the option to file for bankruptcy to be calculated. The results showed that the value of the option to file for bankruptcy is high for households in all portions of the wealth distribution. The high value of the bankruptcy option suggests that one reason why the personal bankruptcy filing rate has risen over time is that, as of the early 1990's, the value of the option to file for bankruptcy was positive for many more households than the number that had already filed.

7.2.4. Bankruptcy and incentives for strategic behavior

A problem with U.S. personal bankruptcy procedures is that they encourage debtors to engage in strategic behavior in order to increase their financial gain from filing for bankruptcy. Under current U.S. law, debtors' financial benefit from filing for bankruptcy under Chapter 7 can be expressed as:

$$\textit{Financial benefit} = \max\{B(1+r) - \max[W - X, 0], 0\} - S \tag{1}$$

Here $B(1+r)$ is the amount of debt discharged in bankruptcy, $\max[W - X, 0]$ is the value of non-exempt assets that debtors must give up in bankruptcy, and S indicates bankruptcy costs, including legal and filing fees, the cost of bankruptcy stigma, the cost of reduced access to credit following bankruptcy. Equation (1) assumes that the fresh start policy is in effect, so that all post-bankruptcy earnings are exempt from the obligation to repay.

White (1998a and 1998b) calculated the financial benefit of filing for bankruptcy for each household in a representative sample of U.S. households—the 1992 Survey of Consumer Finances (SCF). (I assumed that bankruptcy costs, S, were zero.) The results were that approximately one-sixth of U.S. households had positive financial benefit and would therefore benefit from filing. I also examined how the results would change if debtors pursued various strategies to increase their financial gain from bankruptcy. The strategies are: (a) debtors converting assets from non-exempt to exempt by using non-exempt assets to repay part or all of their mortgages, if the additional home equity would be exempt in bankruptcy, (b) debtors moving to more valuable houses, if doing so would allow them to shelter additional non-exempt wealth in bankruptcy, and (c) debtors charging all of their credit cards to the limit, but not obtaining new credit cards. These strategies together increased the proportion of households that benefited

from bankruptcy from one-six to one-third. A final strategy involves debtors moving to Texas before filing, since Texas has an unlimited homestead exemption and also allows debtors to use the Federal bankruptcy exemptions, which are particularly favorable to renters. Combining all of these strategies implies that 61% of all U.S. households could benefit by filing for bankruptcy. These results suggest that, even with high bankruptcy filing rates, many more households in the U.S. could benefit from filing for bankruptcy than have already filed. They also suggest that the bankruptcy filing rate rose rapidly over the decade following 1992 because consumers learned that filing for bankruptcy was financially beneficial and many of them responded by doing so.

7.2.5. Bankruptcy and the social safety net

Personal bankruptcy is not the only source of consumption-smoothing insurance. Government safety net programs, including food stamps, welfare, unemployment insurance, workers' compensation, and the earned income credit, also insure consumption. While bankruptcy provides consumption insurance by forgiving individuals' debts when their wealth or earnings are low, safety net programs provide consumption insurance by giving additional cash or in-kind transfers to individuals whose wealth and earnings are low.

Jackson (1986) and Posner (1995) both pointed out that bankruptcy reduces the cost to the government of providing a social safety net. This is because, when individuals' debts are discharged in bankruptcy, their consumption levels rise and private lenders rather than the government bear the cost. Note that cost reduction for the government may also be an explanation for why bankruptcy law does not allow debtors to waive their right to file for bankruptcy.[57]

8. Research on personal and small business bankruptcy—empirical work

Researchers interested in the empirical research on personal bankruptcy owe a vote of thanks to the U.S. Constitution and to Congress. The U.S. Constitution reserved for the Federal government the power to adopt bankruptcy laws, which means that bankruptcy law is uniform all over the U.S. But in 1978, Congress gave the states the right to set their own wealth exemption levels, so that this aspect of bankruptcy law alone varies among the states. The states have also aided the research cause by adopting widely varying exemption levels and by making relatively few changes in their exemption levels since the early 1980's. This has allowed researchers to treat exemption levels starting

[57] Private lenders in turn shift the burden of bankruptcy onto non-defaulting debtors by raising interest rates. Similarly, the costs of programs such as unemployment compensation and workers' compensation are borne by workers who are not unemployed and not injured on the job, since these programs are financed by premiums paid by employers on behalf of all workers.

in the early 1980's as exogenous to whatever bankruptcy-related decision they are investigating.

In this section, I review research on the effect of bankruptcy exemptions on a variety of behaviors, including the decision to file for bankruptcy, the labor supply decision after bankruptcy, the decision to become an entrepreneur, and the availability of consumer and small business credit. Before doing so, I briefly examine research on the political economy of personal bankruptcy.

8.1. Political economy of bankruptcy

In the 19th century, some of the Western states competed for migrants by offering protection to debtors from their—presumably Eastern—creditors. Texas particularly followed this strategy during its period of independence from 1839 to 1845, because it expected the Mexican leader Santa Ana to re-invade and needed immigrants who could help in its defense. Texas therefore adopted the first property exemption, for homesteads. Texas' pro-debtor laws attracted immigrants from nearby U.S. states and these states responded by adopting generous exemptions of their own in order to compete. While pro-debtor laws presumably attract "deadbeats," they are likely to be entrepreneurial and well-suited to the needs of a frontier economy. Even today, most of the states that have unlimited homestead exemptions form a cluster near Texas. They include, besides Texas, Arkansas, Oklahoma, Kansas, Iowa and South Dakota. In addition, Florida has an unlimited homestead exemption and Minnesota had one from the early 1980's until 1996.

Brinig and Buckley (1996) examined whether states still use bankruptcy policy to attract migrants, using data from the late 1980's. Rather than use exemption levels as their measure of bankruptcy policy, they used bankruptcy filing rates. This means they assume that states with high bankruptcy filing rates have debtor-friendly policies and vice versa. They found that states with higher bankruptcy filing rates had higher immigration rates than states with lower bankruptcy filing rates. To some extent, these results seem surprising, since states with higher bankruptcy filing rates are likely to have scarce and expensive credit. Brinig and Buckley's results suggest that immigrants in general are more concerned about fleeing their old creditors than about obtaining credit to set up new businesses. Brinig and Buckley did not test whether higher exemption levels attract more immigration.

Hynes, Malani, and Posner (2004) examine the determinants of states' bankruptcy exemption levels and test a variety of interest group explanations for exemption levels. The only variable that they found was significantly related to current exemption levels is states' exemption levels in the 1920's. Thus whatever factors determine states' exemption levels, they appear to be very persistent.[58]

[58] See Posner (1997) for discussion of political economy issues in the adoption of the Bankruptcy Code of 1978.

8.2. Studies of the bankruptcy filing decision using aggregate data

The earliest empirical work on the bankruptcy filing decision used aggregate yearly data for the U.S. to show that the passage of the 1978 Bankruptcy Code (the current U.S. bankruptcy law) caused the number of bankruptcy filings to increase. See Shepard (1984), Boyes and Faith (1986), Peterson and Aoki (1984), and Domowitz and Eovaldi (1993). A weakness of these studies is that they could only examine the overall effect of the new Code's adoption on the bankruptcy filing rate. Because the 1978 Code made many changes in bankruptcy law, these studies capture the overall impact of the changes on the bankruptcy filing rate, but cannot isolate which particular features of the Code caused the filing rate to rise. Buckley (1994) used aggregate data for the U.S. and Canada to show that the bankruptcy filing rate in the U.S. is consistently higher. He attributes this result to the fresh start policy in the U.S., which gives U.S. debtors a wider discharge from debt than Canadian debtors receive.

The theoretical model discussed above predicts that consumers are more likely to file for bankruptcy when their financial benefit is higher (see Equation (1) above). Since financial benefit is positively related to the wealth exemption, this implies that filings will be higher in states with higher wealth exemptions. Aggregate data at the national level does not allow this prediction to be tested, but aggregate data at the state or sub-state level does. White (1987) used aggregate county-level data from the early 1980's to test this relationship and found a positive and significant relationship between exemption levels and the bankruptcy filing rate. Buckley and Brinig (1998) did the same type of study using aggregate data for a panel of states during the 1980's, but did not find a significant relationship. The Buckley-Brinig results for exemption levels are not surprising, since they included state dummy variables in their model. In their specification, the state dummies capture the effect of states' initial exemption levels, while the exemption variables themselves capture only the effect of changes in exemptions. The exemption variables were probably found to be insignificant because few states changed their exemptions during the period covered by the study.

8.3. Studies of the bankruptcy filing decision using household-level data

Efforts to estimate models of the bankruptcy filing decision using household-level data were initially hampered by the fact that none of the standard household surveys used by economists asked respondents whether they had ever filed for bankruptcy. In an innovative study, Domowitz and Sartain (1999) used choice-based sampling to get around this limitation by combining two data sources: a sample of households that filed for bankruptcy in the early 1980's and a representative sample of U.S. households—the 1983 Survey of Consumer Finances (SCF)—that included information on households' income and wealth. They found that households were more likely to file for bankruptcy if they had greater medical and credit card debt and less likely to file if they owned a home.[59]

[59] Domowitz and Sartain also estimated a model of debtors' choice between Chapters 7 versus 13.

In 1996, the Panel Survey of Income Dynamics (PSID) ran a special survey that asked households whether they filed for bankruptcy during the previous decade and, if so, in what year. Because the PSID is a panel dataset that surveys the same households every year and collects data on income and wealth, this data allowed a model of the bankruptcy filing decision to be estimated using a single dataset.

The economic model of bankruptcy discussed in the previous section implies that consumers are more likely to file for bankruptcy when their financial benefit from doing so is higher. Specifically, Equation (1) predicts that only wealth, the bankruptcy exemption, the amount owed, and bankruptcy costs affect debtors' filing decisions, since these are the only variables that affect the financial benefit from filing. The economic model also predicts that income will not affect the bankruptcy decision, because it does not enter Equation (1). An alternative, sociologically-oriented model of the bankruptcy filing decision was proposed by Sullivan, Warren, and Westbrook (1989). It argues that debtors never plan for the possibility of bankruptcy nor act strategically to take advantage of it. Instead, they file for bankruptcy only when an unanticipated event occurs that reduces their earnings or increases their expenses to the point where it is impossible for them to repay their debts. In this view, the important factors affecting the bankruptcy decision are ability to repay, as measured by income, and whether adverse events have occurred that reduce ability to repay, such as job loss, illness or divorce.

The PSID data allows the two models of the bankruptcy decision to be tested against each other, since the economic model predicts that wealth rather than income determines whether debtors file for bankruptcy, while the sociological model predicts that income is the most important determinant. But in practice the test of the two models is somewhat imprecise. This is because the PSID asks questions about respondents' non-housing wealth only at five-year intervals. As a result, wealth is unknown in most years and changes in wealth over time tend to be highly correlated with household income.

Fay, Hurst, and White (FHW) (2002) used the PSID to test the two models of households' bankruptcy decisions. Their dataset consisted of PSID households in 1984 to 1995, the years covered by the PSID's 1996 bankruptcy survey. The main explanatory variable was households' financial benefit from filing in each year, calculated according to Equation (1). Other explanatory variables included household income and whether the respondent was divorced or experienced other adverse events during the previous year.

FHW found that consumers are significantly more likely to file for bankruptcy when their financial benefit from filing is higher: if financial benefit increased by $1,000 for all households, then the model predicts that the bankruptcy filing rate in the following year will rise by 7 percent. Thus the empirical evidence supports the economic model of the bankruptcy filing decision. But FHW also found that ability to repay affects the bankruptcy decision, since households with higher incomes are significantly less likely to file. They also tested whether adverse events affect the bankruptcy decision and found that neither job loss nor illness of the household head or spouse in the previous year was significantly related to whether households filed for bankruptcy. But a divorce in the previous year was found to be positively related to the probability of filing and the result

was marginally statistically significant. Thus the results support the economic model of bankruptcy. The results concerning income also support the sociological model of bankruptcy, but they do not support the hypothesis that bankruptcy filings are triggered by adverse events.[60]

FHW also investigated why bankruptcy filings have been rising over time. An additional factor that affects households' filing decision is the level of social disapproval of bankruptcy, or bankruptcy stigma. Surveys of bankruptcy filers suggest that they usually learn about bankruptcy from friends, relatives, or co-workers, who tell them that the bankruptcy process is quick and easy. This information both reduces debtors' apprehension about filing and also passively sends the message that the level of bankruptcy stigma is low, since friends and relatives have filed and are willing to talk openly about their experiences. FHW assumed that the level of bankruptcy stigma in a household's region was inversely proxied by the aggregate bankruptcy filing rate in the region during the previous year, i.e., the higher the aggregate filing rate in the previous year, the lower the level of stigma. They tested this variable in their bankruptcy filing model and found that, in regions with higher aggregate filing rates (lower bankruptcy stigma), the probability of households filing for bankruptcy was significantly higher. This suggests that as households in a region learn about bankruptcy, the filing rate rises.

Another recent study also examined the role of stigma in debtors' bankruptcy decision. Gross and Souleles (2002) used a dataset of credit card accounts from 1995 to 1997 to estimate a model of individual debtors' decisions to default and to file for bankruptcy. Their explanatory variables included measures of each cardholder's riskiness and the length of time since the account was opened. Their measure of bankruptcy stigma was the residual. They found that over the two year period from 1995 to 1997, the probability that debtors filed for bankruptcy rose by 1 percentage point and the probability that debtors defaulted rose by 3 percentage points, holding everything else constant. The authors interpret their results as evidence that the level of bankruptcy stigma fell during their time period.

Ausubel and Dawsey (2004) used credit card data to estimate a model of individual debtors' decisions both to default—which they refer to as "informal bankruptcy"—and to file for bankruptcy. In their model, debtors first decide whether to default and then, conditional on default, they decide whether to file for bankruptcy. Ausubel and Dawsey find that homestead exemptions mainly affect the decision to default; while garnishment restrictions mainly affect the decision to file for bankruptcy conditional on default. These results are not surprising, since homestead and other exemptions apply regardless of whether debtors file for bankruptcy or not, while garnishment restrictions apply

[60] Fisher (2003) re-estimated FHW's model of the bankruptcy decision, adding as an additional explanatory variable individuals' income from government safety net programs. He found that increases in both earned income and income from safety net programs reduce individuals' probability of filing for bankruptcy—a result that supports the Jackson/Posner hypothesis that bankruptcy and government safety net programs are substitutes.

only in bankruptcy. Ausubel and Dawsey argue that researchers have overlooked the importance of informal bankruptcy and the effect of garnishment restrictions on whether households file for bankruptcy, while overemphasizing the importance of exemptions. But their empirical results provide additional support for the economic model of the bankruptcy/default decision. See also Agarwal, Diu, and Mielnicki (2003).

8.4. Empirical research on work effort and the "fresh start"

As discussed above, the Supreme Court justified the "fresh start" in bankruptcy (the 100% exemption for post-bankruptcy earnings) on the grounds that debtors work more after filing for bankruptcy, because they keep all rather than part of their earnings after filing. The Justices did not state precisely what model they had in mind. One possibility is a model in which debtors have already defaulted and are subject to wage garnishment outside of bankruptcy. Then because the fresh start applies in bankruptcy, filing allows debtors to keep all of their earnings at the margin, so that the substitution effect of filing leads to an increase in labor supply. However in this model, filing for bankruptcy also increases debtors' wealth effect by discharging their debt, so that there is an offsetting negative wealth effect on labor supply. Thus the predicted effect of filing for bankruptcy on labor supply is actually ambiguous rather than positive. Alternately, suppose debtors have not defaulted but are considering whether to simultaneously default and file for bankruptcy (the model discussed in Section 7.1). Also suppose the fresh start applies in bankruptcy. Then there is no substitution effect of filing for bankruptcy because debtors keep all of their earnings at the margin regardless of whether they file or not. But filing has a positive effect on debtors' wealth that leads to a reduction in their labor supply. Thus the predicted effect of filing for bankruptcy on labor supply depends on the specifics of the model and could be either ambiguous or negative, rather than positive.

Han and Li (2004) used the special bankruptcy survey and other data from the PSID to test whether debtors' labor supply increases when they file for bankruptcy. Their results are only marginally significant, but they found that filing for bankruptcy is not associated with an increase in labor supply—in other words labor supply either falls or remains constant when debtors file. Han and Li's results suggest that the traditional justification for the fresh start does not hold.

8.5. Bankruptcy and the decision to become an entrepreneur

The U.S. personal bankruptcy system functions as a bankruptcy system for entrepreneurs well as for individuals generally. About one in five personal bankruptcy filings in the U.S. list some business debt, suggesting the importance of bankruptcy to small business owners (Sullivan, Warren, and Westbrook, 1989).

Starting or owning an unincorporated business involves incurring business debts for which the firm's owners are personally liable. This means that the variance of entrepreneurs' wealth is high, because it includes the risk associated with their businesses failing or succeeding. The personal bankruptcy system provides partial insurance for this risk

since, if their businesses fail, entrepreneurs can file for personal bankruptcy under Chapter 7 and both their business and personal debts will be discharged. As a result, personal bankruptcy law makes it more attractive for risk-averse individuals to become entrepreneurs by partially insuring their consumption. Further, states that have higher exemption levels provide more insurance because they allow entrepreneurs to keep additional financial assets—perhaps including their homes—when their businesses fail. This means that risk-averse individuals are predicted to be more likely to own or start businesses if they live in states with higher exemption levels.

Fan and White (2003) examined whether households that live in states with higher exemptions are more likely to start or own businesses, using household panel data from the Survey of Income and Program Participation. They focused on the effect of the homestead exemption, since it is the largest and most variable of the bankruptcy exemptions. They estimated separate models of whether homeowners versus renters own businesses, since only homeowners can use the homestead exemption. They found that homeowners are 35% more likely to own businesses if they live in states with high or unlimited homestead exemptions rather than in states with low homestead exemptions, and the difference was statistically significant. They also found a similarly large and significant effect for renters, which suggests that most renters who own businesses expect to become homeowners. Fan and White also found that homeowners are 28% more likely to start businesses if they live in states with unlimited rather than low homestead exemptions, although the relationship was only marginally statistically significant.

8.6. Bankruptcy and credit markets

The model discussed above suggests that bankruptcy exemptions affect the supply and demand for credit. Creditors are predicted to respond to an increase in wealth exemption levels by raising interest rates, reducing the supply of credit, and tightening credit rationing. But individual debtors—assuming they are risk averse—respond to an increase in the exemption level by demanding more credit, because the additional consumption insurance reduces the risk of borrowing. Debtors raise their credit demand because they benefit from having additional consumption insurance even though borrowing becomes more costly. (However the increase in demand may be reversed at high exemption levels, since even risk averse debtors have declining marginal utility from additional insurance.)

8.6.1. General credit

Gropp, Scholz, and White (1997) were the first to examine the effect of variable wealth exemptions on consumer credit. They used household data from the 1983 Survey of Consumer Finances (SCF), which gives detailed information on debts and assets for a representative sample of U.S. households and also asks respondents whether they have been turned down for credit. The GSW study did not distinguish between different types of credit or different types of exemptions, so that their credit variable was the sum of all

types of loans and their exemption variable was the sum of each state's homestead and personal property exemptions.

GSW found that households were 5.5 percentage points more likely to be turned down for credit if they lived in a state with exemptions in the highest rather than the lowest quartile of the exemption distribution. They also found that interest rates were higher in states with higher bankruptcy exemptions, but the effect depended strongly on borrowers' wealth. In particular, households in the second-to-lowest quartile of the wealth distribution paid interest rates that were 2.3 percentage points higher if they lived in high rather than low exemption states, but households in the third and highest quartiles of the wealth distribution paid the same interest rates regardless of the exemption level.

The authors also examined how the amount of debt held by households varies between high versus low exemption states. Although supply and demand for credit cannot be separately identified, a finding that households hold more debt in high-exemption than low-exemption states suggests that the increase in demand for credit more than offsets the reduction in the supply of credit, and conversely. The authors found that in high exemption states, high-asset households held more debt and low-asset households held less. Thus when high-asset households increase their credit demand in response to higher exemption levels, lenders accommodate them by lending more. But when low-asset households increase their credit demand, lenders respond with tighter credit rationing. GSW calculated that, holding everything else constant, a household whose assets placed it in the highest quartile of the asset distribution would hold $36,000 more debt if it resided in a state with combined bankruptcy exemptions of $50,000 rather than $6,000; while a household whose assets placed it in the second-to-lowest quartile of the distribution would hold $18,000 less debt. Thus higher exemption levels were associated with a large redistribution of credit from low-asset to high-asset households.

The results of the study suggest that, while policy-makers often think that high bankruptcy exemptions help the poor, in fact they cause lenders to redistribute credit from low-asset to high-asset households and raise the interest rates they charge low-asset households.

8.6.2. Secured versus unsecured credit

More recent studies of the effect of bankruptcy on credit markets distinguish between secured versus unsecured loans and between different types of exemptions. Secured credit differs from unsecured credit in that, if the debtor defaults, the lender has the right to foreclose on/repossess a particular asset such as the debtor's house or car. The proceeds of selling the house/car go first to repay the secured debt and then the debtor receives up to the amount of the homestead exemption or the exemption for equity in cars, whichever is relevant. Because the secured creditor must be repaid in full before the debtor benefits from the exemption, the terms of secured loans—unlike unsecured loans—are predicted to be unrelated to wealth exemptions.

However in practice, several factors muddy this prediction. First, when debtors default on secured loans, they often file for bankruptcy under Chapter 13 in order to delay

foreclosure or to reduce the principle amount of the loan (for auto loans). Thus bankruptcy filings by debtors increase creditors' collection costs. Since filing for bankruptcy is more attractive in high-exemption states, secured lending is less attractive in these states. Second, secured loans are often partly unsecured, because the market value of the collateral is less than the amount owed. When sale of the collateral brings in too little to repay the debt in full, the secured lender has an unsecured claim for the unpaid portion of the loan and the value of this claim is negatively related to exemption levels. These factors suggest that the market for secured loans may also be affected by exemption levels.

Berkowitz and Hynes (1999) examined whether higher exemptions were related to individuals' probability of being turned down for mortgages, using the Home Mortgage Disclosure Act data. They found that the probability of being turned down for a mortgage was unrelated to exemption levels. Lin and White (2001) examined the effect of higher exemptions on individuals' probability of being turned down for both mortgage and home improvement loans. Home improvement loans make a useful comparison to mortgages, since they are often unsecured or partially secured. Individuals' probability of being turned down for home improvement loans is therefore predicted to be more strongly related to exemption levels than their probability of being turned down for mortgage loans. Lin and White's study used state dummies to control for differences in exemption levels across states in the initial year and year dummies to control for time trends, so that their exemption variables capture only the effect of changes in exemption levels. They found that applicants for both mortgage and home improvement loans were more likely to be turned down in states with higher homestead exemptions. But the effect of exemptions on debtors' probability of being turned down for home improvement loans was both larger and more statistically significant than their probability of being turned down for mortgages. Finally a recent paper by Chomsisengphet and Elul (2005) argues that exemptions have been found to be a significant determinant of whether applicants were turned down for mortgages only because previous researchers did not control adequately for individual applicants' credit quality, which they argue is correlated with exemption levels. But this argument is difficult to evaluate since the HMDA data includes only very limited information about individual applicants. Overall, the question of whether exemption levels affect markets for secured credit remains unresolved.

8.6.3. Small business credit

Since debts of non-corporate businesses are personal liabilities of business owners, the terms of these loans are predicted to be affected by the exemption levels in the debtor's state of residence. In contrast, debts of incorporated businesses are not liabilities of their owners, so that the terms of loans to small corporations are predicted to be unrelated to exemption levels. But in practice, this distinction is not so clear. Creditors who lend to small corporations often require that the owners of the corporation personally guarantee the loan or give lenders second mortgages on their homes. This abolishes the

corporate/non-corporate distinction for the particular loan and suggests that personal bankruptcy law applies to small corporate credit markets as well.

The model discussed above suggests that, in states with high rather than low exemptions, demand for small business credit will be higher and supply of small business credit will be lower. Although it is impossible to separately identify the effects of exemptions on credit supply versus demand, a finding that the amount of credit held by small businesses is lower in high exemption states would suggest that the reduction in supply more than offsets the increase in demand. Berkowitz and White (2004) used data from the National Survey of Small Business Finance to examine how variations in exemption levels affect whether small business owners are turned down for credit and the size and interest rates on loans they receive. They found that for non-corporate and corporate small businesses, the probabilities of being turned down for credit rise by 32% and 30%, respectively, if firms are located in states with unlimited rather than low homestead exemptions. Both relationships are statistically significant. Conditional on receiving a loan, non-corporate businesses paid interest rates that were 2 percentage points higher and corporate firms paid interest rates that were 0.83 percentage points higher if they were located in states with high rather than low homestead exemptions. Both types of firms also received less credit if they were located in states with high rather than low exemptions.

8.7. Macroeconomic effects of bankruptcy

8.7.1. Bankruptcy and consumption insurance

The model discussed above emphasized the insurance role of bankruptcy and the fact that higher exemption levels provide additional consumption insurance. The model predicts that the variance of household consumption in a state-year will be smaller if the state has a higher exemption level. Grant (2005) tested this hypothesis using data from the Consumer Expenditure Survey, a panel survey of U.S. households. For each state-year in his sample, he computed the average variance of household consumption. Then he regressed the change in the average variance of consumption on the state's exemption level, control variables, and state fixed effects. Because the data cover a 20 year period, there are a large number of changes in exemption levels. Grant found that higher exemption levels are associated with lower variance of consumption, i.e., additional consumption insurance.

8.7.2. Bankruptcy and portfolio reallocation

Because unsecured debts are discharged when individual debtors file for bankruptcy under Chapter 7 but some assets are exempt, debtors who contemplate filing for bankruptcy have an incentive to borrow—even at high interest rates—in order to acquire assets that are exempt in bankruptcy. This behavior is referred to as "borrowing to save." The higher the bankruptcy exemption level in the debtor's state, the stronger is debtors'

incentive to borrow to save. (Similar types of strategic behavior were discussed above in connection with the proportion of households that would benefit from filing for bankruptcy.)

Lehnert and Maki (2002) examined whether households are more likely to borrow to save if they live in states with higher bankruptcy exemptions. Their definition of borrowing to save is that a household simultaneously holds unsecured debt that would be discharged in bankruptcy and liquid assets that exceed 3% of gross income. The authors tested their model using household-level panel data from the Consumer Expenditure Survey. They found that homeowners were 1 to 4 percent more likely to borrow to save if they lived in states with bankruptcy exemptions that were above the lowest quartile of the exemption distribution. The same relationship was not statistically significant for renters, which is not surprising since exemptions for renters are smaller and less variable.

Overall, the results of the empirical studies suggest that bankruptcy has important and wide-ranging effects on individual behavior. Generous bankruptcy exemptions increase demand for credit by reducing the downside risk of borrowing, but reduce the supply of credit by increasing the probability of default. In states with higher bankruptcy exemptions, individuals are turned down for credit more often and pay higher interest rates. In these states, high asset-households hold more credit, while low asset- households hold less credit—suggesting that high exemptions redistribute credit from low-asset to high-asset households. Small businesses are also affected by personal bankruptcy law. They are more likely to be turned down for credit, pay higher interest rates, and borrow less if they are located in high exemption states. In addition to their effects on credit markets, high bankruptcy exemptions also cause individual debtors to file for bankruptcy more often, become entrepreneurs more often, and reallocate their portfolios toward unsecured debt and liquid assets. Contrary to the presumption of the "fresh start," evidence suggests that individual debtors do not change their work hours significantly when they file for bankruptcy. But higher bankruptcy exemptions benefit risk-averse individuals by reducing risk, since they provide partial consumption insurance.

The empirical work on bankruptcy suggests that the increase in the number of personal bankruptcy filings that occurred over the past 20 years could have been due to a combination of households gradually learning how favorable Chapter 7 is and bankruptcy becoming less stigmatized as filing became more common. How the bankruptcy reforms adopted by Congress in 2005 will affect the number of filings remains a subject for future research.

References

Adler, B.E. (1993). "Financial and political theories of American corporate bankruptcy". Stanford Law Rev. 45, 311–346.

Adler, B.E. (1994). "Finance's theoretical divide and the proper role of insolvency rules". S. Cal. Law Rev. 67, 1107–1150.

Adler, B.E., Polak, B., Schwartz, A. (2000). "Regulating consumer bankruptcy: a theoretical inquiry". J. of Legal Studies 29, 585–613.
Agarwal, S., Diu, C., Mielnicki, L. (2003). "Exemption laws and consumer delinquency and bankruptcy behavior: an empirical analysis of credit card data". Quarterly Rev. of Econ. and Finance 43, 273–289.
Aghion, P., Hart, O., Moore, J. (1992). "The economics of bankruptcy reform". J. of Law, Econ. and Org. 8, 523–546.
Aghion, P., Hermalin, B. (1990). "Legal restrictions on private contracts can enhance efficiency". J. of Law, Econ. and Org. 6, 381–409.
Aivazian, V.A., Callen, J.L. (1983). "Reorganization in bankruptcy and the issue of strategic risk". Journal of Banking & Finance 7, 119–133.
Ang, J., Chua, J., McConnell, J. (1982). "The administrative costs of corporate bankruptcy: a note". J. of Finance 37, 219–226.
Athreya, K.B. (2002). "Welfare implications of the Bankruptcy Reform Act of 1999". J. of Monetary Econ. 49, 1567–1595.
Asquith, P., Gertner, R., Scharfstein, D. (1994). "Anatomy of financial distress: an examination of junk bond issuers". Quarterly J. of Econ. 109, 625–658.
Ausubel, L.M., Dawsey, A.E. (2004). "Informal bankruptcy". Working Paper, Department of Economics, University of Maryland, April 2004.
Baird, D.G. (1987). "A world without bankruptcy". Law and Contemporary Problems 50, 173–193.
Baird, D.G. (1993). "Revisiting auctions in Chapter 11". J. of Law & Econ. 36, 633–653.
Baird, D.G. (1986). "The uneasy case for corporate reorganizations". J. of Legal Studies 15, 127–147.
Baird, D.G., Picker, R.C. (1991). "A simple noncooperative bargaining model of corporate reorganization". J. of Legal Studies 20, 311–349.
Bebchuk, L.A. (1988). "A new method for corporate reorganization". Harvard Law Rev 101, 775–804.
Bebchuk, L.A. (1998). "Chapter 11". In: Palgrave Dictionary of Economics and the Law. Macmillan, London, pp. 219–224.
Bebchuk, L.A. (2000). "Using options to divide value in corporate bankruptcy". European Economic Review 44, 829–843.
Bebchuk, L.A. (2002). "The ex ante costs of violating absolute priority in bankruptcy". J. of Finance 57, 445–460.
Bebchuk, L.A., Chang, H. (1992). "Bargaining and the division of value in corporate reorganization". J. of Law, Econ. and Org. 8, 523–546.
Bebchuk, L.A., Fried, J.M. (1996). "The uneasy case for the priority of secured claims in bankruptcy". Yale Law J. 105, 857–934.
Berkovitch, E., Israel, R. (1999). "Optimal bankruptcy law across different economic systems". Rev. of Financial Studies 12, 347–377.
Berkovitch, E., Israel, R., Zender, J.F. (1997). "Optimal bankruptcy law and firm-specific investments". European Econ. Rev. 41, 487–497.
Berkovitch, E., Israel, R., Zender, J.F. (1998). "The design of bankruptcy law: a case for management bias in bankruptcy reorganizations". J. of Financial and Quantitative Analysis 33, 441–467.
Berkowitz, J., Hynes, R. (1999). "Bankruptcy exemptions and the market for mortgage loans". J. of Law and Econ. 42, 908–930.
Berkowitz, J., White, M.J. (2004). "Bankruptcy and small firms' access to credit". RAND J. of Econ. 35, 69–84.
Berglof, E., von Thadden, E.-L. (1994). "Short-term versus long-term interests: capital structure with multiple investors". Quarterly J. of Econ. 109, 1055–1084.
Bergman, Y.Z., Callen, J.L. (1991). "Opportunistic underinvestment in debt renegotiation and capital structure". J. of Fin. Econ. 29, 137–171.
Bester, H. (1994). "The role of collateral in a model of debt renegotiation". J. of Money, Credit, and Banking 26, 72–85.
Betker, B.L. (1995). "Management's incentives, equity's bargaining power, and deviations from absolute priority in Chapter 11 bankruptcies". J. of Business 62, 161–183.

Bolton, P., Scharfstein, D. (1996a). "Optimal debt structure and the number of creditors". J. of Political Economy 104, 1–25.
Bolton, P., Scharfstein, D. (1996b). "A theory of predation based on agency problems in financial contracting". American Economic Rev. 80, 93–106.
Boyes, W.J., Faith, R.L. (1986). "Some effects of the Bankruptcy Reform Act of 1978". J. of Law and Econ. 19, 139–149.
Braucher, J. (1993). "Lawyers and consumer bankruptcy: one code, many cultures". American Bankruptcy Law J. 67, 501–683.
Brinig, M.F., Buckley, F.H. (1996). "The market for deadbeats". J. of Legal Studies 25, 201–232.
Brown, D.T. (1989). "Claimholder incentive conflicts in reorganization: the role of bankruptcy law". Rev. of Financial Studies 2, 109–123.
Buckley, F.H. (1994). "The American fresh start". Southern Cal. Interdisciplinary Law J. 4, 67–97.
Buckley, F.H., Brinig, M.F. (1998). "The bankruptcy puzzle". J. of Legal Studies 27, 187–208.
Bulow, J., Shoven, J. (1978). "The bankruptcy decision". Bell J. of Econ. 9, 437–456.
Carapeto, M. (2000). "Is bargaining in Chapter 11 costly?" EFA 0215; EFMA 2000 Athens Meetings. (http://ssrn.com/abstract=241569.)
Chomsisengphet, S., Elul, R. (2005). "Bankruptcy exemptions, credit history, and the mortgage market". Federal Reserve Bank of Philadelphia Working Paper No. 04-14. (Available at www.ssrn.com.)
Cornelli, F., Felli, L. (1997). "Ex-ante efficiency of bankruptcy procedures". European Ec. Rev. 41, 475–485.
Domowitz, I., Alexopoulos, M. (1998). "Personal liabilities and bankruptcy reform: an international perspective". International Finance 1, 127–159.
Domowitz, I., Eovaldi, T. (1993). "The impact of the Bankruptcy Reform Act of 1978 on consumer bankruptcy". J. of Law and Econ. 26, 803–835.
Domowitz, I., Sartain, R. (1999). "Determinants of the consumer bankruptcy decision". J. of Finance 54, 403–420.
Dye, R. (1986). "An economic analysis of bankruptcy statutes". Economic Inquiry 24, 417–428.
Eberhart, A.C., Moore, W.T., Roenfeldt, R.L. (1990). "Security pricing and deviations from the absolute priority rule in bankruptcy proceedings". J. of Finance 44, 747–769.
Executive Office for U.S. Trustees (2001). "United States trustee program: preliminary report on Chapter 7 Asset Cases 1994 to 2000". (Available at archive.gao.gov/t2pbat2/152238.pdf.)
Fama, E.F., Miller, M.H. (1972). The Theory of Finance. Dryden Press, Hinsdale, IL.
Fan, W., White, M.J. (2003). "Personal bankruptcy and the level of entrepreneurial activity". J. of Law & Econ. 46, 543–568.
Fay, S., Hurst, E., White, M.J. (2002). "The household bankruptcy decision". American Economic Rev. 92, 706–718.
Fisher, J.D. (2003). "The effect of unemployment benefits, welfare benefits and other income on personal bankruptcy". Bureau of Labor Statistics Working Paper.
Franks, J.R., Sussman, O. (2005). "Financial innovations and corporate bankruptcy". J. of Financial Intermediation 14 (3), 283–317.
Franks, J.R., Torous, W.N. (1989). "An empirical investigation of U.S. firms in reorganization". J. of Finance 44, 747–769.
Franks, J.R., Torous, W.N. (1994). "A comparison of financial recontracting in distressed exchanges and Chapter 11 reorganizations". J. of Financial Econ. 35, 347–370.
Franks, J.R., Nybourg, K., Torous, W.N. (1996). "A comparison of U.S., U.K., and German insolvency codes". Financial Management 25, 19–30.
Gertner, R., Scharfstein, D. (1991). "A theory of workouts and the effects of reorganization law". J. of Finance 44, 1189–1222.
Gilson, S.C. (1990). "Bankruptcy, boards, banks and blockholders". J. of Financial Econ. 27, 355–387.
Gilson, S.C., Vetsuypens, M.R. (1994). "Creditor control in financially distressed firms: empirical evidence". Wash. Univ. Law Quarterly 72, 1005–1025.
Gilson, S.C., John, K., Lang, L. (1990). "Troubled debt restructurings: an empirical study of private reorganization of firms in default". J. of Financial Econ. 27, 315–355.

Gordon, R.H., Malkiel, B. (1981). "Corporation finance". In: Aaron, H., Pechman, J. (Eds.), How Taxes Affect Economic Behavior. Brookings Institution, Washington, D.C.

Grant, C. (2005). "Evidence on the effect of U.S. bankruptcy exemptions". In: Disney, R., Grant, C., Bertola, G. (Eds.), Economics of Consumer Credit: European Experience and Lessons From the U.S. MIT Press, Cambridge, MA.

Gropp, R.J., Scholz, K., White, M.J. (1997). "Personal bankruptcy and credit supply and demand". Quarterly J. of Econ. 112, 217–252.

Gross, D.B., Souleles, N.S. (2002). "An empirical analysis of personal bankruptcy and delinquency". Rev. of Financial Studies 15, 319–347.

Han, S., Li, W. (2004). "Fresh start or head start? The effect of filing for personal bankruptcy on labor supply". Federal Reserve Bank of Philadelphia Working Paper.

Hart, O.D., Moore, J. (1998). "Default and renegotiation: a dynamic model of debt". Quarterly J. of Econ. 113, 1–41.

Hart, O.D., La Porta Drago, R., Lopez-de-Silanes, F., Moore, J. (1997). "A new bankruptcy procedure that uses multiple auctions". European Economic Review 41, 461–473.

Hotchkiss, E. (1995). "Postbankruptcy performance and management turnover". J. of Finance 50, 3–21.

Hynes, R.M. (2002). "Optimal bankruptcy in a non-optimal world". Boston College Law Review XLIV (1), 1–78.

Hynes, R.M. (2004). "Why (consumer) bankruptcy?" Alabama Law Review 56 (1), 121–179.

Hynes, R.M., Malani, A., Posner, E.A. (2004). "The political economy of property exemption laws". J. of Law & Economics 47, 19–43.

Jensen, M., Meckling, W. (1976). "Theory of the firm: managerial behavior, agency costs, and capital structure". J. of Financial Econ. 3, 305–360.

Jackson, T.H. (1986). The Logic and Limits of Bankruptcy Law. Harvard University Press, Cambridge, MA.

Kahan, M., Tuckman, B. (1993). "Do bondholders lose from junk bond covenant changes?" J. of Business 66, 499–516.

Lehnert, A., Maki, D.M. (2002). "Consumption, debt and portfolio choice: testing the effects of bankruptcy law". Board of Governors Working Paper.

Lin, E.Y., White, M.J. (2001). "Bankruptcy and the market for mortgage and home improvement loans". J. of Urban Econ. 50, 138–162.

Longhofer, S.D. (1997). "Absolute priority rule violations, credit rationing, and efficiency". J. of Financial Intermediation 6, 249–267.

LoPucki, L. (1983). "The debtor in full control: systems failure under Chapter 11 of the bankruptcy code?" American Bankruptcy Law J. 57, 247–273.

LoPucki, L., Whitford, W. (1990). "Bargaining over equity's share in the bankruptcy reorganization of large, publicly held companies". Univ. of Penn. Law Rev. 139, 125–196.

McConnell, M.W., Picker, R.C. (1993). "When cities go broke: a conceptual introduction to municipal bankruptcy". Univ. of Chicago Law Rev. 60, 425–495.

Myers, S. (1977). "Determinants of corporate borrowing". J. of Financial Econ. 5, 147–175.

Olson, M.L. (1999). "Avoiding default: the role of credit in the consumption collapse of 1930". Quarterly J. of Econ. 114, 319–335.

Peterson, R.L., Aoki, K. (1984). "Bankruptcy filings before and after implementation of the bankruptcy reform law". J. of Econ. and Business 36, 95–105.

Posner, E.A. (1995). "Contract law in the welfare state: a defense of the unconscionability doctrine, usury laws, and related limitations on the freedom to contract". J. of Legal Studies 24, 283–319.

Posner, E.A. (1997). "The political economy of the Bankruptcy Reform Act of 1978". Mich. Law Rev. 96, 47–126.

Povel, P. (1999). "Optimal 'soft' or 'tough' bankruptcy procedures". J. of Law, Econ. and Org. 15, 659–684.

Rasmussen, R.K. (1992). "Debtor's choice: a menu approach to corporate bankruptcy". Texas Law Rev. 71, 51–121.

Rea, S.A. (1984). "Arm-breaking, consumer credit and personal bankruptcy". Economic Inquiry 22, 188–208.

Roe, M.J. (1983). "Bankruptcy and debt: a new model for corporate reorganization". Columbia Law Rev. 83, 527–602.
Roe, M.J. (1987). "The voting prohibition in bond workouts". Yale Law J. 97, 232–279.
Schwartz, A. (1993). "Bankruptcy workouts and debt contracts". J. of Law and Econ. 36, 595–632.
Schwartz, A. (1997). "Contracting about bankruptcy". J. of Law, Econ., and Org. 13, 127–146.
Shepard, L. (1984). "Personal failures and the Bankruptcy Reform Act of 1978". Journal of Law and Economics 27, 419–437.
Shleifer, A., Vishny, R.W. (1992). "Liquidation values and debt capacity: a market equilibrium approach". J. of Finance 47, 1343–1366.
Smith, C.W., Warner, J.B. (1979). "On financial contracting: an analysis of bond covenants". J. of Financial Economics 7 (2), 117–161.
Stiglitz, J.E. (1972). "Some aspects of the pure theory of corporate finance: bankruptcies and take-overs". Bell J. of Econ. 3, 458–482.
Stulz, R., Johnson, H. (1985). "An analysis of secured debt". J. of Financial Econ. 14, 501–521.
Sullivan, T., Warren, E., Westbrook, J. (1989). As We Forgive Our Debtors. Oxford University Press, New York, NY.
Tashjian, E., Lease, R., McConnell, J. (1996). "Prepacks: an empirical analysis of prepackaged bankruptcies". J. of Financial Econ. 40, 135–162.
Triantis, G.G. (1993). "The effects of insolvency and bankrutpcy on contract performance and adjustment". Univ. of Toronto Law J. 43, 679–710.
U.S. Department of Commerce, Bureau of the Census. Statistical Abstract of the United States 1988, 108th Edition. U.S. Government Printing Office, Washington, D.C. (1987).
Wang, H.-J., White, M.J. (2000). "An optimal personal bankruptcy system and proposed reforms". Journal of Legal Studies 39, 255–286.
Warren, C. (1935). Bankruptcy in United States History. Da Capo Press, New York (1972).
Webb, D.C. (1987). "The importance of incomplete information in explaining the existence of costly bankruptcy". Economica 54, 279–288.
Webb, D.C. (1991). "An economic evaluation of insolvency procedures in the United Kingdom: does the 1986 Insolvency Act satisfy the creditors' bargain?" Oxford Econ. Papers 43, 139–157.
Weiss, L.A. (1990). "Bankruptcy resolution: direct costs and violation of priority of claims". J. of Financial Econ. 27, 285–314.
Weiss, L.A., Wruck, K.H. (1998). "Information problems, conflicts of interest, and asset stripping: Chapter 11's failure in the case of Eastern Airlines". J. of Financial Econ. 48, 55–97.
White, M.J. (1980). "Public policy toward bankruptcy: me-first and other priority rules". Bell J. of Econ. 11, 550–564.
White, M.J. (1983). "Bankruptcy costs and the new bankruptcy code". J. of Finance 38, 477–488.
White, M.J. (1987). "Personal bankruptcy under the 1978 Bankruptcy Code: an economic analysis". Indiana Law Journal 63, 1–57.
White, M.J. (1989). "The corporate bankruptcy decision". J. of Econ. Perspectives 3, 129–151.
White, M.J. (1994). "Corporate bankruptcy as a filtering device: Chapter 11 reorganizations and out-of-court debt restructurings". J. of Law, Econ., and Org. 10, 268–295.
White, M.J. (1996). "The costs of corporate bankruptcy: a U.S.–European comparison". In: Bhandari, J., Weiss, L. (Eds.), Corporate Bankruptcy: Economic and Legal Perspectives. Cambridge University Press, Cambridge, UK.
White, M.J. (1998a). "Why don't more households file for bankruptcy?" J. of Law, Econ. and Org. 14, 205–231.
White, M.J. (1998b). "Why it pays to file for bankruptcy: a critical look at incentives under U.S. bankruptcy laws and a proposal for change". Univ. of Chicago Law Rev. 65, 685–732.
White, M.J. (2002). "Sovereigns in distress: do they need bankruptcy?" Brookings Papers on Econ. Activity 1, 287–319.
White, M.J. (2005). "Personal bankruptcy: insurance, work effort, opportunism and the efficiency of the 'fresh start' ". Working Paper. (Available at www.ucsd.edu/~miwhite.)
White, M.J. (2007). "Bankruptcy reform and credit cards". J. of Economic Perspectives, in press.

Chapter 15

ANTITRUST

LOUIS KAPLOW
School of Law, Harvard University, and National Bureau of Economic Research

CARL SHAPIRO
Haas School of Business and Department of Economics, University of California, Berkeley

Contents

Handbook of Law and Economics, Volume 2
Edited by A. Mitchell Polinsky and Steven Shavell

DOI: 10.1016/S1574-0730(07)02015-4

Abstract

This is a survey of the economic principles that underlie antitrust law and how those principles relate to competition policy. We address four core subject areas: market power, collusion, mergers between competitors, and monopolization. In each area, we select the most relevant portions of current economic knowledge and use that knowledge to critically assess central features of antitrust policy. Our objective is to foster the improvement of legal regimes and also to identify topics where further analytical and empirical exploration would be useful.

Keywords

antitrust, competition policy, monopoly, market power, market definition, oligopoly, collusion, cartels, price-fixing, facilitating practices, mergers, horizontal mergers, unilateral effects, coordinated effects, monopolization, exclusionary practices, predatory pricing, exclusive dealing

JEL classification: K21, L12, L13, L40, L41, L42

1. Introduction

In this chapter, we survey the economic principles that underlie antitrust law and use these principles to illuminate the central challenges in formulating and applying competition policy. Our twin goals are to inform readers about the current state of knowledge in economics that is most relevant for understanding antitrust law and policy and to critically appraise prevailing legal principles in light of current economic analysis.

Since the passage of the Sherman Act in 1890, antitrust law has always revolved around the core economic concepts of competition and market power. For over a century, it has been illegal in the United States for competitors to enter into price-fixing cartels and related schemes and for a monopolist to use its market power to stifle competition. In interpreting the antitrust statutes, which speak in very general terms, U.S. courts have always paid attention to economics. Yet the role of economics in shaping antitrust law has evolved greatly, especially over the past few decades. The growing influence of economics on antitrust law can be traced in part to the Chicago School, which, starting in the 1950s, launched a powerful attack on many antitrust rules and case outcomes that seemed to lack solid economic underpinnings. But the growing influence of economics on antitrust law also has resulted from substantial theoretical and empirical advances in industrial organization economics over the period since then. With a lag, often spanning a couple of decades, economic knowledge shapes antitrust law. It is our hope in this essay both to sharpen economists' research agendas by identifying open questions and difficulties in applying economics to antitrust law, and also to accelerate the dissemination of economic knowledge into antitrust policy.

Antitrust economics is a broad area, overlapping to a great extent with the field of industrial organization. We do not offer a comprehensive examination of the areas within industrial organization economics that are relevant for antitrust law. That task is far too daunting for a single survey and is already accomplished in the form of the three-volume *Handbook of Industrial Organization* (1989a, 1989b, 2007).[1] Instead, we focus our attention on four core economic topics in antitrust: the concept of market power (section 2), the forces that facilitate or impede efforts by competitors to engage in collusion (section 3), the effects of mergers between competitors (section 4), and some basic forms of single-firm conduct that can constitute illegal monopolization, namely predatory pricing and exclusive dealing (section 5).[2] In each case, we attempt to select from the broad base of models and approaches the ones that seem most helpful in formulating a workable competition policy. Furthermore, we use this analysis to scrutinize the corresponding features of antitrust law, in some cases providing a firmer rationalization

[1] Schmalensee and Willig (1989a, 1989b) and Armstrong and Porter (2007).

[2] Since the field of antitrust economics and law is far too large to cover in one chapter, we are forced to omit some topics that are very important in practice and have themselves been subject to extensive study, including joint ventures (touched on briefly in subsection 3.5.2), vertical mergers, bundling and tying, vertical intrabrand restraints, the intersection of antitrust law and intellectual property law, and most features of enforcement policy and administration, including international dimensions.

for current policy and in others identifying important divergences.[3] For reasons of concreteness and of our own expertise, we focus on antitrust law in the United States, but we also emphasize central features that are pertinent to competition policy elsewhere and frequently relate our discussion to the prevailing regime in the European Union.[4]

2. Market power

The concept of market power is fundamental to antitrust economics and to the law. Except for conduct subject to *per se* treatment, antitrust violations typically require the government or a private plaintiff to show that the defendant created, enhanced, or extended in time its market power. Although the requisite degree of existing or increased market power varies by context, the nature of the inquiry is, for the most part, qualitatively the same.

It is important to emphasize at the outset that the mere possession of market power is not a violation of antitrust law in the United States. Rather, the inquiry into market power is usually a threshold question; if sufficient market power is established, it is then asked whether the conduct in question—say, a horizontal merger or an alleged act of monopolization—constitutes an antitrust violation. If sufficient market power is not demonstrated, the inquiry terminates with a victory for the defendant.

Here, we begin our treatment of antitrust law and economics with a discussion of the basic economic concept of market power and its measurement. Initially, we define market power, emphasizing that, as a technical matter, market power is a question of degree. Then we explore the factors that determine the extent of market power, first when exercised by a single firm and then in the case in which multiple firms interact. We also consider various methods of inferring market power in practice and offer some further remarks about the relationship between the concept of market power as understood by

[3] There are a number of books that have overlapping purposes, including Bork (1978), Hylton (2003), Posner (2001), and Whinston (2006), the latter being closest to the present essay in the weight given to formal economics.

[4] As implied by the discussion in the text, our references to the law are primarily meant to make concrete the application of economic principles (and secondarily to offer specific illustrations) rather than to provide detailed, definitive treatments. On U.S. law, the interested reader should consult the extensive treatise *Antitrust Law* by Areeda and Hovenkamp, many volumes of which are cited throughout this essay. On the law in the European Union, see, for example, Bellamy and Child (2001), Dabbah (2004), and Walle de Ghelcke and Gerven (2004). A wide range of additional information, including formal policy statements and enforcement statistics, are now available on the Internet. Helpful links are: Antitrust Division, Department of Justice: http://www.usdoj.gov/atr/index.html; Bureau of Competition, Federal Trade Commission: http://www.ftc.gov/ftc/antitrust.htm; European Union, DG Competition: http://ec.europa.eu/comm/competition/index_en.html; Antitrust Section of the American Bar Association: http://www.abanet.org/antitrust/home.html.

economists and as employed in antitrust law.[5] Further elaboration appears in sections 4 and 5 on horizontal mergers and monopolization, respectively.

2.1. Definition of market power

Microeconomics textbooks distinguish between a price-taking firm and a firm with some power over price, that is, with some market power. This distinction relates to the demand curve facing the firm in question. Introducing our standard notation for a single firm selling a single product, we write P for the price the firm receives for its product, X for the firm's output, and $X(P)$ for the demand curve the firm perceives that it is facing, with $X'(P) \leq 0$.[6] When convenient, we will use the inverse demand curve, $P(X)$. A price-taking firm has no control over price: $P(X) = P$ regardless of X, over some relevant range of the firm's output. In contrast, a firm with power over price can cause price to rise or fall by decreasing or increasing its output: $P'(X) < 0$ in the relevant range. We say that a firm has "technical market power" if it faces a downward sloping (rather than horizontal) demand curve.

In practice almost all firms have some degree of technical market power. Although the notion of a perfectly competitive market is extremely useful as a theoretical construct, most real-world markets depart at least somewhat from this ideal. An important reason for this phenomenon is that marginal cost is often below average cost, most notably for products with high fixed costs and few or no capacity constraints, such as computer software, books, music, and movies. In such cases, price must exceed marginal cost for firms to remain viable in the long run.[7] Although in theory society could mandate that all prices equal marginal cost and provide subsidies where appropriate, this degree of regulation is generally regarded to be infeasible, and in most industries any attempts to do so are believed to be inferior to reliance upon decentralized market interactions. Antitrust law has the more modest but, it is hoped, achievable objective of enforcing competition to the extent feasible. Given the near ubiquity of some degree of technical market power, the impossibility of eliminating it entirely, and the inevitable costs of antitrust intervention, the mere fact that a firm enjoys some technical market power is not very informative or useful in antitrust law.

[5] Prior discussions of the general relationship between the economic conception of market power and its use in antitrust law include Areeda, Kaplow, and Edlin (2004, pp. 483–499), Kaplow (1982), and Landes and Posner (1981). For a recent overview, see American Bar Association, Section of Antitrust Law (2005).

[6] For simplicity, unless we indicate otherwise, we assume throughout this chapter that each firm sells a single product. While this assumption is almost always false, in many cases it amounts to looking at a firm's operations product-by-product. Obviously, a multi-product firm might have market power with respect to one product but not others. When interactions between the different products sold by a multi-product firm are important, notably, when the firm sells a line of products that are substitutes or complements for each other, the analysis will need to be modified.

[7] Edward Chamberlin (1933) and Joan Robinson (1933) are classic references for the idea that firms in markets with low entry barriers but differentiated products have technical market power.

Nonetheless, the technical, textbook notion of market power has the considerable advantage that it is amenable to precise measurement, which makes it possible to identify practices that enhance a firm's power to a substantial degree. The standard measure of a firm's technical market power is based on the difference between the price the firm charges and the firm's marginal cost. In the standard theory of monopoly pricing, a firm sets the price for its product to maximize profits. Profits are given by $\pi = PX(P) - C(X(P))$, where $C(X)$ is the firm's cost function. Differentiating with respect to price, we get the standard expression governing pricing by a single-product firm,

$$\frac{P - MC}{P} = \frac{1}{|\varepsilon_F|}, \tag{1}$$

where MC is the firm's marginal cost, $C'(X)$, and $\varepsilon_F \equiv \frac{dX}{dP}\frac{P}{X}$ is the elasticity of demand facing that firm, the "firm-specific elasticity of demand."[8] The left-hand side of this expression is the Lerner Index, the percentage gap between price and marginal cost, which is a natural measure of a firm's technical market power:

$$m \equiv \frac{P - MC}{P}.$$

As noted earlier, some degree of technical market power is necessary for firms to cover their costs in the presence of economies of scale. For example, if costs are given by $C(X) = F + CX$, then profits are given by $\pi = PX - CX - F$ and the condition that profits are non-negative can be written as $m \geq F/PX$, that is, the Lerner Index must be at least as large as the ratio of the fixed costs, F, to the firm's revenues, $R \equiv PX$.

Before proceeding with our analysis, we note that, although anticompetitive harm can come in the form of reduced product quality, retarded innovation, or reduced product variety, our discussion will follow much of the economics literature and most antitrust analysis in focusing on consumer harm that comes in the form of higher prices. This limitation is not as serious as may first appear because higher prices can serve as a loose proxy for other forms of harm to consumers.

2.2. *Single-firm pricing model accounting for rivals*

To aid understanding, we present a basic but flexible model showing how underlying supply and demand conditions determine the elasticity of demand facing a given firm. This model allows us to begin identifying the factors that govern the degree of technical market power enjoyed by a firm. We also note that this same model will prove very useful conceptually when we explore below the impact of various practices on price. Studying the effects of various practices on price requires some theory of how firms set

[8] Strictly speaking, the elasticity of demand facing the firm is endogenous, except in the special case of constant elasticity of demand, since it varies with price, an endogenous variable. All the usual formulas refer to the elasticity of demand at the equilibrium (profit-maximizing) price level.

their prices. The building block for these various theories is the basic model of price-setting by a single, profit-maximizing firm. In addition, as a matter of logic, one must begin with such a model before moving on to theories that involve strategic interactions among rival firms.

The standard model involves a dominant firm facing a competitive fringe.[9] A profit-maximizing firm sets its price accounting for the responses it expects from its rivals and customers to the price it sets.[10] This is a decision-theoretic model, not a game-theoretic model, so it does not make endogenous the behavior of the other firms in the market or of potential entrants. This is the primary sense in which the generality of the model is limited. The model also is limited because it assumes that all firms in the market produce the same, homogeneous product and do not engage in any price discrimination, although the core ideas underlying it extend to models of differentiated products.

The firm faces one or more rivals that, as noted, sell the same, homogeneous product. When setting its price, P, the firm recognizes that rivals will likely respond to higher prices by producing more output. The combined output of the firm's rivals increases with price according to $Y(P)$, with $Y'(P) \geq 0$. Total (market) demand declines with price according to $Z(P)$, with $Z'(P) \leq 0$. If the firm in question sets the price P, then it will be able to sell an amount given by $X(P) \equiv Z(P) - Y(P)$. This is the largest quantity that the firm can sell without driving price below the level P that it selected; if the firm wants to sell more, it will have to lower its price. The firm's so-called "residual demand curve" is therefore given by $X(P)$.

If we differentiate the equation defining $X(P)$ with respect to P, and then multiply both sides by $-P/X$ to convert the left-hand side into elasticity form, we get

$$-\frac{P}{X}\frac{dX}{dP} = -\frac{P}{X}\frac{dZ}{dP} + \frac{P}{X}\frac{dY}{dP}.$$

Next, multiply and divide the dZ/dP term on the right-hand side by Z and the dY/dP term by Y. This gives

$$-\frac{P}{X}\frac{dX}{dP} = -\frac{P}{Z}\frac{dZ}{dP}\frac{Z}{X} + \frac{P}{Y}\frac{dY}{dP}\frac{Y}{X}.$$

Define the market share of the firm being studied by $S = X/Z$. The corresponding market share of the rivals is $1-S = Y/Z$. Replacing Z/X by $1/S$ and Y/X by $(1-S)/S$ in the expression above gives

$$-\frac{P}{X}\frac{dX}{dP} = -\frac{P}{Z}\frac{dZ}{dP}\frac{1}{S} + \frac{P}{Y}\frac{dY}{dP}\frac{(1-S)}{S}.$$

[9] For a recent textbook treatment of this model, see Carlton and Perloff (2005, pp. 110–119). Landes and Posner (1981) provide a nice exposition of this model in the antitrust context.

[10] As with the standard theory of pure monopoly pricing as taught in microeconomics textbooks, the results of this model are unchanged if we model the firm as choosing its output level, with price adjusting to clear the market.

Call the elasticity of supply of the rivals $\varepsilon_R \equiv \frac{P}{Y}\frac{dY}{dP}$, and the absolute value of the elasticity of the underlying market demand curve $|\varepsilon_D| \equiv -\frac{P}{Z}\frac{dZ}{dP}$. The absolute value of the elasticity of demand facing the firm, $|\varepsilon_F| \equiv -\frac{P}{X}\frac{dX}{dP}$, is therefore given by

$$|\varepsilon_F| = \frac{|\varepsilon_D| + (1 - S)\varepsilon_R}{S}. \tag{2}$$

This equation captures the central lesson from this model: the absolute value of the elasticity of demand facing a single firm, given the supply curves of its price-taking rivals and the demand curve of the buyers in its market, is governed by three variables: (1) the underlying elasticity of demand for the product, $|\varepsilon_D|$, which is frequently called the market elasticity of demand; (2) the elasticity of supply of the firm's rivals, ε_R; and (3) the firm's market share, S. The magnitude of the firm-specific elasticity of demand is larger, the larger are the magnitudes of the market elasticity of demand and the elasticity of supply of the firm's rivals and the smaller is the firm's market share. Intuitively, market share is relevant for two reasons: the smaller the firm's share, the greater the share of its rivals and thus the greater is the absolute magnitude of their supply response to a price increase for a given supply elasticity, ε_R; and the smaller the firm's share, the smaller is its share of the increase in industry profits due to a given sacrifice in its own sales.[11]

One polar case in this basic model is that of the traditional monopolist. With no rivals, $S = 1$, so the elasticity of demand facing the firm is just the market elasticity of demand. With rivals, however, the magnitude of the firm-specific elasticity of demand is larger than that of the market elasticity of demand. The other polar case is that of the firm from the theory of perfectly competitive markets. As the firm's share of the market approaches zero, the magnitude of the firm-specific elasticity of demand becomes infinite, that is, the firm is a price-taker.

We can directly translate the firm-specific elasticity of demand given by expression (2) into the profit-maximizing price. As indicated in expression (1), profit maximization involves setting price so that the firm's gross margin, m, equals the inverse of the magnitude of the firm's elasticity of demand. If there are no rivals, $S = 1$ and this relationship simplifies to the standard monopoly formula, $m = 1/|\varepsilon_D|$. For a firm with a tiny market share, $|\varepsilon_F|$ is enormous, so $m \approx 0$, that is, price nearly equals marginal cost. For intermediate cases, as noted, in this model a large market elasticity of demand, $|\varepsilon_D|$, a high elasticity of rival supply, ε_R, and a small market share, S, all lead to a large firm-specific elasticity of demand facing the price leader, $|\varepsilon_F|$, which in turn implies a small margin.

This model provides a guide for studying the types of conduct that may enhance a firm's technical market power and thus allow that firm profitably to raise its price.

[11] It should be noted that statements about the effect of market share must be interpreted carefully. Thus, an outward shift in the supply curve of the rivals, which lowers the firm's market share at any given price, will raise the elasticity of demand facing that firm at any given price. However, more broadly, the firm's market share is endogenous because it depends on the price the firm chooses.

Generically, such conduct will be that which reduces the value of the right side of expression (2): conduct that makes substitute products less attractive, that causes rivals to reduce their supply, and that raises the firm's market share (through the two former means or otherwise). Later we consider how certain types of conduct having these effects should be scrutinized under antitrust law.

This model is quite broad when one undertakes appropriate interpretations and extensions. For example, issues relating to substitute products bear on the market elasticity of demand, as will be noted below. Additionally, one can account for entry by reflecting it in the rival supply elasticity. One particular variant of the model involves infinitely elastic rival supply, perhaps due to entry, at some fixed "limit" price.

2.3. *Multiple-firm models*

The model in subsection 2.2 took the behavior of all but one firm as exogenous. In this section, we consider game-theoretic models that make predictions regarding the degree of market power exercised by interacting firms. First we consider two standard, static, noncooperative models: Cournot's model of oligopoly, for the case with homogeneous products, and Bertrand's model, for the case with differentiated products. Then we consider briefly the possibility of repeated games and the impact of collusive behavior on market power.[12]

2.3.1. *Cournot model with homogeneous products*

The Cournot (1838) model of oligopoly with homogeneous products is similar to the single-firm pricing model in that it identifies how certain observable characteristics of the market determine the degree of a firm's market power, that is, the percentage markup above marginal cost that the firm charges. The Cournot model goes further, however, by providing predictions about how market structure affects the equilibrium price, predictions that will be important for seeing how certain commercial practices and mergers affect price. Specifically, the model predicts that firms with lower costs will have higher market shares and higher markups. The model is frequently employed in markets with relatively homogeneous products, especially if firms pick their output or capacity levels, after which prices are determined such that the resulting supply equals demand.[13] However, one should bear in mind that the Cournot equilibrium is the Nash equilibrium in a one-shot game. As we discuss at length in section 3, many different outcomes can arise as equilibria in a repeated oligopoly game, even if the stage game played each period involves quantity-setting à la Cournot. In antitrust applications, it is generally desirable

[12] There is an enormous literature on oligopoly theory, which we do not attempt to cover systematically. See, for example, Shapiro (1989), Tirole (1988), and Vives (2001). We discuss models of repeated oligopoly at greater length in section 3 on collusion.

[13] Kreps and Scheinkman (1983) use a particular rationing rule to show that capacity choices followed by pricing competition can replicate the Cournot equilibrium.

to test robustness of results to alternative solution concepts as well as to test empirically the predictions of any oligopoly model that is employed.

In a Cournot equilibrium, a single firm's reaction curve is derived as a special case of the basic model of single-firm pricing: the rivals' outputs are all taken to be fixed, so the rival supply elasticity is zero. As we now show, the elasticity of demand facing a single firm is equal to the market elasticity of demand divided by that firm's market share. However, the Cournot model goes beyond the single-firm pricing model because it involves finding the equilibrium in a game among multiple firms.

Suppose that there are N firms, with each firm i choosing its output X_i simultaneously. The Cournot equilibrium is a Nash equilibrium in these quantities. Total output is $X \equiv X_1 + \cdots + X_N$. Industry or market (inverse) demand is given by $P = P(X)$. Given the output of the other firms, firm i chooses its output to maximize its own profits, $\pi_i = P(X)X_i - C_i(X_i)$. The first-order condition for this firm is $P(X) + X_i P'(X) - C_i'(X_i) = 0$. This can be written as

$$\frac{P - MC_i}{P} = \frac{S_i}{|\varepsilon_D|}, \tag{3}$$

where $S_i \equiv X_i / X$ is firm i's market share, and ε_D, as before, is the market elasticity of demand.

To explore this result, consider the special case in which each firm i has constant marginal cost MC_i. Adding up the first-order conditions for all of the firms gives $NP(X) + XP'(X) = \sum_i MC_i$, which tells us that total output and hence the equilibrium price depend only upon the sum of the firms' marginal costs. Moreover, the markup equation tells us that lower-cost firms have higher market shares and enjoy more technical market power. At the same time, the larger is the market elasticity of demand for this homogeneous product, the smaller is the market power enjoyed by each firm and the lower are the margins at all firms. Here we see a recurrent theme in antitrust: a lower-cost firm may well enjoy some technical market power and capture a large share of the market, but this is not necessarily inefficient. Indeed, with constant marginal costs, full productive efficiency would call for the firm with the lowest marginal cost to serve the entire market.

The Cournot model also predicts that total output will be less than would be efficient because none of the firms produces up to the point at which marginal cost equals price; they all have some degree of market power. In the special case with constant and equal marginal costs, each firm has a market share of $1/N$, and the model predicts that each enjoys technical market power according to the resulting equation $(P - MC)/P = 1/N|\varepsilon_D|$. In this simple sense, more firms leads to greater competition and lower prices. However, this model is clearly incomplete for antitrust purposes: presumably, there are fixed costs to be covered (which is why there is a fixed number of firms in the first place), so adding more firms is not costless.[14] This type of analy-

[14] In general, there is no reason to believe that the equilibrium number of firms in an oligopoly with free entry, that is, where equally efficient firms enter until further entry would drive profits below zero, is socially

sis will be directly relevant when we consider horizontal mergers, which remove an independent competitor but may also lead to efficiencies of various types.

One of the attractive theoretical features of the Cournot model is that it generates an elegant formula for the industry-wide average, output-weighted, price-cost margin, that is, the expression $PCM \equiv \sum_{i=1}^{N} S_i \frac{P-MC_i}{P}$. Using equation (3), we get $PCM \equiv \sum_{i=1}^{N} S_i \frac{S_i}{|\varepsilon_D|}$ or

$$PCM = \frac{1}{|\varepsilon_D|} \sum_{i=1}^{N} S_i^2 = \frac{H}{|\varepsilon_D|}, \tag{4}$$

where $H \equiv \sum S_i^2$ is the Herfindahl-Hirschman Index (HHI) of market concentration that is commonly used in antitrust analysis, especially of horizontal mergers.

2.3.2. *Bertrand model with differentiated products*

The Bertrand model with differentiated products is the other key static model of oligopoly used in antitrust. The Bertrand equilibrium is the Nash equilibrium in the game in which the firms simultaneously set their prices. With N firms selling differentiated products, we can write the demand for firm i's product as $X_i = D_i(P_1, \ldots, P_N)$. As usual, the profits of firm i are given by $\pi_i = P_i X_i - C_i(X_i)$. The Bertrand equilibrium is defined by the N equations $\partial \pi_i / \partial P_i = 0$. Writing the elasticity of demand facing firm i as $\varepsilon_i \equiv \frac{\partial X_i}{\partial P_i} \frac{P_i}{X_i}$, firm i's first-order condition is the usual markup equation,

$$\frac{P_i - MC_i}{P_i} = \frac{1}{|\varepsilon_i|}.$$

Actually solving for the Bertrand equilibrium can be difficult, depending on the functional form for the demand system and on the firms' cost functions. In general, however, we know that a firm faces highly elastic demand if its rivals offer very close substitutes, so the Bertrand theory predicts larger markups when the products offered by the various firms are more highly differentiated. In practice, notably, in the assessment of mergers, particular models of product differentiation are used, such as discrete choice models with random utilities, including logit and nested logit models, or models with linear demand or constant elasticities, as we discuss further in section 4 on horizontal mergers.

Here we illustrate the operation of the Bertrand model by explicitly solving a simple, symmetric, two-firm model with constant marginal costs and linear demand. Write the demand curves as $X_1 = A - P_1 + \alpha P_2$ and $X_2 = A - P_2 + \alpha P_1$. Note that the parameter α measures the diversion ratio, that is, the fraction of sales lost by one firm, when it

efficient. See, for example, Mankiw and Whinston (1986). This observation is relevant in assessing certain antitrust policies: if the equilibrium number of firms is "naturally" too small, then exclusionary conduct on the part of the incumbent oligopolists creates an additional social inefficiency. However, if the equilibrium number of firms is "naturally" excessive, different implications would follow.

raises its price, that are captured by the other firm (assuming that the other firm's price is fixed). The diversion ratio, α, will be important when we study horizontal mergers below.[15]

Call the marginal costs per unit MC_1 and MC_2, respectively, and assume that there are no fixed costs. Then we have $\pi_1 = (P_1 - MC_1)(A - P_1 + \alpha P_2)$. Differentiating with respect to P_1 and setting this equal to zero, we get firm 1's best-response curve, $P_1 = (A + \alpha P_2 + MC_1)/2$. Assuming cost symmetry as well, $MC = MC_1 = MC_2$, in the symmetric Bertrand equilibrium we must have $P_1 = P_2 = P_B$ so we get $P_B = \frac{A+MC}{2-\alpha}$.

We can compare the Bertrand equilibrium price to the price charged by a single firm controlling both products. Such a firm would set P to maximize $(P-MC)(A-P+\alpha P)$, which gives the monopoly price of $P_M = \frac{A+MC(1-\alpha)}{2(1-\alpha)}$. The percentage gap between the monopoly price and the Bertrand price is given by $\frac{P_M-P_B}{P_B} = \frac{\alpha}{2(1-\alpha)}\frac{P_B-MC}{P_B}$.[16] This expression tells us that the Bertrand equilibrium price is relatively close to the monopoly price when the two products are rather poor substitutes, that is, when the diversion ratio, α, is low.

This formula will be highly relevant when studying the effect on price of a merger between two suppliers of differentiated products. In that context, the formula measures the price increase associated with the merger, given the prices charged by other firms (and before accounting for efficiencies). The price increase will depend on the pre-merger margin, $\frac{P_B-MC}{P_B}$, and on the diversion ratio.

2.3.3. *Other game-theoretic models and collusion*

Both the Cournot and Bertrand models assume that firms engage in a one-shot noncooperative game. An extensive literature on repeated games explores the possibility that firms may do better for themselves, supporting what are more colloquially described as collusive outcomes, approaching or equaling the industry profit-maximizing price. As suggested by Stigler (1964) and refined in subsequent work, higher prices tend to be sustainable when cheating can be rapidly detected and effectively punished. For a general discussion of models of collusion, see Jacquemin and Slade (1989) and Shapiro (1989).

The possibility that firms can support alternative equilibria featuring higher prices is important to antitrust analysis. First, it suggests that market power may be higher than is otherwise apparent. Second and more important, the possibility of collusion affects the antitrust analysis of other business conduct. For example, a horizontal merger may have

[15] More generally, the diversion ratio from product i to substitute product j is defined as $\alpha_{ji} = (dX_j/dP_i)/(-dX_i/dP_i)$. Converting this equation into elasticity form gives $\alpha_{ji} = \frac{\varepsilon_{ji}}{|\varepsilon_i|}\frac{X_j}{X_i}$, where $\varepsilon_{ji} = \frac{dX_j}{dP_i}\frac{P_i}{X_j}$ is the cross-elasticity from product i to product j.

[16] The details of these calculations are available at http://faculty.haas.berkeley.edu/shapiro/unilateral.pdf.

only a minor impact on price if the merging firms and their rivals are already colluding, but a far greater effect if the reduction in the number of competitors makes collusion easier to sustain. Also, some practices may facilitate collusion, in which case such practices themselves should potentially be subject to antitrust scrutiny. These possibilities are explored further in section 3 on collusion and section 4 on horizontal mergers.

2.4. Means of inferring market power

Assessing the extent of or increase in technical market power in a given situation is often a difficult undertaking. Based upon the foregoing analysis, one can identify a number of potential strategies whose usefulness varies greatly by context. The legal system has tended to rely primarily on a subset of these approaches, focusing mostly on market definition, as discussed below. In recent years, however, it has increasingly considered alternatives when it has perceived that credible economic evidence has been offered.[17]

Although somewhat crude, it is helpful to group means of inferring market power into three categories. First, since market power is technically defined by the extent of the price-cost margin, one can attempt to identify evidence that bears fairly directly on the size of this margin, or to measure profits (which reflect the margin between price and *average* cost). Second, various models, such as the single-firm price-setting model in subsection 2.2, indicate that the extent of market power will be a function of the elasticity of demand, a firm's market share, and rivals' supply response. Accordingly, one can analyze information indicative of the magnitude of these factors. Third, one can make inferences from firm behavior, notably when observed actions would be irrational unless a certain degree of market power existed or was thereby conferred.

2.4.1. Price-cost margin

2.4.1.1. Direct measurement Observing the extent to which price is above marginal cost indicates the degree of technical market power. This direct approach is feasible if one can accurately measure price and some version of marginal cost, usually average incremental cost.[18] Price is often easy to identify, although complications may arise when multiple products are sold together, making it difficult to determine the incremental revenue associated with the product in question. If different customers are charged

[17] For example, the Supreme Court in Federal Trade Commission v. Indiana Federation of Dentists, 476 U.S. 447, 460–461 (1986) (quoting Areeda's *Antitrust Law* treatise) stated: "Since the purpose of the inquiries into market definition and market power is to determine whether an arrangement has the potential for genuine adverse effects on competition, 'proof of actual detrimental effects, such as a reduction of output,' can obviate the need for an inquiry into market power, which is but a 'surrogate for detrimental effects.' "

[18] We use the terms "marginal cost" and "average incremental cost" interchangeably. Both measure the extra cost per unit associated with increased output. Average incremental cost is a somewhat more accurate term, since one is often interested in increments that do not correspond to "one unit" of output. However, if one takes a flexible approach to what constitutes a "unit" of production, the two terms are exactly the same. In practice, average incremental cost is used to determine gross profit margins.

different prices, it may be necessary to calculate the profit margins for sales to different customers (or at different points of time). Complexities also arise when some sales implicitly bundle other services, such as delivery, short-term financing, and customer support; in principle, these factors can be accommodated by redefining the product to include these services (and tracking the costs associated with these services). Marginal cost, by contrast, may be more difficult to measure, due both to difficulties in identifying which costs are variable (and over what time period) and to the presence of common costs that may be difficult to allocate appropriately. In part for this reason, the empirical industrial organization literature, surveyed in Bresnahan (1989), often treats marginal cost as unobservable.

In some cases, approximate measures of price-cost margins may be sufficient and easy to produce, but as evidenced by disputes over cost in predatory pricing cases and in various regulatory contexts, direct measurement of any conception of cost can be difficult and contentious. In any event, as with all measures of technical market power, it is important to keep in mind the distinction between the extent of market power and whether particular conduct should give rise to antitrust liability. For example, as we have already noted, especially in industries in which marginal cost is below average cost and capacity constraints are not binding, nontrivial technical market power may be consistent with what are normally considered competitive industries.

2.4.1.2. Price comparisons Another fairly direct route to assessing the magnitude of price-cost margins, or at least to provide a lower-bound estimate, is to compare prices across markets. For example, if a firm sells its product for a substantially higher price in one region than in another (taking into account transportation and other cost differences), the price-cost margin in the high-price region should be at least as great as the (adjusted) price difference between the regions. This inference presumes, of course, that the price in the low-price region is at least equal to marginal cost. Note that this method can be understood as a special case of direct measurement. It is assumed that the low price is a proxy for (at least an upper bound on) marginal cost, and one then is measuring the price-cost margin directly.

The *Staples* merger case illustrates an application of this method.[19] The government offered (and the court was convinced by) data indicating that prices were higher in regional markets in which fewer office supply superstores operated and that prices fell when new superstore chains entered. This was taken as powerful evidence that a merger of two of the existing three superstores would lead to price increases.

2.4.1.3. Price discrimination Price comparisons often involve a special case of price discrimination, wherein a given firm charges different prices to different consumers, contrary to the implicit assumption in the earlier analysis that each firm sets a single

[19] Federal Trade Commission v. Staples, Inc., 970 F. Supp. 1066 (D.D.C. 1997). For further discussion, see subsection 4.6.1 in our discussion of horizontal mergers.

price for all of its customers. Accordingly, for essentially the same reason as that just given, the ability of a firm to engage in price discrimination implies the existence of market power. If one is prepared to assume that the firm is not pricing below marginal cost to any of its customers, and if one accounts for differences in the cost of serving different customers, the percentage difference between any high price it charges and the lowest price it charges for the same product can serve as a lower bound on the percentage markup associated with the higher price. For example, the substantial price discrimination in sales of pharmaceutical drugs on international markets shows that prices in the United States are very much above marginal cost.

The fact that price discrimination technically implies market power is important because price discrimination is widespread. Familiar examples include airline pricing, senior citizen and student discounts, and the mundane practice of restaurants charging steep price increments for alcoholic beverages (versus soft drinks) and high-end entrees that greatly exceed any differences in marginal cost. For business-to-business transactions, negotiations that typically generate price dispersion and price discrimination are quite common.

Once again, however, it is important to keep in mind that the existence of technical market power does not imply antitrust liability.[20] As is familiar, price discrimination generates greater seller profits yet may well be benign or even favorable on average for consumers. Moreover, the resulting profit margins are often necessary to cover fixed costs, as in models of monopolistic competition. If there are no barriers to entry so that the resulting margins merely provide a normal rate of return on capital, the presence of a gap between price and marginal cost is perfectly consistent with the conclusion that the market is behaving is a competitive fashion, given the presence of fixed costs and product differentiation. Furthermore, in our preceding example of multinational pharmaceutical companies, the margins provide the reward for costly and risky research and development to create and patent new drugs. The ex post market power is necessary to provide the quasi-rents that induce innovation (given that we rely on a patent system rather than a regime that gives direct rewards to innovators from the government fisc).

2.4.1.4. Persistent profits A somewhat different approach to establishing antitrust market power involves looking at a firm's profits, which amounts to comparing price to average (rather than marginal) cost. Under this approach, persistently above-normal

[20] Nor is it the case that price discrimination in itself implies antitrust liability, despite the existence of the Robinson-Patman Act that regulates particular sorts of price discrimination in certain contexts. As presently interpreted, price discrimination may be a violation in so-called primary-line cases, tantamount to predatory pricing, and in secondary-line cases, such as when manufacturers offer discounts (that are not cost justified) to large retailers that are not available to smaller buyers. Notably, the Act does not cover discriminatory prices to ultimate consumers (or to intermediaries that are not in competition with each other) that are nonpredatory. Nevertheless, it seems that defendants in antitrust litigation have been reluctant to rationalize challenged practices that analysts have suggested were means of price discrimination on such grounds, presumably fearing that such explanations would be to their detriment. Of course, one way this could be true is that the existence of some technical market power would thereby be conceded.

profits indicate a high price-cost margin and thus the existence of technical market power. This method shares difficulties with any that rely on measures of cost. In particular, it is often very hard to measure the return on capital earned for a given product, or in a given market, especially for a firm that is engaged in many lines of business and has substantial costs that are common across products.[21] Another problem with this approach is that the return on capital should, in principle, be adjusted for risk. Frequently, one is looking at a successful firm, perhaps one that has been highly profitable for many years following some initial innovation that, *ex ante*, may not have turned out as well.

In addition, average costs often differ from marginal costs. When average costs are higher, this approach may mask the existence of technical market power. In such circumstances, however, marginal-cost pricing may be unsustainable in any event; that is, although there may be technical market power, there may not be any way (short of intrusive regulation that is not contemplated) to improve the situation. When average cost is below marginal cost, profits can exist despite the absence of any markup. In such cases, entry might be expected. If profits are nevertheless persistent, there may exist entry barriers, a subject we discuss below.

2.4.2. *Firm's elasticity of demand*

In the single-firm pricing model, the price-cost margin (Lerner Index) equals the inverse of the (absolute value of the) firm's elasticity of demand, as indicated by expression (1). Furthermore, as described in expression (2), this elasticity depends on the market elasticity of demand, the firm's market share, and rivals' supply elasticity. In the Cournot, Bertrand, and other oligopoly models, many of the same factors bear on the extent of the price-cost margin and thus the degree of market power. Accordingly, another route to inferring market power is to consider the magnitude of these factors.

2.4.2.1. Direct measurement One could attempt to measure the elasticity of demand facing the firm in question.[22] A possible approach would be to estimate the market elasticity of demand and then make an adjustment based on the firm's market share. Alternatively, one might directly observe how the firm's sales have varied when it has changed its price. As a practical matter, both of these methods may be difficult to implement. However, they may nevertheless be more reliable than the alternatives.

2.4.2.2. Substitutes, market definition, and market share In antitrust analysis, both by agencies (notably, in examining prospective horizontal mergers) and by the courts, the dominant method of gauging the extent of market power involves defining a so-called relevant market and examining the share of a firm or group of firms in that market. In defining product markets, the focus is on which products are sufficiently good demand

[21] See, for example, Fisher and McGowan (1983).
[22] See, for example, Baker and Bresnahan (1988).

substitutes for the product in question to be deemed in the same market. Likewise, in defining the extent of the geographic market, the question concerns the feasibility of substitution, for example, by asking how far patients would travel for hospitalization. Although we have discussed the economic analysis of market power at some length, the concept of market definition has not yet appeared directly. Hence it is useful to consider the relationship between the most common method used in antitrust law to assess market power and the implications of the foregoing economic analysis.

The connection is easiest to see by examining expression (2), which relates the firm-specific elasticity of demand to the market elasticity of demand, the firm's market share, and rivals' elasticity of substitution. Consider the case in which the firm produces a homogeneous product, has a high share of sales of that product, and faces a highly elastic market demand curve due to the existence of many close substitutes. The firm-specific elasticity of demand is high and thus the extent of technical market power is small even though the firm's market share is high in the narrowly defined market consisting only of the homogeneous product sold by the firm. One could redefine the "market" to include the close substitutes along with the homogeneous product sold by the firm. The market elasticity of demand in this broader market is presumably smaller, but since the firm's market share in this market is also necessarily lower, we would again conclude that the firm-specific demand elasticity is large and thus that the degree of technical market power is low.

Courts—and thus lawyers and government agencies—traditionally equate high market shares with a high degree of market power and low shares with a low degree of market power. This association is highly misleading if the market elasticity of demand is ignored, and likewise if rivals' elasticity of supply is not considered. In principle, as just explained, the paradigm based on market definition and market share takes the market elasticity of demand into account, indirectly, by defining broader markets—and thus producing lower market shares—when the elasticity is high. As should be apparent from the foregoing discussion, the standard antitrust approach is more indirect than necessary and, due to this fact plus its dichotomous structure (substitutes are either in the market or not), will tend to produce needlessly noisy conclusions.[23] We discuss market definition at greater length in subsection 4.5 on horizontal mergers and subsection 5.2.1 on monopolization.

Frequently, it is useful to decompose the elasticity of demand for a given product into various cross-elasticities of demand with other products. For example, if the price of soda rises, consumers will substitute to other drinks, including, perhaps, beer, juice, milk, and water. Naturally, the analysis in any given case will depend upon exactly how these various products are defined (soda could be broken into regular soda and diet soda, or colas and non-colas, etc.). But the underlying theory of demand does not vary with such definitions. To illustrate, suppose that consumers allocate their total income of I across N distinct products, so $\sum_{i=1}^{N} P_i X_i = I$. To study the elasticity of

[23] This point is elaborated in Kaplow (1982).

demand for product 1, suppose that P_1 rises and the other prices remain unchanged. Then we get $X_1 + P_1 \frac{dX_1}{dP_1} + \sum_{i=2}^{N} P_i \frac{dX_i}{dP_1} = 0$. Converting this to elasticity form gives $-\frac{P_1}{X_1}\frac{dX_1}{dP_1} = 1 + \sum_{i=2}^{N} \frac{P_1}{X_i}\frac{dX_i}{dP_1}\frac{P_i X_i}{P_1 X_1}$. Defining the cross-elasticity between product i and product 1 as $\varepsilon_{i1} = \frac{dX_i}{dP_1}\frac{P_1}{X_i}$, and the revenues associated with product i as $R_i = P_i X_i$, this can be written as

$$|\varepsilon_{11}| = 1 + \sum_{i=2}^{N} \varepsilon_{i1} \frac{R_i}{R_1}. \tag{5}$$

In words, the (absolute value of the) elasticity of demand for product 1 is equal to one plus the sum of the cross-elasticities of all the other products with product 1, with each cross-elasticity weighted by the associated product's revenues relative to those of product 1. If we define each product's share of expenditures as $s_i = R_i/I$, then expression (5) can be written as $|\varepsilon_{11}| = 1 + \frac{1}{s_1}\sum_{i=2}^{N} s_i \varepsilon_{i1}$, so the cross-elasticity with each rival product is weighted by its share of revenues.[24]

This decomposition of the market elasticity of demand is instructive with regard to the standard practice in antitrust of defining markets by deciding whether particular products are sufficiently good substitutes—generally understood as having sufficiently high cross-elasticities of demand—to be included in the market. The expression makes clear that even a substitute with a very high cross-elasticity may have much less influence than that of a large group of other products, no one of which has a particularly high cross-elasticity. Moreover, products' shares of total revenues are not ordinarily considered in an explicit way, yet the formula indicates that a substitute with half the cross-elasticity of another can readily be more important, in particular, if its associated revenues are more than twice as high. More broadly, this representation of the relationship between overall elasticity and individual cross-elasticities reinforces the point that the effect of substitutes is a matter of degree and thus not well captured by the all-or-nothing approach involved in defining antitrust markets.

Some further comments concerning market share are in order, particularly in light of the fact that a persistently high market share is very frequently presented as compelling evidence that a firm has market power. No doubt this inference is often valid, specifically, if the market demand elasticity and rivals' supply elasticities are low in magnitude and the market conditions are reasonably stable. However, a firm with only a modest cost advantage may profitably maintain its high share by pricing low enough to capture most of the market. This occurs, for example, in the model of the dominant firm facing a competitive fringe if the fringe supply is very elastic at a price just above the firm's own marginal cost. Consider, for example, a trucking firm that provides 100% of

[24] Cross-elasticities need not be positive. For example, when the weighted summation equals zero, we have the familiar case of unit elasticity—that is, as price rises, expenditures on the product in question remain constant—and when the summation is negative, we have an elasticity less than one in absolute value, often referred to as inelastic demand.

the freight transportation on a particular route but would quickly be displaced by nearby rivals (whose costs are essentially the same but who suffer a slight disadvantage due to a lack of familiarity with the route's customers) if it were to raise its price even a few percent. Additionally, a firm may have a 100% share in a market protected by a patent, but if there are sufficiently close substitutes, its market power is negligible. Conversely, even a firm with a low share of sales of a particular product may have quite a bit of technical market power if the magnitude of the market elasticity of demand and rivals' elasticity of supply for that product are very low. Gasoline refining and electricity generation are two examples of products for which this latter situation can arise. In sum, the right side of expression (2) indicates that market share is only one factor that determines the elasticity of demand facing a firm, so the magnitude of market share is a relevant component of market power but not a conclusive indicator.

2.4.2.3. Rivals' supply response: barriers to expansion, mobility, and entry In examining the right side of expression (2) for the firm's elasticity of demand, the preceding subsection focused on the market elasticity of demand and market share. However, the elasticity of supply by rivals is also relevant, as indicated by the just-mentioned contrasting examples of trucking, on one hand, and gasoline refining and electricity generation, on the other hand. The concept of rivals' supply should be understood broadly, to include expanded output from existing plants, shifting capital from other regions or from the production of other products, introducing new brands or repositioning existing ones, and entry by firms in related businesses or by other firms. If market power is significant, it must be that the aggregate of these potential supply responses—often referred to as expansion, mobility, and entry—is sufficiently limited, at least over some pertinent time period. Gilbert (1989) provides an extended discussion of such barriers, Berry and Reiss (2007) survey empirical models of entry and market power, and Sutton (2007) discusses the relationship between market structure and market power.

In some cases, the elasticity of rivals' supply may be measured directly, by measuring output responses to previous changes in price by the firm in question, or by other firms in similar markets. Often, however, some extrapolation is required, such as in predicting whether a hypothetical increase in price to unprecedented levels following a merger would generate a significant supply response. For internal expansion by existing rivals, the question would be whether there exist capacity constraints, steeply rising marginal costs, or limits on the inclination of consumers of differentiated products to switch allegiances. In the case of new entry, timing, possible legal restrictions (intellectual property, zoning, and other regulatory constraints), brand preferences, the importance of learning by doing, and the ability to recoup fixed costs, among other factors, will determine the extent of restraint imposed.

Particularly regarding the latter, it is common to inquire into the existence of so-called barriers to entry (sometimes taken as a shorthand for all forms of supply response by rivals). In some instances, such as when there are legal restrictions, the meaning of this concept is fairly clear. However, in many cases, it is difficult to make sense of the notion of entry barriers in a vacuum. For example, there is much debate about whether

economies of scale should be viewed as a barrier to entry. If minimum efficient scale is large and incumbent producers have long-term exclusive dealing contracts with most distributors, entry may be rendered too costly, and existing firms might enjoy high price-cost margins (more than necessary to cover fixed costs). If instead there merely exist fixed costs and marginal costs are constant, in a free-entry equilibrium there will be positive price-cost margins yet no profits. The positive margins will not induce further entry because their level post-entry would be insufficient to recover fixed costs. As we have observed repeatedly, although market power would exist in the technical sense, the situation should not be viewed as problematic from an antitrust perspective.

Many structural features of markets have been identified as possible entry barriers: economies of scale, learning by doing, reputation, access to capital, customer switching costs, lack of product compatibility, network effects, patent protection, and access to distribution channels. Because the implication of so-called entry barriers depends on the context—and because some degree of market power is sometimes unavoidable yet many are reluctant to state or imply its existence, such as by deeming something to be an entry barrier in a setting where antitrust intervention seems inappropriate—there is no real consensus on how the term "barriers to entry" should be defined or applied in practice.[25] We do not see clear benefits to formulating a canonical definition of the concept. It may be best simply to keep in mind the purpose of such inquiries into the existence of entry barriers: to assess rivals' supply response as an aspect of an inquiry into the existence of market power, noting that market power is often relevant to antitrust liability but not sufficient to establish it. Beyond that, it may be more helpful to defer further analysis until considering specific practices in specific settings.

2.4.3. Conduct

In some situations, one may be able to infer the presence of market power from the challenged conduct itself. If we observe a firm engaging in a practice that could not be profitable unless it enhanced the firm's market power to some certain degree, we may then infer that market power would indeed increase to that degree. For example, if a firm pays large amounts to retailers to agree not to deal with prospective entrants or spends large sums to maintain tariffs, we may infer that these practices create or enhance that firm's market power.[26] If one accepts the premise that a firm's expertise in assessing its own market power is likely to be more reliable than that produced by a battle of experts before an agency or in litigation, then the firm's own conduct may be a sound basis for inferring the existence of market power.

[25] See Carlton (2004), McAfee, Mialon, and Williams (2004), and Schmalensee (2004) for recent discussions of how to apply the concept of entry barriers in antitrust analysis.

[26] As we discuss in subsection 5.4.2 in our analysis of exclusive dealing contracts with retailers, we would need to rule out pro-competitive justifications, such as those based on free riding. In the case of lobbying to erect tariff barriers, even if the conduct enhances market power, it would not violate U.S. antitrust laws because petitioning government, even to restrict competition, is exempt activity under the *Noerr-Pennington* doctrine.

Two caveats should be noted. First, the amount of market power that may be inferred will sometimes not be very great. A firm with billions of dollars in sales would happily spend millions lobbying for tariffs even if the resulting degree of market power were trivial. On the other hand, if a firm engages in a plan of below-cost pricing that sacrifices hundreds of millions of dollars in current profits, in the absence of other explanations one might well infer that it anticipates a substantial degree of market power, at least sufficient to recoup its investment.

Second, the reliability of the inference depends greatly on the lack of ambiguity regarding the character of the practice under consideration. If one is certain that the conduct would only be undertaken if it could enhance the firm's market power (to some requisite degree), then the inference is sound. However, often it will be contested whether the conduct in question was designed to and will have the effect of increasing market power rather than constituting a benign or even beneficial practice that increases welfare. For example, prices below cost may be profitable because they are predatory, or because they are introductory offers that will enhance future demand for an experience good, or because they stimulate the demand for other products sold by the firm at a healthy price-cost margin. If pro-competitive explanations are sufficiently plausible, no inference of market power is warranted, at least without further investigation.

Recognizing the possibility that the conduct at issue may be pro-competitive is especially important given the role that market power requirements often play in antitrust, namely, as a screening device. That is, we may require a plaintiff to prove the existence of market power because we do not want to subject a wide range of behavior to the costs of antitrust scrutiny and the possibility of erroneous liability. When the conduct that provides the basis for inferring market power is the very same conduct under scrutiny, and furthermore when the purpose and effect of such conduct is ambiguous, permitting an inference of market power from the conduct somewhat undermines the screening function of the market power threshold. This concern may be especially great when juries serve as the finders of fact.[27]

2.5. *Market power in antitrust law*

As noted, in antitrust law the notion of market power is frequently used as a screen: a firm (or group of firms) must be shown to have some level of market power as a prerequisite to considering whether the conduct in question gives rise to antitrust liability. As a result, antitrust investigations and adjudications devote substantial attention

[27] This concern may help to explain the Supreme Court's decision in Spectrum Sports v. McQuillan, 506 U.S. 447 (1993), where the Court held in an attempted monopolization case that the plaintiff had to meet the market power requirement independently of proving predatory conduct. Although the holding on its face seems illogical (if, as the plaintiff argued, it would have been irrational to have engaged in the conduct unless the requisite contribution to market power were present), the actual practice under consideration may well have appeared to the Court to be nonpredatory, so it wished to heighten the plaintiff's required proof before it would allow the case to be considered by the jury.

to whether or not the requisite market power exists. In rhetoric and often in reality, this legal approach of viewing market power as something either present or absent—a dichotomous classification—is at odds with the technical economic notion of market power as a matter of degree. Because some degree of technical market power is ubiquitous, it is evident that the term "market power" as used in antitrust law has another meaning. Nevertheless, the law's notion of market power is quite closely related to that of economists. A legal finding of market power constitutes not merely a declaration of the existence of technical market power, however trivial, but rather a conclusion that the degree of existing or increased market power exceeds some threshold, a benchmark that, as we will see, varies with the type of conduct under consideration and that in most instances is not clearly specified.

This feature of antitrust law's use of a market power requirement is well illustrated by the law of monopolization. As will be elaborated in subsection 5.2, under U.S. antitrust law "[t]he offense of monopoly ... has two elements: (1) the possession of monopoly power in the relevant market and (2) the willful acquisition or maintenance of that power as distinguished from growth or development as a consequence of a superior product, business acumen, or historic accident."[28] The requirement of "monopoly power" is conclusory in that it merely signifies that degree of market power deemed minimally necessary and also sufficient to satisfy the first element of the offense of monopolization. It is understood that this level of market power is higher than that required in other areas of antitrust law. Notably, the market power requirement is highest in monopolization cases, somewhat lower in attempted monopolization cases, and lower still in horizontal merger cases, as will be discussed in subsections 4.4.2 and 5.2.1. However, these requirements typically are not stated quantitatively, making it difficult to know very precisely what is the threshold in any of these areas.

In principle, the fact that market power is a matter of degree should be recognized in designing antitrust rules. A monopolistic act that is unambiguously undesirable might be condemned even if the incremental impact on market power is modest, whereas for conduct that is ambiguous, with a high risk of false positives, it may be appropriate to contemplate condemnation only when the potential effect on market power is substantial. If one were minimizing a loss function in which there was uncertainty about the practices under scrutiny, and if the degree of harm conditional on the practices being detrimental was rising with the extent of market power, an optimal rule could be stated as entailing a market power requirement that was highly contextual.

For practical use by agencies and in adjudication, however, a more simplified formulation may economize on administrative costs, provide clearer guidance to parties, and reflect the limited expertise of the pertinent decision-makers. Nevertheless, some greater flexibility may be warranted and is indeed increasingly reflected in antitrust doctrine. The early emergence of a *per se* rule against price-fixing, which dispenses with proof of market power, is one illustration. Another is the increasing use of intermediate levels of scrutiny under the rule of reason (see subsection 3.5.2) and the implicit reliance

[28] United States v. Grinnell Corp., 384 U.S. 563, 570–71 (1966).

on different market power thresholds by the antitrust agencies in reviewing horizontal mergers in different industries, despite the existence of official guidelines that purport to be uniform.

In addition to differences in the magnitude of market power thresholds and whether there is some flexibility regarding the requisite degree of market power, there is variation across contexts in whether the question posed concerns the extant level of market power or the amount by which the actions under scrutiny would increase it. In a monopolization case, the standard question is often whether a firm's past practices have improperly created or maintained monopoly power, so the inquiry is usually into whether significant market power already exists, as reflected in the previously quoted formulation. By contrast, in examining horizontal mergers, the focus is on whether the proposed acquisition would significantly increase market power.[29]

We believe that this distinction is overstated and potentially misleading and that the correct inquiry should focus largely on contributions to market power. Even in the typical monopolization case, the relevant question is how much the past practices contributed to the existing situation. If the contribution is large (and if the practices are not otherwise justifiable), it seems that there should be a finding of liability even if the resulting total degree of power is not overwhelming. (In such a case, the initial level of market power presumably will have been rather low.) Likewise, even if the degree of existing market power is great, in cases in which the practices in question did not plausibly contribute significantly to that result, one should be cautious in condemning those practices, that is, they should be condemned only if they are unambiguously undesirable.

As an example, consider a firm selling a relatively homogeneous product, such as in the chemical industry, that enjoys a significant cost advantage over its rivals based on patented process technology. That firm might well enjoy a nontrivial degree of technical market power. Neither good sense nor existing law ordinarily condemns the discovery of a superior production process. Let us assume that the firm's technical market power was legally obtained and suppose further that the firm prices against a perfectly elastic rival supply at some trigger price that is below the firm's monopoly price. Antitrust issues could arise if this firm attempts to acquire its rivals or if the firm engages in conduct that drives its rivals out of business. In considering such cases, the degree of the firm's initial market power is of secondary importance (although if it were near zero, further inquiry would probably be pointless). Instead, the central question should be whether and to what extent the acquisition or exclusionary conduct will augment that firm's market power and thus harm consumers. For example, however great is the initial level of market power, the firm would gain no additional power by acquiring (or destroying) one of its rivals as long as numerous others that remain still have highly elastic supply at that same trigger price. However, the firm might well gain market power by acquiring or destroying a rival with uniquely low costs, thereby raising the price at

[29] See subsection 4.4.2, where we discuss the point that the extant level of market power is also important.

which substantial competing supply would be triggered. We return to the question of the relevance of extant market power versus challenged practices' contribution to power in subsection 5.1.2 with regard to monopolization and exclusionary practices.

A further possible deviation between economic analysis and antitrust law with regard to market power concerns the benchmark against which the height of price-cost margins is assessed. The U.S. antitrust enforcement agencies in the Horizontal Merger Guidelines (1992) define market power as "the ability profitably to maintain prices above competitive levels for a significant period of time," and the Supreme Court has similarly stated that "As an economic matter, market power exists whenever prices can be raised above the levels that would be charged in a competitive market."[30] If one understands the competitive price to refer to the price that would be charged in a hypothetical, textbook, perfectly competitive market in which firms have constant marginal costs equal to the marginal cost of the firm in question at the prevailing equilibrium, then the legal and economic concepts are essentially the same. However, the hypothetical competitive scenario that underlies such statements is rather vague: the counterfactual is not explicit, and some specifications that may implicitly be contemplated may not yield sensible answers. For example, what is meant by the perfectly competitive price in a market with fixed costs?

Courts have struggled with these issues for many years. The Supreme Court has stated that "Monopoly power is the power to control prices or exclude competition."[31] This is not a meaningful screen, however, since any firm with technical market power has some ability to control prices. Conversely, in the European Union, the European Court of Justice has said that a "dominant position" corresponds to "a position of economic strength enjoyed by an undertaking which enables it to hinder the maintenance of effective competition on the relevant market by allowing it to behave to an appreciable extent independently of its competitors and customers and ultimately of consumers."[32] This test is not especially useful either, since even a firm with great market power does not rationally behave independently of its competitors or customers. That is, there is some monopoly price, P_M, which—however high it may be—implies that a price of, say, $2P_M$ would be less profitable due to far greater consumer substitution away from the product at that higher price.

3. Collusion

We now turn to collusion, including price-fixing cartels and other arrangements that may have similar effects, such as the allocation of customers or territories to different suppli-

[30] Jefferson Parish Hospital District No. 2 v. Hyde, 466 U.S. 2, 27 n. 46 (1984).
[31] U.S. v. E.I. du Pont de Nemours & Co, 351 U.S. 377, 391 (1956).
[32] Case 322/81, Michelin v. Commission [1983] ECR 3461 §30.

ers.[33] For concreteness, we will ordinarily focus on price-fixing. There is an enormous literature on the economics of collusion that we do not attempt to review systematically. Existing surveys include Shapiro (1989), Jacquemin and Slade (1989), Motta (2004, ch. 4), and of particular note Whinston (2006, ch. 2). For in-depth discussion of some especially interesting price-fixing cases, see Borenstein (2004), Connor (2004), Elzinga and Mills (2004), Motta (2004, pp. 211–219), and Porter and Zona (2004).

The focus here, as in the rest of this survey, is on the intersection of economics and the law. We begin by noting the core elements from each field and posing questions about their relationship. Next, we explore the economics of collusion, focusing on the necessary elements for successful collusion, lessons from game-theoretic models of oligopoly, and the various factors that bear on the likelihood of successful collusion. Finally, we examine legal prohibitions in light of the basic teachings of economics.

3.1. Economic and legal approaches: an introduction

3.1.1. Economic approach

For as long as there has been commercial competition, rivals have been tempted to short-circuit it because self-interest favors their own profits at the expense of customers' interest in lower prices and the overall social interest in allocative efficiency. No less a champion of the free-market system than Adam Smith ([1776] 1970, bk 1, ch. X) considered collusion an ever-present danger. "People of the same trade seldom meet together, even for merriment and diversion, but the conversation ends in a conspiracy against the public, or in some contrivance to raise prices." If one thinks in terms of a homogeneous product, firms seek to establish and maintain the monopoly price, which exceeds the price that would prevail in the absence of the agreement. With differentiated products or price discrimination, although there is no single monopoly price, the same idea applies: firms seek to elevate prices and thus raise their collective profits at the expense of consumers. In so doing, the firms typically increase the gap between price(s) and marginal cost(s) and thus raise deadweight loss and lower total welfare, defined as the sum of supplier profits and consumer surplus. Thus, collusion is unwelcome, whether one is seeking to maximize overall efficiency or consumer welfare.

Colluding firms use a variety of methods to achieve the basic goal of raising prices. In some cases, firms agree to minimum prices. In others, they agree to limit their production levels, since output restrictions translate into elevated prices. Alternatively, firms can allocate customers or territories among themselves, with each firm agreeing not to compete for customers, or in territories, assigned to others. These customer and territorial allocation schemes effectively grant each firm a monopoly over some portion of

[33] We do not explicitly address the full range of "horizontal agreements," which includes group boycotts as well as arrangements among buyers, notably, to suppress the prices of inputs, the latter of which are subject to similar analysis as that presented here, although they have received less antitrust scrutiny.

the overall market, so they lead to higher prices and reduced output, even though these schemes do not directly specify price or output.

Economists studying collusion, and more generally oligopoly, tend to inquire into the factors that determine the market equilibrium outcome in an industry. Economists typically focus on whether the outcome is relatively competitive, with prices close to marginal cost, or at least some measure of average cost, or relatively collusive, with prices close to the level that a monopolist would pick to maximize industry profits. This approach is consistent with economists' traditional emphasis on market outcomes and their implications for the allocation of resources.

This approach focuses on description or prediction, not on policy prescriptions regarding how the government should mitigate the costs of collusion. There is a general consensus that clearly identifiable attempts to engage in collusive behavior should be prohibited, so explicit cartel agreements should not be legally enforceable and private attempts to agree upon and enforce supra-competitive prices should be punished. It is widely recognized, however, that it is not always possible to determine whether collusion is occurring or, even when it is, which specific practices should be proscribed. One approach in such settings would be price regulation, which is often undertaken in the case of natural monopolies but is generally thought to be inferior to decentralized competition when such is feasible. In the past, there have been recommendations to deconcentrate certain industries in order to achieve more competitive outcomes.[34] Such proposals have not been implemented, except in some cases of monopolization, and have not of late been actively considered in the United States. Another structural approach is more prevalent: enjoining horizontal mergers that make collusive outcomes more likely, a topic we explore in section 4. Finally, for cases in which collusion can be identified but the specific practices enabling it cannot, Posner (1969, 2001) interestingly proposes the imposition of monetary penalties on oligopolists if the market equilibrium outcome is collusive.[35] The idea is that, just as Pigouvian taxes induce firms to refrain from inefficient behavior, the details of which might be difficult for a regulator to observe or proscribe, so too would appropriate fines or damage awards in private litigation lead firms to abstain from collusive behavior. This approach assumes, importantly, that it is possible to measure the extent to which prices exceed non-collusive levels, which

[34] Legislation was introduced repeatedly in the early 1970s that would have authorized the dissolution of firms in concentrated industries that had not engaged in substantial price competition over three consecutive years; see, for example, S.1167, March 12, 1972. This legislation was based on the White House Task Force Report on Antitrust Policy (1968), commonly known as the Neal Report.

[35] On its face, present practice appears to differ significantly from Posner's proposal. Although the United States and most other competition enforcement regimes do provide for fines or private damage remedies in cases of price-fixing, to trigger such payment obligations, the government or private parties typically need to show that in some sense there is an "agreement." Furthermore, satisfaction of this requirement is generally understood to entail more than demonstrating that the observed outcome involves a "collusive price," although as we shall discuss, just how much more must be shown and what constitutes an adequate demonstration is unclear. See subsections 3.1.2 and 3.4.

poses both conceptual and practical challenges of a sort that are encountered in imposing conventional price-fixing sanctions, the magnitude of which depends on the extent of collusive overcharges. However, Posner's approach has not been embraced by the courts.

3.1.2. Legal approach

The legal approach to collusion, at least on its face, differs from the economic approach. As just described, the economic approach begins with a diagnosis of the problem, then tries to ascertain whether and when collusion occurs, and finally assesses the efficacy of competing remedies. Although one would like to believe that the legal approach is at some level grounded in such analysis, on the surface it appears to focus instead on particular behavioral elements. As will be seen in the course of our analysis in section 3, the extent to which the legal approach can ultimately be rationalized on economic grounds depends on how legal tests are interpreted.

In the United States, the European Union, and many other jurisdictions, the structure of legal prohibitions revolves around the distinction between unilateral and group behavior. Unilateral behavior is circumscribed to a limited degree by anti-monopolization law (see section 5) and various other provisions but is not subject to a regime of price regulation or other forms of internal micro-management of firms' dealings.[36] Thus, firms are purportedly free to set prices and other conditions of trade.[37] This freedom, however, is restricted to unilateral behavior. Independent firms are expected to compete, conferring the benefits of competition on consumers and on society as a whole.

The central legal question with which we will be concerned—and will elaborate in subsection 3.4—is how courts or other regulators are to determine when supposedly competing firms are instead conspiring. Legal prohibitions are typically triggered by certain types of conduct rather than by outcomes themselves. For concreteness, we will discuss the prohibition in U.S. antitrust law, Sherman Act §1, which makes illegal "[e]very contract, combination . . . , or conspiracy, in restraint of trade." In practice, the standard term of art is "agreement," even though that term does not appear in the statute.[38] Thus, the legal question is whether firms' pricing is the result of an agreement. If not, there is no violation. If so, there is a violation, and penalties in the United States include having to pay treble damages to injured customers, being subject to injunctions on prohibited behavior, and criminal penalties, under which firms' executives convicted of price-fixing serve prison terms and firms pay fines.

[36] It is true, however, that remedies in monopolization cases and some others can entail what is tantamount to fairly detailed regulation.

[37] There are important qualifications, notably with regard to proscriptions on predatory pricing (see subsection 5.3), but the focus in this section is on prices that are too high and thus directly harm customers rather than on prices that are too low and thus directly harm competitors.

[38] Interestingly, this is the language of Article 81 of the competition law in the European Union. Like in the United States, the concept in the European Union embraces more than formal contracts yet it is uncertain just how much more.

What, then, is an agreement? And how does this concept relate to the economic analysis of collusive behavior? These questions will occupy much of the remainder of section 3 of our survey. To provide some guidance in the interim, a few preliminary observations are offered. First, there are clear cases. At one extreme, if competitors meet in the proverbial smoke-filled room, negotiate a detailed cartel arrangement, sign it, and implement it—and, importantly, this all can be proved in a legal proceeding—an agreement and hence a legal violation will undoubtedly be found to exist. At the other extreme, no agreement would presumably exist and no violation would be found due to the mere fact that competitors' prices are equal—as one expects with homogeneous products and perfect competition, for example—or that they sometimes move together—as tends to occur when there are shocks to input prices (think of retail gasoline stations changing sale prices when prices from refineries change).

The difficult cases fall at various points in between, in terms of what actually transpired and what can be proved before a tribunal. Consider a simple example. Suppose that just two firms, A and B, supply a particular product. Let the monopoly price of that product be $100 and the fully competitive price (that is, the price at which the industry marginal cost curve crosses the demand curve) be only $40. Suppose further that the actual industry price persists at $100, with sales split evenly between the two firms. This is clearly a collusive outcome, but have the firms entered into an agreement in restraint of trade?

As noted, if a written agreement, negotiated and signed in a smoke-filled room, is produced as evidence, a violation will be found. Suppose that no such agreement is directly in evidence. One possibility is that such an agreement nevertheless exists, and if a tribunal can be convinced of this by circumstantial evidence, a case will have been made. But what sort of evidence would be necessary to make this inference? The answer depends importantly on the competing hypotheses—and on which alternative explanations are likewise deemed to involve agreements and hence would also constitute violations. What interactions short of a meeting in a smoke-filled room that results in a written document will suffice? Is a face-to-face meeting required or would a conference call or an e-mail exchange be enough? What about other forms of communication, such as statements relayed through third parties or in various codes? Or nonverbal communication (hand signals, winks and nods, posting signs with proposed prices, and so forth)? Must there be a written document? Presumably not. Must there be a formal agreement tantamount to a legally enforceable contract? Well, since a contract would not be legally enforceable in any event, presumably this too is not required.

In sum, we can be certain that agreements may be deemed to exist when something well short of the formal meeting and written cartel document exists. But it is not clear how much less will give rise to liability, or, put in the affirmative, just what is (are) the core underlying element(s) of an "agreement." For now, we will leave this question, returning to it in subsection 3.4, after we have surveyed some key aspects of the pertinent economic theory, which one might hope would illuminate the legal inquiry.

3.2. Oligopoly theory

3.2.1. Elements of successful collusion

Economists have long recognized that there exist certain prerequisites to successful collusion. The classic modern reference is Stigler (1964). Green and Porter (1984) embed these issues in a supergame context. The key elements are (1) *reaching consensus*: some understanding must be reached among the otherwise-competing firms regarding what conduct is permitted under the terms of the collusive agreement, such as the prices that the firms will charge; (2) *detection*: some reliable means must exist by which departures from the agreement can be detected; and (3) *punishment*: some credible mechanism must be established by which such departures are punished if and when they are detected. Specifically, the prospect of detection and punishment must be sufficient to deter individual firms' proclivity to cheat on the agreement, typically by cutting prices in the short-term, hoping to reap greater profits through a higher market share at the expense of the other firms, before they can respond. Related to the need to reach an agreement is the problem of (4) *inclusion*: a means of inducing participation by a sufficiently large number of incumbent suppliers so that competition from non-participants does not undermine the profitability of the collusive agreement. Lastly and relatedly, the incumbent firms must be protected by (5) *entry barriers*: there must not be so much competition from quickly arriving new entrants so as to undermine the effectiveness of collusion.

Some economists consider these requirements to be so daunting that cartels are either unable to form or quick to collapse, even in the absence of antitrust laws designed to stop collusion. For example, when OPEC first arose, some confidently predicted its immediate demise. However, the experience with OPEC and empirical evidence on price-fixing more broadly does not support this optimistic view. For example, in the past decade, the Antitrust Division of the U.S. Department of Justice has broken up many large, international cartels that had operated for years (despite the fact that they were illegal under the antitrust laws) and successfully reaped hundreds of millions if not billions of dollars in profits.[39] For recent extensive surveys of the evidence of collusive activity, see Connor (2007), Harrington (2006b), and Levenstein and Suslow (2006). Whinston (2006, pp. 26–38) offers a more selective discussion of the empirical evidence regarding the effects of price-fixing conspiracies.

[39] See Litan and Shapiro (2002) for a discussion of U.S. cartel enforcement activities during the 1990s. The Antitrust Division of the Department of Justice regularly announces enforcement actions against cartels. For example, as of September 2006, some $731 million of fines had resulted from the Division's investigation into price-fixing of DRAM (Dynamic Random Access Memory). See "Samsung Executive Agrees to Plead Guilty, Serve Jail Time, for Participating in DRAM Price-Fixing Conspiracy," http://www.usdoj.gov/atr/public/press_releases/2006/218462.htm.

3.2.2. Repeated oligopoly games and the folk theorem

The basic theoretical framework used to evaluate the presence, absence, or efficacy of collusion is that of dynamic or repeated oligopoly, that is, situations in which an identifiable group of suppliers offering substitute products interact over time.[40] This framework includes infinitely repeated oligopoly games, so-called supergames. Cartel theory requires dynamic analysis because the central elements of detection and punishment inherently take place over time.

One of the central findings in the theory of oligopoly supergames is that there are many—indeed, infinitely many—noncooperative equilibrium outcomes, including outcomes that maximize the joint profits of the oligopolists, even if one restricts attention to subgame-perfect Nash equilibria.[41] To give a flavor for why there are so many equilibria in supergames, consider a game in which each of N firms, selling the same homogeneous good and incurring a constant cost per unit of C, sets its price every period. Suppose that the stage game is a classic Bertrand pricing game, so the firm with the lowest price in a given period serves the entire market in that period, and if multiple firms charge the same lowest price output is shared equally among them. (One can also think of the firms bidding each period to serve the single large customer who buys that period.) For the moment, suppose also that, as soon as a given period ends, each firm immediately observes all the prices charged by the other firms during that period.

In this simple repeated Bertrand game, the competitive outcome, by which we mean the Nash equilibrium in the one-shot (stage) game, involves each firm setting its price equal to cost, C, each period. This would be the only noncooperative equilibrium outcome if this pricing game were played only a single time, or indeed any finite number of times (a familiar consequence of backward induction). However, if the game continues indefinitely, it is easy to construct a subgame-perfect equilibrium in which the price each period is $P > C$, for many different values of P ranging from the competitive price all the way up to the monopoly price (or, in fact, higher). The trick is to postulate that, should any firm ever charge below P, all the other firms will set a price of C in all subsequent periods. This punishment strategy thus entails reversion to the one-shot Nash equilibrium. Because this behavior supports an equilibrium price of P every period, the equilibrium profit of each firm is $\pi(P)/N$ each period, where $\pi(P)$ denotes the profits earned by a firm setting price P and serving the entire market. Adding up over all periods and discounting using the one-period discount factor δ, the equilibrium profits of each firm are $\pi(P)/N(1-\delta)$. In contrast, in any given period a single firm could defect by slightly discounting its price, thereby capturing the entire market that

[40] There is a very large literature on oligopoly theory and supergames. Since our focus in on antitrust, we will draw on this literature but not review it in any systematic way. See Shapiro (1989) for a survey of oligopoly theory. Ivaldi et al. (2003a), Motta (2004, ch. 4), and Whinston (2006, ch. 2) provide recent discussions of the application of oligopoly theory to collusion cases in antitrust.

[41] Friedman (1971) showed that the full cartel outcome can be supported in a repeated oligopoly if the players are sufficiently patient.

period. This would, by hypothesis, condemn all of the firms—importantly, including itself—to the competitive outcome for all future time. The payoff to this defecting firm is $\pi(P)$ in the current period and zero in all subsequent periods. Defecting is therefore unattractive if and only if $\pi(P)/N(1-\delta) > \pi(P)$, that is, if and only if $\delta > 1-1/N$. If the periods are sufficiently short, that is, if price cuts are detected rapidly enough, then δ is close to unity and this condition is met, even when N is large.

The frequency of sales, and hence the speed of detection, is implicitly built into this simple model. As expected, the faster that rivals can learn of one firm's defection and respond by reducing their own prices, the easier it is to sustain collusion. Formally, the model assumes that price cuts are observed after one period, but the length of time comprising one period has not been specified. If one period takes time T, and if the interest rate per unit of time is r, then the discount factor is $\delta = e^{-rT}$. Therefore, a longer detection lag corresponds to a lower discount factor, making it less likely that collusion will be sustainable, *ceteris paribus*.

This highly simplified example illustrates that there exists a subgame-perfect equilibrium in which all the firms charge the same price, P, repeatedly, for any $P > C$, as long as detection is sufficiently rapid. For example, if the length of a period is one month, and we imagine that the interest rate is around 1% per month, then the discount factor δ is roughly 0.99, so the monopoly price can be sustained as a perfect equilibrium in the repeated oligopoly game as long as the number of firms is fewer than one hundred.

While the specific calculations here depend upon the particular oligopoly game being played each period, the basic idea—that for plausible discount factors there exist perfect equilibria with high (well above competitive) prices even with many firms—is by no means specific to this example. Consider, for example, Cournot oligopoly with a homogeneous product. On one hand, a firm that defects cannot gain nearly as much as in the Bertrand game. (The ratio of profits from defecting to the per-period profits from following the equilibrium strategy depends upon the shape of the demand curve.) On the other hand, reverting to the static Cournot equilibrium is not as severe a punishment as reverting to the static Bertrand equilibrium. Shapiro (1989) shows that, with $\delta = 0.99$, the monopoly price can be sustained in a repeated Cournot oligopoly with constant marginal cost and linear demand as long as there are no more than four hundred firms.

Furthermore—one might say "worse yet"—this example is hardly anomalous. To the contrary, the "folk theorem" for repeated games tells us that, quite generally, there exists a plethora of equilibria in repeated games, including equilibria that correspond to full cooperation (the profit maximizing cartel outcome), as long as the players are sufficiently patient, that is, if the discount factor is sufficiently close to unity. Fudenberg and Maskin (1986) provide general conditions under which any feasible and individually rational payoff vector in the stage game can be achieved as a perfect equilibrium in the repeated game if players are sufficiently patient.[42] Fudenberg, Levine, and Maskin

[42] Abreu, Pearce, and Stacchetti (1986, 1990) consider punishments stronger than simply reverting to the one-shot Nash equilibrium. In the simple repeated Bertrand game, no punishment can be stronger than reversion

(1994) extend this result to games in which the players do not observe each others' actions but only a public outcome (such as price) that signals those actions.

These folk theorems pose two related and fundamental challenges for the analysis of oligopoly. First, these strong results do not comport with the observation that rivals often compete rather than cooperate and in particular with the prevailing view that collusion is difficult if the number of firms is moderately large—a view that Farrell (2000) calls the "structural consensus." Clearly, some important things are missing in models of repeated oligopoly that predict collusive outcomes for a very wide range of market structures and industry conditions. Put simply, supergame theory, at least with sufficiently patient players and without further modifications, is of limited use because it proves too much. The economics literature has responded to this criticism by probing the assumptions underlying the folk theorem and by exploring equilibrium outcomes when the relevant discount factor is not very close to unity, that is, when the players are not "sufficiently patient." In particular, a large literature, some of which we examine in subsection 3.3, explores the conditions that make it more or less difficult to support collusive outcomes in repeated oligopoly.

Second, and of particular relevance to antitrust law, there is no explicit role for communications in the basic models of repeated games, so these models do not help us understand the impact of meetings and other communications among oligopolists. Some research, however, does explore aspects of this limitation, as we consider next.

3.2.3. Role of communications

One of the specific shortcomings of standard models of oligopolistic supergames is that they do not help us understand how the firms initially determine which of the plethora of equilibria to play. One interpretation of these games is that the firms engage in extensive communications and perhaps negotiations *before* the game begins, in order to agree upon the equilibrium that they will play. Under this interpretation, the equilibria in oligopolistic supergames represent self-enforcing outcomes that can arise once an agreement is reached. The alternative explanation of the observed conduct is that the firms somehow find their way to a relatively collusive outcome without engaging in any communications other than through their actions in the market, such as their setting of prices.

Ambiguity about the role of communications is inherent in the standard solution concept of Nash equilibrium (and thus perfect equilibrium, a refinement of Nash equilibrium). When the strategies are highly complex, and especially when there are multiple equilibria, the perfect equilibrium (or Nash equilibrium) methodology does not explain

to the one-shot Bertrand equilibrium, since it involves zero profits, which by definition is the level of profits a firm can obtain by exiting. For the repeated Cournot game, however, stronger punishments are possible by targeting the defecting firm and punishing again a firm that refuses to go along properly with its own punishment. When available, these stronger punishments make it possible to sustain the monopoly price even if the firms are somewhat less patient.

how the firms were able to coordinate to select an equilibrium. Yet common sense indicates that communications can play a role in such coordination, and complex strategies supporting a collusive outcome would seem to constitute evidence in favor of the hypothesis that the firms in fact met and reached some sort of agreement at some point in time.

The literature on "cheap talk" asks whether communications affect the equilibrium outcome in a game-theoretic setting. Farrell and Rabin (1996) provide an excellent overview of this literature.[43] In general, cheap talk, that is, communications that do not directly affect payoffs, can affect equilibrium outcomes.[44] Farrell and Rabin give the example of one Cournot duopolist telling another: "You cut your output and I'll cut mine." While this might be a trick—the speaker gains if the listener cuts output, whether or not the speaker does so—this also might be an effective way of initiating output reductions that sustain collusion in repeated play. In the end, Farrell and Rabin conclude that cheap talk about intended play can make a big difference when the players' interests are well aligned, but the gains available from coordination can easily be lost due to dispute and bargaining problems.

A different strand of the literature studies the role of communications to convey private information. Athey and Bagwell (2001, 2006) permit the firms to communicate about private cost information in a repeated pricing game. Compte (1998) and Kandori and Matsushima (1998) study communications when firms observe private but imperfect signals about past play. Athey, Bagwell, and Sanchirico (2004) study a model with private cost shocks and publicly observed prices.

Facing the rather ambiguous theoretical results in the "cheap talk" literature reported by Farrell and Rabin (1996), economists have conducted experiments to learn how communications affect the strategies adopted by players. This literature is surveyed by Crawford (1998). As an early example, Cooper et al. (1989) find that in the "battle of the sexes game," where the players can gain from cooperation but do not agree about which outcome is best, communications greatly increase the chance that the players will successfully coordinate. Kühn (2001) discusses the antitrust implications of this experimental literature, emphasizing the role of communications in achieving coordination by reducing uncertainty about what other firms will do.

The critical role of communications in sustaining collusion is revealed in the fascinating study by Genesove and Mullin (2001) of the Sugar Institute, a trade association that operated from 1927 to 1936. They examine in detail how sugar refiners established a set of rules to facilitate collusion. The Sugar Institute experience shows how weekly meetings among sugar refiners were used to establish and interpret rules that enforced business practices making price-cutting more transparent. In contrast to the theories

[43] Much earlier, Schelling (1960) recognized the importance of communications and discussed the role of "focal points" in coordinating outcomes in strategic situations.

[44] Communications also may have no effect. There always exists a "babbling" equilibrium in which players ignore the statements made by others. As emphasized by Farrell and Rabin (1996), however, many of these equilibria are implausible.

described above, which involve no cheating in equilibrium, cheating did occur, but retaliation was carefully limited. This example illustrates a number of functions served by regular communications, functions that could not be served simply by initial communications. Related lessons can be found in Harrington (2006b), who reports on some twenty European Commission cartel cases from 2000 to 2004.

3.3. Industry conditions bearing on the likelihood of collusive outcomes

As noted, to get beyond the folk theorem, the theoretical literature on oligopoly has extensively explored the conditions under which the joint profit-maximizing outcome can be achieved as a perfect equilibrium in a repeated oligopoly game when the discount factor is *not* close to unity. One way this question is posed is to ask how various factors affect the critical discount factor, δ^*, such that the fully collusive outcome can be supported as a perfect equilibrium for $\delta \geq \delta^*$.[45] While a survey of the enormous literature on oligopoly theory is well beyond the scope of this chapter, we mention here selected results that are especially relevant to antitrust. We also examine a number of other factors, not strictly part of the standard oligopoly supergame framework, that bear on the feasibility of collusion.

In the first three subsections, we relax several extreme and unrealistic assumptions made in the simple model used above. First, the simple model assumed that a defecting firm could capture the entire market with even a slight price cut. Second, it assumed that even a tiny price cut would surely be observed by rivals. Third, it assumed that even the slightest defection would be punished severely, with all firms pricing at marginal cost in perpetuity, leading to zero profits for all firms. We then consider a variety of other factors that make it more or less difficult for collusive outcomes to be sustained as perfect equilibria in repeated oligopoly.

3.3.1. Limited growth for defecting firm

There are many reasons why a firm that defects from collusive prices may not be able to capture the entire market, including upward sloping marginal cost (in the limit, capacity constraints), customer loyalty, customer switching costs, and product differentiation. Clearly, if the gains from defection are limited, collusion will be easier to sustain, *ceteris paribus*.

To illustrate this idea in the simple model introduced above, suppose that capacity constraints only permit a single firm to grow its sales by a factor $1 + g$ in a single period.[46] In the simple model above, $1 + g = N$, but if a firm can only, say, double in size in one period, then $g = 1$. For now, we retain the assumption that all price

[45] Ivaldi et al. (2003a) take this approach in their accessible and informative overview paper.

[46] A fully specified model would relate this growth limit to underlying economic variables, such as capacity or the degree of product differentiation, and to magnitude of the defecting firm's price cut.

cuts are detected by rivals and punished strongly, leading to zero future profits for the defecting firm. With these assumptions, optimal defection involves a tiny price cut and yields profits of $\pi(P)(1+g)/N$. As a result, charging the collusive price is optimal so long as $\delta > \delta^* = \frac{g}{1+g}$. The smaller is g, the smaller is δ^*, indicating that collusion is easier to sustain. For example, if $g = 1$, then $\delta^* = 0.5$, compared with $\delta^* = 0.9$ in the case where $N = 10$ and a defecting firm could capture the entire market. However, the $\delta^* = 0.5$ calculation makes the assumption that a tiny price cut by one of ten firms, doubling its market share from 10% to 20%, would surely be observable by all of the other firms.

3.3.2. *Imperfect detection*

As emphasized by Stigler (1964), Green and Porter (1984), and much of the subsequent literature, when one firm cuts its price, its rivals may not be able to observe the price cut. Imperfect detection unquestionably makes it more difficult to sustain collusive outcomes. In fact, collusion cases in antitrust law frequently revolve around whether the firms have sufficient ability accurately to observe price cuts so as to enforce a collusive outcome. Additionally, many of the factors considered in subsequent subsections take on importance because they affect the ability of the firms to detect and punish those who defect from a collusive arrangement.

To illustrate the fundamental importance of detection, consider how the calculus of defection changes if there is an exogenous probability, θ, that the price cut is observed by rivals.[47] Retaining our assumption that a single firm can only grow its sales by a factor $1+g$ in a single period, the payoff from cutting price for a single period is $\frac{\pi(P)}{N}(1+g) + (1-\theta)\frac{\pi(P)}{N}\frac{\delta}{1-\delta}$.[48] Collusion is sustainable if and only if this payoff is less than that from maintaining the collusive price, $\frac{\pi(P)}{N}\frac{1}{1-\delta}$. Simplifying, collusion is sustainable if and only if $\delta > \delta^* = \frac{g}{\theta+g}$. This expression captures a basic tradeoff: collusion cannot be sustained if detection is very unlikely (low values of θ), especially if a firm can grow rapidly before detection would take place (large values of g). With $g = 1$ and $\theta = 0.25$, $\delta^* = 0.8$, far higher than $\delta^* = 0.5$ when $\theta = 1$.

Stigler (1964) emphasizes the role of price transparency and secret price cutting. Spence (1978) argues that uncertainty about demand conditions makes it more difficult for suppliers to distinguish shifts in demand from defections by their rivals, and thus makes collusive outcomes more difficult to sustain. Green and Porter (1984) derive trigger strategies when prices are only observed with noise, in which case there is a tradeoff: entering the punishment phase more readily provides a stronger deterrent but can lead to price wars even when no firm has defected. Harrington and Skrzypacz (2007)

[47] Immediately below, we explain how the analysis changes if the probability of detection is a function of the price charged by the price-cutting firm.

[48] This expression assumes, as above, that the firm earns zero profits in the future if its price cut is detected, but also that if the price cut is not detected promptly, it is never observed.

study a model in which sustaining collusion in the presence of demand uncertainty requires asymmetric punishments if the firms observe each other's outputs but not their prices.

In the defection calculations presented so far, there was no reason for a defecting firm to cut its price by more than a small amount since deeper price cuts were not needed for that firm to sell either to the entire market (in our initial model) or to as many customers as the firm can serve given its capacity (in the modified model). The defection calculations are more complex if the firm-specific demand curve is such that a defecting firm's profits are decreasing in its own price at the collusive price, in which case the firm's immediate profits would be higher with a discrete rather than incremental price cut. This variation alone could easily be accommodated using standard pricing theory: the defecting firm would maximize its immediate profits by pricing at the point where marginal revenue and marginal cost are equal.

The analysis becomes more complex, and more interesting, however, if we combine this idea with imperfect detection. More specifically, suppose that deeper price cuts are more likely to be detected by rivals. Then optimal defection involves a tradeoff: lower prices lead to higher profits in the immediate term but a higher probability of detection. The lesson is that the sustainability of collusion may depend in a complex way on the interaction between the ability of a defecting firm to gain customers in the short run and the ability of that firm's rivals to detect that the firm has departed from collusive behavior.

3.3.3. Credibility of punishment

The simple model effectively assumed that it was credible for the oligopolists, as a group, to punish any defection by reverting in perpetuity to the "competitive" outcome, defined as the one-shot equilibrium. In the case of a pricing game with homogeneous products, this implied zero profits for all of the firms. Clearly, this is an extreme assumption, one that seems to dismiss the temptation of the firms to relent on their punishments and try again to achieve a more profitable collusive outcome.

To explore the role of the punishment strategies, consider first the importance of the magnitude of punishment. Suppose that, after one firm defects, instead of prices forever after being set at cost, the other firms respond by merely matching the initial price cut. With this assumption, a firm that cuts price to $P_{cut} < P$ earns total profits of $\pi(P_{cut}) + \frac{\delta}{(1-\delta)}\frac{\pi(P_{cut})}{N}$. How does this expression compare with $\frac{\pi(P)}{N} + \frac{\delta}{(1-\delta)}\frac{\pi(P)}{N}$, the profits from maintaining the collusive price? The profits from defection will exceed the profits from continued collusion for values of P_{cut} near P since cutting price increases the first term by a factor of N and has an effect on the second term that vanishes as $P_{cut} \to P$. After all, the full benefit of defection—capturing the entire monopoly profit—is obtained with a small price cut in the period of defection, and the future punishment is negligible since the price only declines very slightly from the initially collusive price. This analysis illustrates that threats to merely match price cuts are insufficient to maintain collusion.

Second, consider whether the punishment of cutting price to cost in all future periods is credible. One might think that it is because the strategy consists of playing the stage-game Nash equilibrium over and over again, and by construction no single firm would find it optimal to depart from such conduct, given the behavior of the other firms. However, there is something fishy about strategies that call for smooth initial coordination followed by a perpetual price war if any one firm departs from the agreed-upon price, however brief the period of time. In particular, would the firms not be tempted to relent on the price war at some point and return to cooperation? The tension arises because the logic of perfect equilibrium does not consider collective departures from the punishment regime. But ruling out collective departures is hard to defend in a theory that postulates from the outset that the firms can find a way to coordinate to select an equilibrium that is mutually beneficial.

One simple but *ad hoc* way of dealing with this point is to limit the duration of punishments that are allowed after one firm defects by cutting its price below the initially specified level. In the simple supergame above, suppose then that punishments are limited to K periods. With this limitation, the payoff to a single firm from defecting is equal to $\pi(P) + \delta^{K+1}\pi(P)/N(1-\delta)$, where the first term represents the profits during the period when the firm defects and captures the entire market and the second term measures the profits this firm earns once the collusive outcome is restored after K periods of punishment (during which no profits are earned). For collusion to be sustainable, this expression must be no larger than the profits from indefinitely charging the collusive price, P, which as before equal $\pi(P)/N(1-\delta)$. Continuing to use $\delta = 0.99$, for moderate values of K, this condition implies that collusion is sustainable with up to roughly $N \approx K$ firms.[49] So, if we think of one period as corresponding to one month and if we believe that the firms can credibly enter into a price war for one year following a defection, collusion is sustainable so long as there are no more than twelve firms.

A much deeper, but far more complex, way of addressing the credibility of punishments can be found in Bernheim and Ray (1989) and Farrell and Maskin (1989). These papers have proposed refinements of the perfect-equilibrium solution concept that rule out continuation play, including punishment strategies, that is not collectively credible. At the very least, a perfect equilibrium is rejected if the continuation play in any subgame is Pareto dominated by the continuation play in any other subgame. Using the terminology from Farrell and Maskin, who study two-person games, an equilibrium that does *not* contain one subgame that is Pareto dominated by another subgame is called "weakly renegotiation proof" (WRP), a refinement of the subgame-perfect equilibrium concept. In the example above, the continuation game after one firm defected involved

[49] Defection is unattractive so long as $\pi(P) + \delta^{K+1}\pi(P)/N(1-\delta) < \pi(P)/N(1-\delta)$, which is equivalent to $N(1-\delta) + \delta^{K+1} < 1$. For values of δ near 1, $\delta^K \approx 1 - K(1-\delta)$, so this expression is approximately $N(1-\delta) + 1 - (K+1)(1-\delta) < 1$, which can be written as $N - 1 < K$. Intuitively, defecting gives the firm an extra $N-1$ times its profits right away, which is balanced against the loss of those profits for K periods. For $\delta = 0.99$ and moderate values of K, a good approximation for the maximum number of firms is $N \approx K$.

all firms pricing at cost forever. In this subgame, all firms earn a continuation payoff of zero. This subgame is thus Pareto dominated by the subgame consisting of the game itself, in which each firm earns a payoff of $\pi(P)/N(1-\delta)$. Therefore, the perfect equilibrium used above to support the monopoly price is not weakly renegotiation proof, even though it is subgame perfect. All of the firms would prefer to resume cooperating rather than carry out the punishment.

Farrell and Maskin (1989) show that the WRP condition, when applied to repeated-duopoly Bertrand and Cournot games, rules out highly asymmetric equilibria, even if the discount factor is close to unity. However, the WRP condition does not rule out many equilibria involving the monopoly price, including the symmetric equilibrium in which the two firms split the monopoly profits. If asymmetric punishments are specified that favor one firm, that firm will block renegotiations, which WRP required to be a Pareto improvement. Therefore, the WRP concept alone does not successfully resolve the paradox associated with the folk theorem as applied to oligopolies. Farrell (2000) suggests a further refinement, which he calls quasi-symmetric WRP, motivated by the notion that all innocent firms will be treated symmetrically. This concept requires that all innocent firms prefer to carry through with the punishment of a defecting firm rather than revert back to the original equilibrium strategies. Farrell shows that monopoly prices cannot be supported, regardless of the discount factor $\delta < 1$, for moderate numbers of firms in Bertrand or Cournot oligopoly, if this condition must be satisfied. This approach is promising, but relies on the assumption that the firms would find it difficult to establish punishments that treat innocent firms asymmetrically. Further work is required before these ideas can be put to practical use to help identify industry conditions under which collusion is most likely to be effective while accounting for the collective credibility of responses to defections.[50]

3.3.4. Market structure

We now consider a series of factors relating to market structure that affect the incentive and ability of oligopolistic suppliers to sustain a collusive outcome in repeated play.

3.3.4.1. Market concentration Collusive outcomes are less likely to occur in industries with more firms because greater numbers make it more difficult to satisfy the first four conditions necessary for successful collusion. Reaching consensus is harder with more parties involved. Detection is more difficult since price cutting by one small firm

[50] McCutcheon (1997) even suggests that the Sherman Act may *help* firms collude, stating (p. 348): "Government policies that are designed to stop price-fixing may benefit firms by making it worthwhile for them to meet to set up collusive agreements, while making it costly enough for them to avoid undesirable future negotiations." However, this view is not supported by the *Sugar Institute* case reported in Genesove and Mullin (2001). In any event, evidence from the past decade shows clearly that large fines and treble damage awards in price-fixing cases can and do impose substantial penalties on firms engaged in price-fixing, not the weaker sanctions necessary for McCutcheon's logic.

may be very difficult for some or all of the other firms to discern. Punishment is less likely to be effective for two reasons. First, a defecting firm is harder to deter because it has more to gain from cheating: its market-share gain during the period of initial defection is likely to be greater, and its loss from punishment smaller, the more firms were sharing in the collusive profits. Second, punishment may be more difficult to coordinate because of the free-rider problem. Inclusion is also harder due to free-rider problems, as each individual firm may believe that the others will coordinate, whether or not it participates.[51] For precisely these reasons, one of the concerns underlying merger enforcement policy is that mergers between rivals that increase concentration can raise the likelihood that the remaining firms will coordinate after the merger.

Our simple supergame model with homogeneous goods and repeated price competition illustrates how symmetric collusive outcomes are more difficult to sustain when the number of suppliers is larger. We showed above in a very simple model that the monopoly price could be supported in a perfect equilibrium if and only if $\delta > \delta^* = 1 - 1/N$. The larger the number of firms, the larger is δ^*, meaning that the firms must be more patient to sustain the collusive outcome.

Asymmetries in market shares tend to make it more difficult to sustain collusion. For illustrative purposes, suppose that firm i has a market share of s_i.[52] The condition for this firm to cooperate rather than defect is $s_i\pi(P)/(1-\delta) > \pi(P)$, which can be written as $s_i > 1-\delta$. This condition will be most difficult to meet for the firm with the smallest market share, since this firm has the greatest temptation to gain share before the other firms can respond and also the least profits to lose from punishment that renders all firms' profits equal to zero.[53] Defining $s_{\min}$ as the market share of the smallest firm, we get $\delta^* = 1 - s_{\min} > 1 - 1/N$, so the firms must be more patient to sustain the collusive outcome.[54] The smallest firm plays the role of the maverick, that is, the firm most prone to defection from the collusive outcome.

While instructive, this simple model is unable to capture the other factors noted above, which tend to be even more important in practice. Of particular note is the temptation of one relatively small firm to decline to participate in the collusive arrangement or secretly to cut prices to serve, say, 4% rather than 2% of the market. As long as price cuts by a small firm are less likely to be accurately observed or inferred by the other firms than are price cuts by larger firms, the presence of small firms that are capable of expanding significantly is especially disruptive to effective collusion.

[51] This point assumes that, due to capacity limits, rising marginal cost, or other factors, some degree of collusion is possible even if some firms do not participate.

[52] The analysis here is incomplete because it does not explain the underlying sources of the differences in market shares. We address that below when we consider cost asymmetries among the firms.

[53] The analysis assumes that the smallest firm is nevertheless able to capture the market; if, however, capacities are proportional to existing shares, the conclusion may not follow.

[54] Note that, when market shares are equal, $s_{\min} = 1/N$.

3.3.4.2. Cost asymmetries The calculation just given with $s_{\min}$ is incomplete because it does not explain *why* the firms have different market shares. A common explanation is cost asymmetries, and these also make it more difficult for firms to sustain collusive outcomes.[55]

Reaching consensus is clearly more difficult with cost asymmetries since there is less likely to be a focal point for pricing and since the firms may well disagree about the price they would like to see prevail. Furthermore, if collusion is to maximize potential industry profits, production efficiency requires that the low-cost firms be allocated a greater share of sales, but this may require contentious negotiations and/or side payments, thereby limiting somewhat the potential gains from collusion. These problems are exacerbated if cost information is private, since each firm may have an incentive to represent to the others that its costs are low in order to receive a higher allocation of output or other more favorable treatment.[56]

The conventional wisdom states that enforcement of the collusive outcome is also more difficult with cost asymmetries. For example, Ivaldi et al. (2003a) argue that cost asymmetries hinder collusion, stating (p. 36): "even if firms agree on a given collusive price, low-cost firms will again be more difficult to discipline, both because they might gain more from undercutting their rivals and because they have less to fear from a possible retaliation from high-cost firms." Ivaldi et al. show in a simple duopoly example that a higher δ^* applies to the lower-cost firm than to the higher-cost firm, if the firms divide the market equally, because the lower-cost firm earns positive profits in the punishment phase. Assigning a larger share of the market to the lower-cost firm is one way to overcome this obstacle and restore the collusive outcome. However, allocating a lower share to the higher-cost firm necessarily makes it more attractive for that firm to deviate from collusion. Colluding firms thus face a tradeoff: lower-cost firms can be assigned larger market shares, which reduces their incentive to defect, but doing this increases the incentive of the higher-cost firms to defect.

These issues are explored in greater detail in Vasconcelos (2005), who studies repeated quantity competition among firms with heterogeneous quadratic cost functions, where firms differ in their ownership of an underlying asset that lowers the cost function.[57] He shows that, in the optimal collusive equilibrium, output is shifted away from the less efficient firms and toward the more efficient firms. In this equilibrium, the less efficient, smaller firms have the greatest incentive to depart from the collusive outcome, while the more efficient, larger firms have the greatest incentive to depart from the punishments specified by the equilibrium strategies. His results are relevant for the analysis of horizontal mergers since he shows how a merger affects the scope for collusion by changing not only the number of firms but also the distribution of holdings of the underlying asset and thus the distribution of costs among the firms.

[55] See Mason, Phillips, and Nowell (1992) for experimental results showing that cooperation is more likely in a duopoly if the firms have symmetric costs.

[56] Athey and Bagwell (2001, 2006) study repeated oligopoly with private cost information.

[57] See also Rothschild (1999).

Firms also may differ in their cost of capital and hence in the discount rate they use to compare current and future profits. A firm that is under financial pressure, for example, may have a high discount rate (low discount factor) and be especially tempted to defect. Harrington (1989) studied collusion among firms with different discount factors, showing again how market shares must be allocated to support a collusive outcome. As another example, a firm may believe it "deserves" a greater market share than it has historically enjoyed, perhaps because it believes it would greatly increase its market share under competitive conditions. Such maverick firms may be especially disruptive to collusive pricing.

3.3.4.3. Buyer concentration and auction markets Collusion is generally thought to be more difficult to sustain in markets where the buying side is highly concentrated. Apart from the fact that larger buyers may have a more credible threat to vertically integrate upstream than smaller buyers, buyers who purchase a large share of the output of the colluding firms can act strategically and internalize many of the benefits of disrupting collusion. For example, Snyder (1996) shows how a large buyer can strategically accumulate a backlog of unfilled orders to create a bulge in demand that can undermine or destabilize collusion. More generally, a large buyer can strategically create variations in demand over time. For example, by curtailing purchases in one period, the buyer may lead some or all of the suppliers to suspect that others have cheated on the pricing agreement.

Additional strategies are available for a buyer who is setting up the rules by which the suppliers will bid for business. Klemperer (2002) reports enormous variation in the prices received in auctions of third-generation mobile telephone licenses across different European countries, arguing that some auction designs facilitated collusion and thus led to far lower prices being paid for these licenses than was paid for other, comparable licenses. Marshall and Meurer (2004) discuss some of the unique issues that arise when considering collusion in a bidding context, including a discussion of spectrum and timber auctions, arguing that collusion is much more difficult in sealed-bid, first-price auctions than in oral ascending-bid auctions, a point proven more formally in Robinson (1985).

3.3.4.4. Collective market power including entry barriers If the firms have little collective market power, so they collectively face rather elastic demand for their products, their incentive to collude is correspondingly low. Collective market power may be small because the colluding firms are just a subset of the incumbent suppliers, because of low barriers to entry into the sale of the products the firms offer, or because the products they sell face competition from close substitutes sold by other firms. The smaller is the collective market power of the firms that are allegedly colluding, the closer will be the firms' price to the competitive price, and the smaller the damages imposed on consumers by effective collusion.

3.3.4.5. Multi-market contact Multi-market contact refers to situations where firms interact in more than one market at the same time. Much of the literature suggests that multi-market interaction tends to make it easier for the firms to sustain collusion. The standard reference here is Bernheim and Whinston (1990). Bernheim and Whinston first prove an irrelevance result: when identical firms with constant marginal cost meet in identical markets, multi-market contact does not aid in sustaining collusion. While defection in one market can be punished in other markets, a firm can simply defect in all markets simultaneously. However, Bernheim and Whinston go on to show how multi-market contact can sustain collusion in many other settings. For example, multi-market contact can mute market level asymmetries, for example, if each firm has a major competitive advantage in one market (which could include one geographic area of a single product market). Suppose, for example, that Firm A is the leader in Market A and Firm B is the leader in Market B, but both firms compete in both markets. Firm A will be especially tempted to defect in Market B, where Firm A has a smaller share, but may be deterred if Firm B would respond in Market A. Mutual forbearance may well result. Furthermore, multi-market contact increases the frequency of interaction, permitting one firm to discipline another more rapidly than would otherwise be possible.

There is some evidence to support the proposition that multi-market contact makes it easier for firms to sustain collusive outcomes. In the airline industry, Evans and Kessides (1994) find that fares are higher on routes for which the carriers interact on multiple routes. In the mobile telephone industry, Parker and Röller (1997) find higher prices in markets where carriers have multi-market contact. Cramton and Schwartz (2000) look at signaling to support collusion in FCC spectrum auctions, where multiple auctions for licenses were conducted simultaneously.

The value of multi-market contact in sustaining collusive outcomes is less clear, however, once one accounts for the noisiness of the signals that the firms receive regarding possible defections by others. Green and Porter (1984) show that, with noisy signals, limited punishments are optimal; in their model, the firms revert to cooperation after a limited period of time. With multi-market contact, spreading punishment across markets may simply not be desirable, just as engaging in a longer price war, while feasible, may not be optimal in the Green and Porter model. After all, in models where punishments actually occur in equilibrium, stronger punishments are costly. This important idea is absent from the many supergame models in which punishment never takes place in equilibrium. Thus, in more realistic models in which defections and/or punishments actually occur, multi-market contact may have no effect on the ability of the firms to collude. This view is supported by the *Sugar Institute* case described by Genesove and Mullin (2001); the Sugar Institute was very careful to calibrate punishments to the violation and certainly did not employ the maximum possible punishment. Had they done so, the cartel would have collapsed early on. In fact, the Sugar Institute steered away from multi-market linkages, carefully limiting punishment to the same geographic region where the violation occurred.

3.3.5. Product differentiation

The traditional view in antitrust circles has been that collusion is easier to sustain among firms selling homogeneous products rather than highly differentiated products. Reaching consensus should be easier when agreement only requires that one price, not many, be established. However, there is no compelling theoretical reason to believe that detection and punishment are more difficult if the products are more differentiated. On one hand, with highly differentiated products, a single firm that cuts its price is likely to gain relatively few sales, since many customers will still prefer the other brands. Therefore, defecting is less attractive. On the other hand, punishments are weaker, since price cuts by the other firms have a smaller effect on the profits of the defecting firm if products are highly differentiated. Ross (1992) presents two models of oligopolistic supergames with differentiated products that capture these ambiguities.

Another reason that collusion is more difficult to maintain when products are differentiated is that dimensions of competition other than price and cost cutting can take center stage. Collusion along such dimensions as product design and marketing can be very difficult to establish and sustain. Even if an initial agreement is reached, the most tempting way to defect from a collusive agreement may be to improve one's product or to expand one's marketing budget rather than to cut one's price. The firms may find it very difficult to restrain competition on marketing because of the difficulty in drawing the line between permissible and impermissible marketing activities. Likewise, collusion on product design may be hard to sustain due to the difficulties of defining what types of product improvements are permissible and the fact that product improvements include an element of commitment, tempting firms to preempt their rivals to capture more market share on a sustained basis.

3.3.6. Capacity constraints, excess capacity, and investment in capacity

In many industries, certainly including traditional manufacturing, capacity constraints are an important aspect of the competitive environment. In fact, capacity investment decisions can be the most important dimension along which competition occurs in the long run. We now address capacity decisions and their interaction with pricing and output decisions. We begin with a short-run analysis, which takes capacities as fixed, and then go on to the long-run analysis, which includes capacity investment decisions.

3.3.6.1. Collusion on prices with capacity constraints We have already observed that collusion on prices is easier to sustain if any single firm could only gain limited sales by cutting its price. As we noted, one reason a firm may not be able to increase its sales much by cutting price is that the firm may face capacity constraints.[58] Therefore,

[58] A defecting firm might be able to relax this constraint by building inventories in anticipation of cutting its price.

it would appear that collusion on prices is very easy to sustain if the firms all have little excess capacity.

This argument, however, is seriously incomplete. Most fundamentally, if all firms are producing at capacity, it is hard to say that they are effectively colluding on prices. Full-fledged price competition could not cause prices to be lower than the level at which demand and supply would be equated given full capacity utilization. Effective collusion on prices must, therefore, go hand-in-hand with some degree of output restriction, that is, excess capacity. With this clarification, one can ask how the presence of capacity constraints affects the analysis already provided in which such constraints were absent. Put differently, does the presence of excess capacity make it easier or more difficult to sustain collusion?

Following the literature, we frame this discussion in terms of firms that can produce at constant marginal cost up to some well-defined capacity level and not beyond that point (in the short run). More generally, one could study models in which each firm has a smoothly increasing marginal cost curve. The resulting analysis would be considerably more complex but lead to similar tradeoffs and conclusions.

The effects of symmetric capacity constraints on collusion are theoretically ambiguous. The greater is the excess capacity at each firm, the more each firm can gain by defecting. However, by the same token, greater excess capacity means that the other firms can expand output more to punish the defecting firm. In a price-setting supergame with capacity constraints, Brock and Scheinkman (1985) show that collusion is more difficult to sustain in the presence of capacity constraints than in their complete absence, but the relationship between δ^* and the per-firm excess capacity at the monopoly price is not monotonic. Lambson (1987) generalizes these results to optimal cartel punishment strategies. Abreu (1986) obtains similar results for repeated quantity-setting games with capacity constraints.

Notwithstanding these theoretical ambiguities, in practice symmetric capacity constraints may well facilitate collusion, at least in comparison with a situation in which all firms can produce at constant marginal cost. After all, a far greater percentage expansion of output is likely to be needed for a lone defecting firm fully to benefit from price cutting than is needed for all of the firms to meet the expanded demand at the lower competitive price (especially if N is not very small). Plus, expansion for the latter purpose can take place over time.

Asymmetries *per se* in capacity constraints are likely to hinder collusion. More precisely, for a given level of total capacity, collusion is more difficult if capacity is distributed unevenly across the firms. If one firm has greater excess capacity, that firm has a greater incentive than others to cut its price, and its rivals have less of an ability to discipline that firm.[59]

[59] See Compte, Jenny, and Rey (2002), Davidson and Deneckere (1984, 1990), and Lambson (1994, 1995).

3.3.6.2. Capacity investment decisions In the longer run, the firms can adjust their capacities. One can think of these capacity choices much like quantity choices and thus interpret the results from quantity-setting supergames as applying to capacity choices over time. This approach is most reasonable if capacities are relatively short-lived, so there is little commitment value associated with capacity, or if the market is growing, so that firms are routinely adding capacity. With this interpretation, the discount factor reflects the time over which one firm can observe other firms' capacity choices and respond with its own. Since it takes longer to change capacity than to change price, the discount factor relevant for capacity decisions is lower than that for pricing decisions, making collusion on capacities more difficult to sustain, *ceteris paribus*. On the other hand, the initial capacity expansion by the defecting firm can itself take time and may be difficult to hide, making it hard for one firm to gain much of an edge on its rivals before they are able to respond.

However, treating capacity decisions just like output decisions may fail to reflect accurately some of their distinctive aspects: capacity investments tend to be lumpy and involve significant sunk costs. The irreversible nature of capacity choices is emphasized by the literature on preemptive capacity investment, which predicts outcomes that are *more* competitive than the static Cournot equilibrium.[60] The commitment aspect of building capacity tends to make collusion on capacity more difficult: after one firm adds capacity, it may not be credible for the other firms to add capacity as well, or not as much as would be needed to deter the initial expansion.

Capacity choices can interact with pricing choices over time in complex ways. Benoit and Krishna (1987) show that firms will choose to build and maintain excess capacity to support a collusive pricing outcome. Davidson and Deneckere (1990) study a "semi-collusive" equilibrium in which the firms first pick capacities and then play a repeated pricing game, setting prices at the highest sustainable level.

3.3.7. Market dynamics

3.3.7.1. Demand growth, demand shocks, and business cycles In the simple models of repeated pricing competition, demand growth makes collusion easier because a defecting firm sacrifices more in future profits in exchange for a short-term increase in its market share. Likewise, if demand is declining, defection is more tempting. One can easily incorporate these ideas into the simple model presented above by adding a market growth factor.

However, these results rely on several assumptions that may not be justified in the presence of market growth or decline. First, they assume that defection today will forever disrupt collusion and lead to highly competitive outcomes in perpetuity. We already observed that the firms will be tempted to renegotiate to avoid this unpleasant outcome. The incentive to renegotiate is greater if the market is growing. We have also emphasized that applying very strong punishments is not optimal in the presence of imperfect

[60] See the models of two-stage competition in Shapiro (1989) and the citations therein.

detection and that punishments proportional to the deviation are attractive. The lack of proportionality between today's defection and perpetual punishment is even greater in the presence of growing demand. Second, the results ignore the possibility that growing demand will induce new firms to enter the market. The prospect of future entry makes it more tempting to defect in the present and less valuable to maintain cooperation among the current incumbents. Third, the simple model of repeated price-setting does not account for the fact that growing demand may tempt the firms to engage in preemptive capacity additions.

The logic of collusion implies that the temptation to defect depends upon the relative size of current versus expected future demand. This has implications for short-term demand shocks, which are distinct from secular growth or decline in demand. Rotemberg and Saloner (1986) study a model in which demand shocks are independently and identically distributed, so demand today conveys no information about demand in the future. A positive demand shock thus makes defection relatively more attractive. This same logic can be applied to collusion over the business cycle. Haltiwanger and Harrington (1991) show that collusion is more likely to break down in the portion of the business cycle during which demand is declining. Bagwell and Staiger (1997) generalize these results to a model in which demand alternates stochastically between boom and recession phases. Porter (1983) and Ellison (1994) apply some of these ideas to the Joint Economic Committee, a railroad cartel from the 1880s. In a nice empirical application, Borenstein and Shepard (1996) find that retail gasoline margins are higher when future demand is expected to be higher or future costs are expected to be lower. Lastly, we stress that when demand is unpredictable, as in Green and Porter (1984), collusion is more difficult to sustain because the firms have greater difficulty distinguishing demand fluctuations from cheating on the collusive agreement.

3.3.7.2. Disruptive innovation The more likely it is that the market will experience a disruptive innovation, the harder it is to sustain collusion. To see this, suppose that each period there is some probability, ψ, that a major new technological innovation will be introduced into the market, disrupting the collusive agreement. (A similar analysis applies to other factors that might disrupt the agreement.) For example, a major innovation may disrupt the collusive agreement because it is introduced by a new entrant or because it introduces such a sharp asymmetry among the existing firms that cooperation is no longer sustainable. Suppose that the innovation ends the profit flows for the incumbent suppliers. Under these conditions, the payoff from defecting remains at $\pi(P)$ but the payoff from cooperating is reduced because future profits must also be discounted by the probability that disruption occurs. Formally, this is equivalent to changing the discount factor from δ to $\delta(1-\psi)$, making collusion more difficult to sustain.

3.3.7.3. Switching costs, network effects, and learning by doing Defection is more tempting if the defecting firm can gain a lasting advantage over its rivals, either in terms of market share or cost. With consumer switching costs, at least some of the customers gained today from a price cut will remain in the future even if prices fall once the defec-

tion has been observed. Capturing customers today has lasting value for the defecting firm: in models of competition with switching costs, a firm's installed base is a valuable asset, even if the firms compete vigorously to gain new customers. The logic here is similar to that of cutting price when customers' demand is high: the defecting firm captures more sales by cutting its price today. On the other hand, in the presence of customer switching costs it can be more difficult to attract customers in the first place.

Collusion also can be difficult to sustain in the presence of strong network effects, at least if the firms sell incompatible products. In the clearest case, where the market is bound to tip toward one product standard or another, collusion between incompatible products is difficult to maintain since the firm that is losing the standards battle may be very tempted to engage in price-cutting, or some other tactic, to avoid entering a downward spiral.

A similar dynamic arises in the presence of learning by doing. If learning is based on cumulative output, a firm that expands its production today will experience lower costs tomorrow, thereby gaining a lasting advantage. Due to the commitment and preemption aspects of higher current production, a firm that is more aggressive today captures more profits in the future, making collusion more difficult to sustain in the presence of strong learning-by-doing effects.

3.4. Agreements under antitrust law

3.4.1. On the meaning of agreement

As described briefly in subsection 3.1, there seems to be a contrast between the economic and legal approaches to the regulation of collusive behavior. Under the economic approach, one first attempts to determine the existence of collusion and the magnitude of its effects and then considers which if any remedies are appropriate. Under the legal approach taken by antitrust, the first step is the determination of whether there exists an agreement, and, if there is, certain legal sanctions apply: in the United States, these are treble damages to injured customers, criminal penalties on perpetrators including fines and imprisonment, and possibly injunctions against particular practices.

The extent to which these approaches diverge depends importantly on the legal concept of agreement. One standard definition—found in dictionaries and common usage in many contexts—is that an agreement signifies harmony of opinion or action.[61] Under that straightforward notion, collusion seems nearly synonymous with agreement. Indeed, a typical dictionary definition of collusion is a secret agreement or cooperation, suggesting further that the terms have the same meaning.[62]

[61] The definitions throughout are taken from Merriam-Webster's Collegiate Dictionary (10th ed. 1993), without quotation marks or ellipses. Sometimes other definitions are listed as well.

[62] If the legal term "agreement" was interpreted to require secrecy, then the law would in essence offer a complete defense whenever price-fixers were willing to reveal their plans, which they would have every incentive to do if that insulated them from legal liability.

It seems, however, from legal materials—court opinions, agency pronouncements, and commentary—that the law's notion of agreement is different, in particular, narrower. Nevertheless, it has remained somewhat mysterious just what more is required. Return to the classic example of an undoubted agreement: the secret meeting in a smoke-filled room at which competing firms suggest prices to each other, settle on a particular price, and indicate their assent to adhere to that price. Suppose we remove the smoke from the room, and then the room itself—for example, the firms might use a conference call or e-mail (or, as in one antitrust case, enter fares and symbols on a common electronic airline reservation system). Now, let us dispense with the secrecy: perhaps the firms might speak to each other through sequential press conferences. At the conclusion of this sequence, we have a sort of behavior that is often observed and is generally considered to be legal, that is, not to constitute an agreement. But why? Which step has anything to do with whether or not the firms *agreed* to anything?[63] (As already mentioned, it is the economist's term, collusion, not the legal term, agreement, that often denotes secrecy.)

As one reads legal statements on the subject, it appears that communication is central to the inquiry. Again resorting to standard definitions, communication refers to a process by which information is exchanged between individuals through a common system of symbols, signs, or behavior. By that standard, press conferences surely involve communication. So does virtually any other means of effective collusion.

Consider another simple example. In a somewhat remote area, there are two retail gasoline stations located on opposite corners of an intersection. Each posts its price on large signs readily visible from the road—and, of course, from the other station. The competitive price is $2.00 and the monopoly price $3.00. One can easily construct a sequence of interactions—wherein each station owner posts various prices, waits to see the other's response, then adjusts his or her own price, and so forth. We would predict that, even if neither benefited from a formal course in game theory, they might readily settle on a price near $3.00. The time during which a defector could reap profits without response might be a matter of minutes, not months. Hence, successful collusion seems quite likely.

The legal question is whether the two owners have "agreed" to price at $3.00. Suppose, as suggested, that the legal system gives content to the term agreement by asking whether the parties communicated with each other. Well, they did not *speak* to each other; they may not even speak the same language. However, in the relevant sense, they did speak to each other in a common language, that of price. The absence of words may have slightly lengthened the time it took to settle (agree?) on the price of $3.00. And, should one station cut its price to $2.90 (in the absence of any change in market conditions, such as a drop in the price of fuel from refineries), the other station owner's quick

[63] It may matter for other reasons whether communications are public. For example, buyers may value having information sooner. (However, buyers do not value means of communication that make collusion against them possible, even if one consequence is that they learn of adjustments to collusive prices somewhat sooner.) In any event, it is not clear how this consideration bears on whether there exists an agreement.

response, cutting its price, say, to $2.80, will be pretty unambiguous; it will be understood as an invitation to raise prices, an invitation that would be accepted by posting a $3.00 price.

Examples like these seem to suggest that there is little, if any, difference between the legal requirement of agreement and the economist's notion of collusion. Yet it also seems that few think that this is actually the case. Surely, it is believed, the law requires more: more evidence of agreement, usually through more evidence of communications.[64] Yet, as should now be clear, it is hard to tell what more is being sought. It seems that some different sort of evidence is required, but evidence of what?

Some legal utterances distinguish between "express" agreements and "tacit" agreements. Tacit ordinarily means that the communication does not use words or speech (which, by contrast, is what is meant by express). By that definition, the press conferences, being conducted using words, would constitute express rather than tacit agreements, but the gasoline station owners, using signs, would not be express agreements—unless, of course, one pointed out that a sign showing "$3.00" is functionally equivalent to a sign showing "three dollars," the latter, containing words rather than numerals, constituting an express rather than tacit communication. Likewise, one could consider sign language, other hand signals, winks and nods, and so forth. Indeed, it is hard to believe that a sensible legal regime would make legality—and heavy consequences—turn on subtleties of modes of expression and taxonomic disputes over which constitute "expressions" or "communications."[65]

Official legal pronouncements, although sometimes seemingly clear, are not that helpful either. U.S. Supreme Court opinions include famous statements such as the following:[66] " '[C]onscious parallelism' has not yet read conspiracy out of the Sherman Act entirely."[67] But this merely indicates that purely independent action—such as different gasoline stations raising their prices in parallel when the price of oil rises—does

[64] The "statutory language [of Sherman Act Section 1] is broad enough ... to encompass a purely tacit agreement to fix prices, that is, an agreement made without any actual communication among the parties to the agreement. ... Nevertheless, it is generally believed ... that an express, manifested agreement, and thus an agreement involving actual, verbalized communication, must be proved in order for a price-fixing conspiracy to be actionable under the Sherman Act." In re High Fructose Corn Syrup Antitrust Litigation, 295 F.3d 651, 654 (7th Cir. 2002) (opinion by Judge Posner). See also the further discussion of this case in note 72 below.

[65] Not only do legal authorities devote little attention to defining "agreement," but when other terms like "express" are employed, these too are not elaborated. Furthermore, although one standard definition of express is to represent in words, other standard meanings include to make known (regardless of the mode), to reveal impulses artistically, and to represent by signs and symbols, which covers the full gamut, including presumably most meanings that many of those who use the term "express" intend to exclude.

[66] As noted earlier, the European Union has a similar agreement requirement that likewise extends beyond formal contracts and is imprecise. In *Dyestuffs*, the European Court of Justice elaborated the concept of a concerted practice as "a form of co-ordination between undertakings which, without having reached the state where an agreement properly so called has been concluded, knowingly substitutes practical co-operation between them for the risk of competition." ICI Ltd. v. Commission, Case 48/69 [1972] ECR 619, ¶64. Although some sort of contact between the parties seems to be required, the Commission seems inclined to find behavior illegal even when the contact is indirect. See Bellamy and Child (2001).

[67] Theatre Enterprises v. Paramount Film Distributing Corp., 346 U.S. 537, 541 (1954).

not constitute an agreement. Or consider: "The essential combination or conspiracy in violation of the Sherman Act may be found in a course of dealing or other circumstances as well as in an exchange of words. . . . [A conspiracy may be found where] the conspirators had a unity of purpose or a common design and understanding, or a meeting of the minds in an unlawful arrangement. . . ."[68] This (partially question-begging) expression aligns substantially with the idea that successful collusion is sufficient. More recently (and more commonly quoted in modern cases), the Supreme Court has stated that evidence must be presented " 'that tends to exclude the possibility' that the alleged conspirators acted independently . . . [that is,] that the inference of conspiracy is reasonable in light of the competing inference[] of independent action. . . ."[69] Here, the interpretation depends on the meaning of "independent." If taken to mean "without regard to others," then collusive behavior is not independent action and thus is sufficient to trigger liability. Yet another pronouncement (in a more recent case, but not one directly addressed to the agreement question) is that "[t]acit collusion, sometimes called oligopolistic price coordination or conscious parallelism [is] not in itself unlawful."[70] Tacit collusion, however, is undefined and is not generally understood to be the same as conscious parallelism. What all of these court decisions and most other statements have in common is that key terms are not defined, the subject is not directly discussed in any depth (that is, for more than a paragraph), and no rationale is offered for deeming one set of scenarios to be legal and another illegal.

A further important complication is that it is well accepted that, whatever is required to establish an agreement, it is allowable (and typical) for the demonstration to be indirect, through circumstantial evidence. So-called "smoking guns" are not required. For example, if the law demands proof of direct verbal communications on the specific price and pattern of punishment, it might be argued that near-simultaneous price increases, and then declines in response to defections, are evidence of such communications and hence sufficient to establish a violation. The implicit logic is, "How else could this behavior be explained?" This perplexing question and some of the earlier discussions on the possible meanings of agreement require further attention to the role of communications in the economic theory of collusive behavior.

3.4.2. Agreement, economics of collusion, and communications

Suppose, as seems to be believed by most, that the legal requirement of an agreement is satisfied only by certain types of communication: perhaps verbal statements or close equivalents, sufficiently directed at competitors, that relate closely to pricing behavior, and that may be responded to reasonably promptly, precisely, and directly. What, then, is

[68] American Tobacco Co. v. United States, 328 U.S. 781, 809–10 (1946).
[69] Matsushita Electric Industrial Co. v. Zenith Radio Corp., 475 U.S. 574, 588 (1986), quoting Monsanto Co. v. Spray-Rite Service Corp., 465 U.S. 752, 764 (1984).
[70] Brooke Group Ltd. v. Brown & Williamson Tobacco Corp., 509 U.S. 209, 227 (1993).

the relationship between these more explicit sorts of communication and the economic theory of collusion?

Reviewing the economic theory of collusion as summarized in subsections 3.2 and 3.3, communication may be relevant at a number of points. First, the nature of communications may bear on the ease of reaching consensus. In the example with two gasoline stations, rather simple communications seem sufficient. But if there are more firms, greater heterogeneity (in costs, products, or other features), more uncertainty about buyer behavior, or other complicating factors, greater negotiation may be required, which in turn might be facilitated by more explicit (direct, head-to-head, simultaneous, prolonged) communications. This view is not entirely obvious, however, for if all parties knew that they were limited to a few rounds of simple price suggestions, after which they must have reached agreement, it is possible that agreements would be reached more quickly and with greater likelihood (although perhaps they would also be less durable, due to misunderstandings). While this discussion is largely outside simple models of repeated oligopoly, which typically ask whether a price P can be sustained, these questions are addressed in the literature on "cheap talk," cited above.

Second, in the detection of cheaters, explicit, detailed communication might also be helpful. If firm A's cheating is noticed by firm B, firm B could tell others. Compte (1998) and Kandori and Matsushima (1998) address this possibility in a model where firms observe and can communicate their private information about past play. Alternatively, if other firms suspect that firm A is cheating, discussions with firm A (perhaps supported by firm A presenting original invoices or other information) might help clear up the matter, avoiding price wars due to mistaken inferences. The Sugar Institute operated very much in this manner, as described by Genesove and Mullin (2001). Oligopoly theory is slowly moving more in the direction of modeling these types of issues, at least by exploring the role of communications about private cost information. See Athey and Bagwell (2001, 2006).

Third, punishment might better be coordinated with more explicit communication. Determining the magnitude of the price cut and its duration, perhaps focusing punishment when firms' product lines and regions of operations vary, and other aspects of strategy might be worked out more effectively. As with reaching consensus, however, greater opportunity for detailed communication may be a double-edged sword. As noted, the opportunity for renegotiation can undermine punishment. In any case, it is generally assumed in formal models that some particular punishment strategy has been chosen and will be pursued; the question explored is whether the strategy, if pursued, would deter cheating ex ante, or whether the strategy is credible, not what communications may be necessary to select or effectuate the strategy.

Fourth, inclusion might be enhanced through detailed negotiations. This consideration is based on reasoning similar to that of reaching consensus and is likewise outside standard formal analysis.

In all, there are many reasons to believe—and it generally is believed—that greater opportunity for freer, more detailed, explicit communication tends to facilitate collusion (although there are some countervailing factors). If this is indeed the case, it follows that

it is more important to prohibit more explicit forms of communication. Still, the question remains, why not prohibit all communication? The answer must be that various forms of communication—such as making price information available to customers—serve other, legitimate purposes, and that less explicit communications—such as sharing aggregated and lagged sales information through a trade association—are more likely to promote socially valuable functions than to facilitate collusion. This statement, too, is not obvious, for many socially valuable functions, such as the setting of compatibility standards for emergent technologies or the sharing of information about industry conditions, require highly explicit communication. Furthermore, as the case of the two gasoline stations illustrates, in some instances facilitating collusion requires very little explicit communication.

Additionally, directing the legal inquiry at the nature of communications—which themselves often cannot be observed by the tribunal but must be inferred from circumstantial evidence—raises what might be called a paradox of proof. Suppose available evidence indicates that, in the situation under scrutiny, collusion is especially easy and the danger of supra-competitive pricing is accordingly very high. Moreover, evidence conclusively demonstrates that we have experienced a collusive outcome—at roughly the monopoly price—for years. Who wins? Arguably, the defendants. They could argue that, precisely because collusion is so easy, they were able to achieve monopolistic results—and, they gleefully concede, will be able to continue to do so for the foreseeable future—without any meetings in smoked-filled rooms, elaborate negotiations, and so forth. Just a few public pricing signals and they were off. Moreover, since all have taken courses in strategy at business school and all are advised by the leading consulting firms and their affiliated game theory experts, coordinating punishment with only minimal, indirect communications is a snap. Hence, the very strength of the evidence of the ease and success of collusion makes it implausible to infer that the defendant firms must necessarily have met and had long discussions about price-fixing.

Reflecting on this case and other possibilities, it would seem that the relationship between the ease of collusion and the likelihood that there were sufficiently explicit communications to trigger liability under the agreement requirement (whatever it turns out to be) is not monotonic. Put differently, we are asking just how should the factors listed above, which make it more or less difficult to sustain collusive outcomes, be incorporated into a price-fixing case in which the existence of an agreement is proved through circumstantial evidence.

Beginning at one end of the spectrum, suppose that industry conditions are such that it is extremely difficult for the firms to sustain a collusive outcome because there are many firms, low entry barriers, price-cutting by one firm is very difficult for rivals to observe, and demand and cost are highly variable. Under these industry conditions, we would not expect the firms to have engaged in unobserved meetings that satisfy an explicit communication requirement simply because such meetings would likely be futile. Moreover, if it is nevertheless asserted that collusion occurred, we just will not believe that an effective price-fixing agreement was reached. The evidence on pricing and cost could not have been certain, and any uncertainty is naturally resolved against

an inference of collusion because it would be nearly impossible under the observed industry conditions.

Consider next an industry in which conditions are such that collusion is somewhat easier to sustain, perhaps because the industry is more concentrated, has moderate entry barriers, pricing is more transparent, and demand and cost are less volatile. These industry conditions make it more likely, but still far from inevitable, that a collusive outcome could arise. Under these circumstances, if collusion indeed seems to have occurred, its very difficulty (but not such high difficulty as to blend toward impossibility) suggests that explicit communications may well have been employed to carry it off.

Toward the opposite end of the spectrum, consider an industry in which the conditions are highly conducive to collusion: highly concentrated, no prospect of entry, transparent pricing, and stable demand and cost. Think of the two gas stations. We now have the case with which we began, presenting the paradox of proof: the very ease of collusion negates the inference that there must have been elaborate, explicit communications.

In sum, as industry conditions move from those that make collusion nearly impossible to those that make it incredibly easy, the inference that there must have been highly detailed communications first becomes stronger and then weaker. It is rather hard to say where on this continuum the maximum inference arises, or in what intermediate range some given proof standard is satisfied.[71]

How does this paradox of proof square with the law and what we observe in practice? U.S. courts typically insist on the presentation of various so-called "plus factors." Yet these factors are often little more than indicators that collusion rather than purely independent behavior is likely to have occurred.[72] As just explained, such factors indeed

[71] The more one pushes the logic underlying the inference of agreement, the more complex it becomes. For example, in the region in which collusion is moderately difficult, a slight increase in the ease of collusion makes it more likely that collusion was attempted, which raises the likelihood of a given type of explicit communication, but, conditional on collusion having been attempted, reduces the likelihood that communication was more explicit because, by hypothesis, collusion is becoming easier. The depiction in the text, which assumes a single peak, may be overly simplistic. Moreover, one supposes that different industry conditions in different combinations that contribute to the ease or difficulty of collusion may have varying effects on the need for more explicit communication and the forms that it will take.

[72] Some of the most common factors seem to go little beyond requiring interdependent rather than independent behavior. For example, prominent plus factors include various sorts of evidence showing that the firms' actions are "against self-interest" in the absence of collusion. For a survey and critical commentary, see Areeda and Hovenkamp (2002, vol. 6, 241–250). Courts also frequently rely on evidence that purports to directly indicate the existence of an agreement. For example, Judge Posner in In re High Fructose Corn Syrup Antitrust Litigation, 295 F.3d 651, 662 (7th Cir. 2002) offers, among others, the following quotations from the alleged conspirators as evidence of the existence of the requisite agreement: "We have an understanding within the industry not to undercut each other's prices." "[O]ur competitors are our friends. Our customers are the enemy." A competitor's president is called a "friendly competitor" and mention is made of an "understanding between the companies that … causes us to … make irrational decisions." As the discussion in the text explains, however, there can exist such an "understanding" and firms can view competitors cooperatively as a result of education about collusion, good advice, common sense and experience, and open communications (such as the gas stations' posting of prices), so it is difficult to discern in what sense more than the existence of consciously interdependent, collusive interaction is required.

favor the inference of an illegal agreement if but only if we are on the "difficult" side of the maximum, where additional evidence indicating the ease or benefits of collusion makes the likelihood of the requisite communications higher. On the other side of the maximum, they make collusion more likely but explicit communications less likely.

We are unaware of any cases (nor have we ever heard anyone suggest the existence of any cases) in which parties and courts acted as if they were in what we are referring to as the paradox region, that is, past the peak, such that evidence that collusion is more feasible makes the inference of detailed communications less likely. How can one explain this one-sidedness? One possibility is that, even though the law has had this character for over half a century in the United States (and for shorter, although significant, periods in many other jurisdictions), no one has really understood the nature of the legal requirement.

Another possibility is that all cases are in fact at the difficult side of the maximum. That is, there are no industries where successful collusion is at all likely in the absence of highly explicit communications. Observe, however, that if this were true, the agreement requirement would be superfluous. That is, if there exists collusion, there must have occurred the requisite communications to trigger the agreement requirement. Were this always true, nothing would need to be proved beyond the mere existence of collusion. (This would suggest that Posner's aforementioned prescription would be implied by existing law, and thus not constitute a significant departure from it.)

Yet another possibility is that there are cases past the maximum, in the paradox range, but defendants are reluctant to advance the argument that the proof against them implies the absence of any agreement and hence victory. The reason is that, in conceding that collusion is easy, likely, and probably in fact has occurred and will continue, they fear that they will hurt their case. Defendants may suffer in the determination of liability because, as a practical matter, a fact finder (whether a jury, judge, or expert tribunal) is more likely to condemn them if they in fact operate in a situation inherently conducive to collusive outcomes and are likely taking advantage of it. They win on the formal law but lose because they show themselves to be greedy and behaving in an antisocial manner. In that event, it may be that *de facto*, the greater the danger of collusion, the greater the likelihood of liability, without regard to any inference that does or does not follow about explicit communications and the satisfaction of the agreement requirement. Additionally, if there is a sufficient prospect that liability will be found, defendants may be worried about penalties. The more they argue that collusion is easy, the more plausible will be high estimates of overcharges (in amount and duration) and thus the greater will be fines and damage payments.

This discussion may raise more questions than it answers, but we believe that it is, ultimately, clarifying. The economic analysis of collusion, although quite complex, is at least fairly straightforward in stating the question it addresses and the motivation for the inquiry it undertakes. Upon examination, the same cannot be said about the law's requirement of an agreement and the role of industry conditions in inferring that such an agreement exists. We hope to have advanced understanding in two ways: by being more precise about what agreement might mean, and, for a given definition, by being

more explicit about the relationship between the economics of collusion and whether such agreement requirement is satisfied.

Two additional observations about the interplay between the economics of collusion and antitrust law are in order. First, under antitrust law it is possible for there to be a violation even when it is clear that no successful collusion occurred. If competitors meet formally, enter into a written agreement, but ultimately fail miserably in executing it, most legal regimes would find a violation. If fines or damage awards were limited to a multiple of the overcharge, this finding would be moot. However, other sanctions may be employed; notably, those engaged in the attempt may be put in prison. It is sometimes efficient to punish unsuccessful attempts (especially when detection is difficult and limits on sanctions may make it impossible to punish violators sufficiently to achieve effective deterrence), and examining direct communication may help to identify unsuccessful attempts. Of course, evidence about pricing patterns of the sort that might be deciphered by economic experts may also aid in the task, especially if there were efforts to put the agreement into effect.

Second, separate from the agreement requirement, penalties may depend substantially on the extent and duration of overcharges. Undertaking these measurements requires expert economic analysis. The greatest difficulty, of course, is in determining what would have been the price but for the collusion. It is necessary both to specify conceptually the nature of the equilibrium that would otherwise have prevailed (perfect competition? monopolistic competition in price with differentiated products?) and to calculate just what price would have prevailed in that equilibrium.[73] This inquiry is very closely related (in some respects, identical) to that necessary to identify whether collusion existed in the first place.

3.5. Other horizontal arrangements

Our analysis has focused almost entirely on collusion that involves arrangements purely concerned with the fixing of prices. Simple price-fixing, in turn, is unambiguously—"per se"—illegal in the United States and subject to similar prohibitions elsewhere. There exists, however, a variety of horizontal entities—partnerships, trade associations, joint ventures, standard-setting bodies, to name a few—and such entities engage in myriad forms of conduct.

Certain horizontal arrangements can serve as substitutes for direct price-fixing. As noted, firms might agree to divide territories or customers so as to eliminate competition. Although no particular price has been set, each firm is left to act as a monopolist with respect to its portion of the market, so the result is similar to that of a price-fixing cartel. The economic analysis is analogous: firms must be able to agree on the market

[73] If damages were based not on the overcharge times the quantity purchased, as is ordinarily the case, but instead or also on losses of consumer surplus regarding units not purchased, information on the entire relevant segment of the demand curve would be required.

allocation (instead of the price), cheating (selling to other firms' allotted customers) must be detectable and subject to effective punishment, firms with significant capacity need to be included in the agreement, and entry must be limited. Likewise, legal scrutiny tends to be similar: pure horizontal divisions of the market among competitors are also *per se* illegal in the United States.

Not all horizontal arrangements involve pure schemes to fix prices or divide the market (often called "naked restraints"). Nevertheless, many horizontal arrangements pose some risk to competition. Accordingly, antitrust laws need to draw distinctions. Under the law in the United States, this is done under the rubric of the "rule of reason": reasonable schemes are permissible; unreasonable ones are prohibited. Obviously, this concept needs to be fleshed out and related to economic analysis, and that task will be our central focus in this section. But first we will consider a particular class of horizontal arrangements that is closely related to our foregoing discussion of collusion and the legal prohibition on price-fixing.

3.5.1. Facilitating practices

In our consideration in subsection 3.3 of conditions bearing on the likelihood of successful collusion, we largely took such conditions to be exogenous. Some factors, however, are within the firms' control, individually or collectively. Antitrust scrutiny has focused primarily on the latter.[74] In this regard, two lines of attack must be distinguished. First, is horizontal agreement on some practice that facilitates collusion itself an illegal agreement in restraint of trade and thus an independent basis for liability? Most challenges, and our own discussion, emphasize this inquiry. Second, does the use of facilitating practices constitute evidence of the existence of an underlying agreement directly to fix prices? In some respects, the distinction may be immaterial, notably if both agreements on the facilitating practice itself and agreements on price-fixing are illegal and if the remedy is the same. (A remedial difference is that a facilitating practice might independently be enjoined.) Nevertheless, in the context of evaluating the evidence in a particular case, it clarifies thinking to keep this difference in mind.

An important facilitating practice that has long been the subject of antitrust regulation concerns information exchanges among competitors, sometimes in the context of trade association activity and other times conducted independently.[75] For example, in

[74] In E.I. du Pont de Nemours & Co. v. Federal Trade Commission, 729 F.2d 128 (2nd Cir. 1984), the FTC unsuccessfully challenged what it asserted to be facilitating practices that were unilaterally adopted (although employed by all four firms in the industry), claiming authority under Section 5 of the Federal Trade Commission Act, which does not require the existence of an agreement. To the extent that facilitating practices can only be challenged when their existence is attributable to an agreement, emphasis is placed on the issues considered in subsection 3.4, notably, under what circumstances an agreement can be inferred when multiple firms employ a facilitating practice.

[75] As elsewhere, our discussion focuses on U.S. antitrust law. EU law also encompasses facilitating practices, including exchanges of detailed information among competitors in industries prone to collusion. See, for example, Bellamy and Child (2001, §4-042).

the *American Column & Lumber* case, a violation was found where firms exchanged information on prices in individual transactions and this information was subject to audit for accuracy.[76] Such information greatly eases the detection of cheaters, whereas such details do not have an obvious and substantial productive use. In *Container Corporation*, firms were deemed to have violated the antitrust laws when they called competitors to verify the accuracy of buyers' assertions of having been offered lower prices elsewhere.[77]

Another class of interest involves firms' contracts with their customers. For example, a firm's use of a most-favored-customer clause—under which it agrees to give all customers under contract the benefits of any price cut extended to a subsequent customer—may greatly reduce its incentive to defect from a collusive price since it must sacrifice profits on its existing customer base that was otherwise locked in for a period of time at a higher price. (This disincentive would be immaterial for an infinitesimal price cut, but if, as previously discussed, greater price cuts are necessary to attract substantial new business, the disincentive could be substantial.) Some firms employ price-matching (meeting competition) clauses, under which they promise to lower their price if the buyer can find a competitor that charges less. (Some clauses promise to equal the price, perhaps even retroactively, that is, on previous sales, and others promise to exceed the competitor's price reduction.) This arrangement deters other firms from lowering their prices. Moreover, it facilitates detection because buyers offered lower prices are more likely to reveal otherwise secret price cuts. Observe that under these arrangements buyers as a whole are disadvantaged—if effective, the market price is sustained at a higher level—but individual buyers are subject to the free-rider problem: each may well gain (if there is any chance that some seller will lower price), but its contribution to a higher market price will be negligible if it is a small purchaser.[78] These cross-currents are explored in the economic literature on most-favored-customer clauses and meeting competition clauses.[79]

Other types of practices are directed at coordination problems caused by product heterogeneity and competition along dimensions other than price. Quality or grading standards may promote uniformity—or at least reduce variety to manageable proportions—facilitating agreement on price. Agreements may limit credit (and other) terms, lest firms cheat on the price by offering favorable interest rates.

More broadly, any factor that may inhibit collusive pricing is potentially subject to firms' creativity in devising means of avoiding its detrimental effect. There are, of course, limits on what is feasible. Furthermore, to the extent that the use of facilitating practices itself requires collusion, firms must overcome any difficulties of coordination,

[76] American Column & Lumber Co. v. United States, 257 U.S. 377 (1921).

[77] United States v. Container Corporation of America, 393 U.S. 333 (1969).

[78] This type of free-riding problem also arises when a monopolist employs exclusive dealing provisions with its customers, as discussed in subsection 5.4.1 below.

[79] See Cooper (1986), Edlin (1997), Edlin and Emch (1999), and Salop (1986). Borenstein (2004) applies this idea to price matching in the airline industry and the airline tariff publishing case.

detection, and enforcement with regard to the facilitating practices themselves, as illustrated by the *Sugar Institute* case. Some facilitating practices may be readily formulated and observed; others may be complex or hidden. Accordingly, the successful use of facilitating practices will vary greatly.

There remains another important consideration with many facilitating practices: they may have redeeming virtues. Some information exchange enhances planning. Forcing trading into formal markets (which was permitted in *Chicago Board of Trade*[80]) produces benefits that flow from public prices. Contractual arrangements with buyers regarding the sellers' and competitors' prices reduce search costs. Exchange of cost information may enhance productive efficiency by shifting output to more efficient firms; see Shapiro (1986). Even what may seem literally to be price-fixing will often be efficient, such as when productive partnerships or joint ventures are formed and the resultant entity fixes a single price for its common product. Likewise, many other arrangements that may seem beneficial may also have effects on the feasibility of collusion. Accordingly, it is necessary to formulate a means of balancing the costs and benefits, which is the subject of the next subsection.

3.5.2. Rule of reason

In the United States, the "rule of reason" was formally announced nearly a century ago, in the monopolization case of *Standard Oil*.[81] Shortly thereafter, it was given more content in *Chicago Board of Trade* in language that is routinely quoted (or paraphrased) to this day: "The true test of legality is whether the restraint imposed is such as merely regulates and perhaps thereby promotes competition or whether it is such as may suppress or even destroy competition."[82] Although more specific than the almost-completely question-begging inquiry into "reasonableness," the meaning of this test is hardly self-evident. Just what is meant by "competition"? Is it valued purely as a means or as an end unto itself? It is useful to begin with a few modern invocations of the rule of reason's promoting-competition test, followed by some reflection on the broader question of interpretation and its relationship to economic analysis.[83]

[80] Board of Trade of City of Chicago v. United States, 246 U.S. 231 (1918).

[81] Standard Oil Co. of New Jersey v. United States, 221 U.S. 1 (1911).

[82] Board of Trade of City of Chicago v. United States, 246 U.S. 231, 238 (1918). The Court continued: "To determine that question the court must ordinarily consider the facts peculiar to the business to which the restraint is applied; its condition before and after the restraint was imposed; the nature of the restraint and its effect, actual or probable. The history of the restraint, the evil believed to exist, the reason for adopting the particular remedy, the purpose or end sought to be attained, are all relevant facts. This is not because a good intention will save an otherwise objectionable regulation or the reverse; but because knowledge of intent may help the court to interpret facts and to predict consequences."

[83] In the European Union, conduct may be deemed exempt from the prohibition in Article 81(1) on anticompetitive agreements if it meets certain criteria in Article 81(3) that bear resemblance to the rule of reason in the United States. There are both block (general) exemptions and those granted individually. To enhance clarity, the Commission has issued Guidelines on the Applicability of Article 81 of the EC Treaty to Horizontal Cooperative Agreements (European Union, 2001).

In *National Society of Professional Engineers*, the Society had an ethics rule prohibiting engineers from bargaining about price until after they were selected for a project.[84] The proffered justification was that otherwise customers might be induced to focus excessively on the price of professional services at the expense of concerns about quality and safety. The Supreme Court found a violation. Safety was not deemed unimportant, but rather something that ultimately was for customers to decide. They could employ the Society's approach if they wished, but the Society could not impose this choice on all customers. Competition meant free and open choice, not one side of the market collectively dictating terms to the other.

In *Indiana Federation of Dentists*, cost-conscious insurance companies employed an internal procedure for reviewing submissions for reimbursements.[85] The dentists objected (to non-dentists passing professional judgment on their work, so they claimed) and agreed as a group not to supply the necessary documentation. They too lost. Once again, it was for the customer—or, in essence, the customer's agent, the insurance company—to make whatever judgments it wished. Any individual dentist was free not to deal with any insurer if the dentist thought the insurer's practices inappropriate (or for any other reason or for no particular reason), but dentists could not agree, as a group, to impose their judgment.

Consider also *National Collegiate Athletic Association*, involving agreements among universities regarding college football.[86] The Supreme Court was not bothered by their agreements on rules of the game (size of playing field, scoring, and so forth)—rules that were not challenged—but did find their restrictions on schools selling television rights independently of the Association's scheme to constitute a violation.[87]

For the most part, cases such as these seem to see competition as a process. The view seems to be that competition consists of buyers and sellers each deciding for themselves—or, more precisely, in individual buyer-seller pairs—with whom they will deal and on what terms. Independent decisions are a central feature of competition, whereas groups (typically of sellers) who attempt to impose some regime regarding the proper terms of dealing are subverting the process. They may or may not be right, but that is not the question. Put another way, what is right is essentially taken to be whatever is the outcome of the competitive process, much like how one accepts the equilibrium price in a competitive market as "reasonable."

Perhaps competition is viewed as good in itself. Or instead the view may be that competition is valued for its results, whether those understood by economists, in terms of

[84] National Society of Professional Engineers v. United States, 435 U.S. 679 (1978).

[85] Federal Trade Commission v. Indiana Federation of Dentists, 476 U.S. 447 (1986).

[86] National Collegiate Athletic Association v. Board of Regents of the University of Oklahoma, 468 U.S. 85 (1984).

[87] This case also nicely illustrates what is sometimes referred to as the "ancillary restraints" doctrine. Namely, an anticompetitive restraint is not deemed permissible merely because it is associated with an otherwise legitimate venture; however, the restraint may well be allowed if it is reasonably necessary to accomplish the legitimate objectives of the venture.

allocative efficiency, or other notions concerning freedom of choice. Under this second, instrumental view of competition, the antitrust laws are nevertheless interpreted to relate only to the process: perhaps the integrity of the competitive process is much easier to assess than the outcome of that process, and benefits are assumed to flow as long as competition is assured. In cases where it is alleged that the competitive process is not providing the expected benefits, courts in the United States repeatedly state that the appropriate remedy is to seek legislative or regulatory action. Even when there are market imperfections, it is plausible to distrust the collective schemes of self-interested market actors, schemes that they allege to be correctives in the public interest—Adam Smith's warning being apropos.

This process view, however, is problematic. Although economists routinely use the term "competition," it does not readily bear the weight that it must under the rule of reason in judging industry practices. Is the formation of a joint venture between two firms that might otherwise compete with each other, although less effectively or with a somewhat different product, an enhancement to or a detraction from competition? How about a partnership or a horizontal merger? What of curing a market failure? Even if the result of coordinated action is unambiguously more efficient, is it more competitive? Does the competitive process include competition among institutional forms, including various forms of cooperation among groups of firms that operate in the same industry? More broadly, when the conditions for perfect, textbook competition fail (that is, pretty much always), is there an unambiguous way to describe one or another arrangement or outcome as more competitive?[88]

Economists do not traditionally answer such questions. Instead, they undertake positive analysis of behavior and outcomes under various market arrangements. For normative purposes, the ordinary metric is welfare, or efficiency, or perhaps utility to each party or class of parties, not the degree of competition according to some competition index. Yet, if the rule of reason is legally defined in terms of competition itself—that which promotes competition is legal, that which suppresses competition is illegal, end of story—then economics cannot directly address the legal test.

As it turns out, no matter how often the promote-versus-suppress-competition test is invoked, it is not adhered to uniformly, and legal authorities seem to depart from it fairly readily in many of the cases in which its application seems problematic. As noted, in *National Collegiate Athletic Association*, the Supreme Court finds horizontal agreement on rules of the game to be unproblematic. There is a sense in which this flexibility may have benefited from a fortuitous play on words, in that such rules were seen as creating competition—sports competition, that is. But even the case that first announced the now-canonical language on competition, *Chicago Board of Trade*, was

[88] As with "agreement," little aid comes from standard definitions. Competition is ordinarily taken to mean the act or process of competing, rivalry, or specifically the effort of parties to secure business of a third party. Under that rubric, even a simple partnership of two individuals who otherwise might produce (however inefficiently) on their own can readily be seen as "anticompetitive." This definition is reasonably clear, but as will be discussed it is one that antitrust tribunals often disregard, and with good reason.

one that condoned restrictions on individual actors' freedom of action to produce a greater good, trading in the public market.[89]

Other modern cases reinforce a more complex interpretation. Notably, in *Broadcast Music*, there were two large entities (BMI and ASCAP) that, between them, licensed the rights to nearly all domestic copyrighted music to various users (for example, radio and television stations).[90] The entities set a single fee for block licenses, which was challenged, among other reasons, as constituting illegal price-fixing. And in fact, each of these two entities did set single prices for bundles of millions of musical compositions that otherwise might be priced independently. Yet, economies of scale in contracting and copyright enforcement (that is, monitoring the illegal use of the entities' portfolios of music by unlicensed parties) induced the Court to find no violation. The result in the market was nothing like atomistic competition under which individual composers paired voluntarily with individual buyers, an alternative the Court found to be cumbersome. Instead, huge collections of otherwise-competitors used a sales agent to dictate price and other terms of dealing. Supposing one accepts that the permitted arrangements were on-balance desirable on efficiency grounds, there remains the question of whether the arrangements involved more or less "competition."

One device employed in *Broadcast Music* and in some other cases is to treat the challenged venture as a single entity: once viewed in this manner, there is no longer a horizontal agreement and thus no violation of Sherman Act Section 1. Looking ahead to section 4, horizontal mergers are not themselves viewed as price-fixing cartels—even though the merged firms presumably fix a common price—but rather as single entities. In such cases, however, there remains the question whether the agreement *creating* what is subsequently viewed as a single entity constitutes a violation. Carte blanche would authorize formal cartels, say, incorporated as a single firm. Of course, jurisdictions do not freely permit formal cartel arrangements or horizontal mergers. Nor do they automatically approve even loose trade associations if, for example, member firms engage in information exchanges of a sort that facilitate collusion and generate little offsetting benefit. That is, when trade association activity has been challenged successfully, no single-entity defense has been recognized. Thus, the single-entity characterization is more of a conclusion than a reason to decide one way or the other.

What, then, is the underlying meaning of the rule of reason? On one hand, antitrust law does not insist on pure atomistic competition, prohibiting all combinations from small partnerships to trade associations to joint ventures to mergers. On the other hand, horizontal arrangements are not freely permitted. Instead, they are subject to some sort of balancing test, whether under the rule of reason in the United States or under other rubrics elsewhere. When the arrangement looks like little more than a pure

[89] Our point is not to agree with the Court's analysis in *Chicago Board of Trade*, which was problematic in a number of respects, but rather to indicate that the pure, atomistic, hands-off process view of competition was never the complete story.

[90] Broadcast Music, Inc. v. Columbia Broadcasting System, Inc., 441 U.S. 1 (1979).

interference with ordinary competition, it is likely to be condemned with little further inquiry. In many cases in the United States, for example, market power need not be demonstrated, and adverse effects need not be proved (although for an award of damages these considerations will be important). Examples include *National Society of Professional Engineers*, *Indiana Federation of Dentists*, and the television marketing restrictions in *National Collegiate Athletic Association*—in addition to naked price-fixing and related practices. When, however, there appear to be benefits—from combining production, conducting research, setting standards, or otherwise—condemnation is not guaranteed, as demonstrated by *Broadcast Music*. And some horizontal arrangements, like partnerships and mergers that do not produce substantial market power, are routinely allowed.

The primary area of ambiguity concerns the many practices that fall in between the extremes. Economists can analyze the arrangements' effects and assess their efficiency. But how do such assessments relate to the legal test? Most modern antitrust rule-makers and adjudicators seem to pay substantial attention to economic considerations, at least in many settings. But under formulations like the rule of reason, the conception of reasonableness—whether or not concretized as a determination of promotion versus suppression of competition—is not well specified. We know from cases like *Broadcast Music* that pros and cons will sometimes be balanced, but what counts as a benefit or cost of an arrangement, what metric is employed for measurement and conversion to a common denominator (if this is done at all), and what is the ultimate decision rule remain somewhat of a mystery.[91] A purely economic criterion has not been explicitly embraced; nor has it been rejected.[92]

3.6. Antitrust enforcement

We close this section by commenting briefly on some of the law and economics issues that arise in antitrust enforcement.[93]

[91] "Courts sometimes describe their task under the rule of reason as one of 'balancing' potential harms against likely gains or defenses. But balancing implies that one places some measurable quantity of something on one side of a scale, a quantity of something else on the other side, and determines which side outweighs the other. The set of rough judgments we make in antitrust litigation does not even come close to this 'balancing' metaphor. Indeed, most courts do not even define a unit of measurement in which the quantities to be balanced can be measured. Assuming the relevant unit is dollars, one would need to place at least a rough dollar estimate on the dangers to competition … and a similar estimate on likely cost savings, output increases, or other benefits. To the best of our knowledge, this has never been done in any antitrust case." Hovenkamp (2005, vol. 11, p. 339). Hovenkamp, it should be noted, does not offer this depiction as a criticism. Instead, he sees such balancing as beyond the institutional competence of courts and believes that in practice they employ a structured sequence of (essentially dichotomous) inquiries that usually enables them to resolve cases one way or the other without ever having to balance costs and benefits.

[92] One question of particular interest is, supposing that the criterion is economic, whether it involves efficiency as a whole or only consumer surplus. Compare our discussion of this issue in the context of horizontal mergers, in subsection 4.4.3.

[93] For discussion of additional issues, see, for example, Posner (2001).

3.6.1. Impact of antitrust enforcement on oligopolistic behavior

The workhorse model of oligopoly used to study collusion, namely the model of repeated price- or quantity-setting, does not explicitly include any antitrust enforcement. At first blush, this seems rather peculiar, at least from the perspective of law and economics. However, if this basic model captures conduct that is believed by the parties to be beyond the reach of the antitrust laws—repeated price-setting without any other communications—then the omission is justified. This is another reminder that economic theory may be most relevant in determining the existence of price-fixing when it helps us understand whether additional conduct, such as communications or facilitating practices, significantly increases the likelihood that a collusive outcome will occur.

In contrast, wherever antitrust law is applicable, it is important to consider the influence of expected sanctions on firms' behavior. Harrington (2004a, 2004b, 2005) introduces enforcement policy into oligopoly supergames. He posits that a newly formed cartel will be more likely to attract the attention of antitrust enforcers (perhaps based on complaints by customers) if it rapidly raises price from the competitive level to the cartel level. He shows how the price path adopted by the cartel and the steady-state cartel price are affected by antitrust enforcement. He also studies the relationship between damages rules in price-fixing cases and cartel pricing. In the process, he identifies some complex and even perverse effects of antitrust enforcement on cartel pricing.

3.6.2. Determinants of the effectiveness of antitrust enforcement

A number of other aspects of antitrust enforcement have recently been illuminated by economic analysis. One increasingly active approach to enforcement is the government's attempt to strategically induce some colluding firms to turn on their peers. Enhanced leniency toward cooperating firms (as well as increased international cooperation) led to successful prosecutions in a series of major international price-fixing cases during the 1990s. Harrington (2006a) discusses the impact of corporate leniency programs on collusion. See also Motta and Polo (2003) and Motta (2004, p. 194) on the European Commission's newly adopted leniency policy, and Litan and Shapiro's (2002) discussion of cartel enforcement during the 1990s.

Another important enforcement supplement that is particularly important in the United States involves private lawsuits for (treble) damages. When the Department of Justice brings a price-fixing case, there typically are immediate follow-on private actions brought by parties claiming to have been overcharged. Frequently, these cases are brought as class actions, and many have resulted in large payments. Although only direct purchasers can claim damages under U.S. federal antitrust laws, many states allow indirect purchasers to recover damages as well. In all of these settings, economists are relied upon to estimate damages for overcharges. As previously noted, the challenge they confront—determining what prices would have existed but for the illegal collusion—is closely related to the underlying analysis of collusive behavior.

There is also a growing economic literature on cartel detection—addressing what patterns of pricing, or bidding, are indicative of collusion—that is important for setting enforcement priorities, determining liability, and assessing damages. Porter (2005), Harrington (2007), and Whinston (2006, pp. 38–52) provide highly informative surveys. Bajari and Summers (2002) discuss the detection of collusion among bidders in an auction setting.

4. Horizontal mergers

The primary concern about horizontal mergers—that is, mergers between direct competitors—is that they may lead to anticompetitive price increases, either because the merged entity on its own will find it profitable to raise prices from pre-merger levels (so-called unilateral effects) or because the increase in concentration enhances the prospects for successful collusion (coordinated effects).[94] Accordingly, we begin by offering an economic analysis of these possibilities, drawing on our analysis in sections 2 and 3. Next, we briefly review empirical evidence on the actual effects of horizontal mergers.[95]

Antitrust enforcement plays an active role with regard to horizontal mergers because in the United States nontrivial mergers must be reviewed by one of the two federal authorities with overlapping jurisdiction, the Antitrust Division of the Department of Justice (DOJ) and the Federal Trade Commission (FTC). (Similar merger review takes place in other jurisdictions, such as the European Union.) These reviews are governed by the Horizontal Merger Guidelines, the current version of which was (mostly) promulgated in 1992 by the DOJ and FTC.[96] We describe the pertinent procedures, from initial filing to the agencies' analysis to court challenges and remedies. Although our focus is on the Merger Guidelines due to their current centrality, we also discuss the pertinent antitrust statutes and the evolution of horizontal merger doctrine in the courts. We pay particular attention to the role of prospective merger synergies, usually referred to in antitrust discussions as merger efficiencies. These benefits are important in determining the threshold of anticompetitive effects that must be present to challenge a merger—that is, the law implicitly presumes mergers to be advantageous to some degree—and also in offering a possible affirmative defense to a merger that otherwise would be prohibited. Furthermore, in assessing the role of efficiencies in justifying horizontal mergers, it is

[94] A price increase often serves as a proxy for other possible anticompetitive effects, such as a reduction in product quality or service or a decrease in the pace of innovation.

[95] A subject related to horizontal mergers that we do not consider here is the tendency of partial cross-ownership to soften competition and thus increase price. See Bresnahan and Salop (1986), Reynolds and Snapp (1986), Farrell and Shapiro (1990a), O'Brien and Salop (2000), and Gilo, Moshe, and Spiegel (2006).

[96] Federal Trade Commission and U.S. Department of Justice, Horizontal Merger Guidelines (April 2, 1992) (as revised April 8, 1997 with respect to Section 4, relating to Efficiencies). See also the enforcement agencies' detailed commentary on the guidelines. FTC and DOJ (2006).

necessary to specify more precisely the goals of the antitrust laws, in particular, whether the objective is to maximize total economic welfare or instead just consumer surplus.

Finally, we consider in greater depth the economics underlying the analysis dictated by the Merger Guidelines, particularly with regard to market definition, relating the Guidelines approach to the economic analysis of market power presented in section 2. In this regard, we also discuss the growing body of empirical methods for predicting the effects of particular horizontal mergers. Part of the challenge is theoretical: given that there are a number of theories of oligopoly, with rather different predictions, which one should be used in a given merger? Presumably, the one that best fits the facts of that merger. But all of these theories are highly simplified in comparison with the inevitable complexity of real world competition, so picking the most suitable model of oligopoly is far from straightforward.

4.1. Oligopoly theory and unilateral competitive effects

The basic idea underlying theories of unilateral effects is that the merged firm will have an incentive to raise its price(s), in comparison with the pre-merger price(s), because of the elimination of direct competition between the two firms that have merged. The examination of specific oligopoly models makes it possible to quantify the effects, which is important for merger enforcement. First, quantification can help to identify the mergers that are most likely to have significant price effects and thus cause significant harm to consumers. These are the mergers that presumably warrant further scrutiny, if not prohibition. Second, quantification allows us to estimate the merger efficiencies necessary to offset the loss of competition and thereby allow the merger to pass muster according to the consumer surplus or total welfare standard.

4.1.1. Cournot model with homogeneous products

We begin by studying the effects of mergers in the Cournot oligopoly model described in subsection 2.3.1. The Cournot model seems like a good starting place since it generates a number of sensible predictions relating market structure to the equilibrium outcome. In particular, we derived equation (3) $\frac{P-MC_i}{P} = \frac{S_i}{|\varepsilon_D|}$ that relates a firm's price-cost margin to its market share and the market elasticity of demand. In the special case with constant and equal marginal costs, each firm has a market share of $1/N$, and the Cournot model predicts that the margin of each firm will be given by $(P-MC)/P = 1/N|\varepsilon_D|$. We also derived an expression for the industry-wide average, output-weighted, price-cost margin, that is, $PCM \equiv \sum_{i=1}^{N} S_i \frac{P-MC_i}{P}$, namely expression (4): $PCM = \frac{1}{|\varepsilon_D|}\sum_{i=1}^{N} S_i^2 = \frac{H}{|\varepsilon_D|}$, where, recall, $H \equiv \sum S_i^2$ is the Herfindahl-Hirschman Index (HHI) of market concentration.

The idea that a firm with a large share will have more market power, and thus will charge a higher price (but still less than the monopoly price) has been very influential in horizontal merger enforcement. So has the idea that margins are higher in more concentrated industries. In fact, based partially on the expression for the PCM, the Merger

Guidelines measure market concentration using the HHI. At the same time, it is recognized that the margins of all firms in a given market are lower if the elasticity of demand in that market as a whole is large, a subject to which we will return in discussing market definition in subsection 4.5 below.

While all of these expressions accurately characterize the Cournot equilibrium, none of them actually tells us what happens to price, consumer surplus, profits, or total welfare as a result of a merger between two firms in a Cournot oligopoly. To answer those questions, which are central to the analysis of horizontal mergers, it is necessary to compare the Cournot equilibria before and after the merger and, in particular, to specify what is involved when two formerly independent firms become one.

Salant, Switzer, and Reynolds (1983) address this question, emphasizing the peculiar result that mergers in a Cournot oligopoly can be unprofitable. The reduction in the number of firms raises price. Initially, the merging firms reduce their output because they internalize more of the effect of their output on price than they did previously. In turn, non-merging firms raise output somewhat, leading the merging firms to cut output further. At the new equilibrium, price is higher. (Indeed, this is why price is ordinarily higher in Cournot equilibrium when there are fewer firms.) But the merged firm's combined share of total industry profits is lower; after all, other firms' quantities rise and the output of the merging firms falls. Salant et al. focus on the symmetric case with constant marginal costs; in this setting, a merger of any number of firms is equivalent to all but one of the merging firms shutting down. As a result, the cost due to the smaller profit share will exceed the benefit from a higher industry price unless, in their example, the merging firms constitute 80% or more of the industry! In this simple model, a merger does not lead to a "stronger" firm in any sense—as noted, it is as if the acquired firm simply exits. If this story depicted how mergers work, few mergers (short of mergers to monopoly) would be observed. Accordingly, a theory that plausibly explains mergers that actually occur requires that the merging firms own assets that can be usefully combined in some way.

Perry and Porter (1985) pursue this point using a model in which each firm owns a certain amount of capital. In their model, each firm's marginal cost increases linearly with that firm's output, and the slope of the marginal cost curve is lower, the larger is the firm's capital stock. Thus, firms that own more capital are larger in the resulting Cournot equilibrium. Perry and Porter assume that when two firms merge, the merged entity owns their combined capital stock and thus has a lower marginal cost curve than either of the constituent firms. In addition, since the marginal cost of each rival firm rises with its output, the ability of rival firms to expand in response to the merger is not as great as in the prior example in which marginal cost is constant. As a result, horizontal mergers are much more likely to be profitable in this model. Levin (1990) generalizes the Salant, Switzer, and Reynolds (1983) model in a different direction by allowing the firms to differ in their (constant) marginal costs.[97] McAfee and Williams

[97] Levin also allows the merged firm to behave other than as a Cournot oligopolist, for example, as a Stackelberg leader.

(1992) further explore models with quadratic cost functions where the marginal cost of a firm is proportional to the ratio of its output to its capital stock, showing how the magnitude of the price increase resulting from a merger depends on the capital stocks of the merging and non-merging firms.

Farrell and Shapiro (1990b) significantly generalize these results and provide an analysis of the price and welfare effects of horizontal mergers in Cournot oligopoly. They start with a Cournot equilibrium among N firms, where the cost function of firm i is given, as before, by $C_i(X_i)$. A merger in Cournot oligopoly can be modeled as the replacement of two existing firms with cost functions $C_1(X_1)$ and $C_2(X_2)$ by a single merged firm with its own, new cost function, $C_{12}(X_{12})$.

Farrell and Shapiro say that a merger generates no synergies if the merger simply allows the merging firms to rationalize output between their existing operations or facilities, that is, if $C_{12}(X_{12}) = \min_{X_1, X_2} [C_1(X_1) + C_2(X_2)]$ subject to $X_1 + X_2 = X_{12}$. Define the pre-merger outputs of the two merging firms as $\overline{X}_1$ and $\overline{X}_2$, and the pre-merger price as $\overline{P}$. Label the two merging firms so that firm 2's pre-merger output is at least as large as firm 1's pre-merger output, $\overline{X}_2 \geq \overline{X}_1$. Using the pre-merger Cournot equilibrium relationship (3), $\frac{\overline{P}-MC_i}{\overline{P}} = \frac{S_i}{|\varepsilon_D|}$, we know that larger firms have higher markups, so firm 1's marginal cost in the pre-merger equilibrium, $\overline{MC_1} = MC_1(\overline{X}_1)$, is at least as large as firm 2's, $\overline{MC_2} = MC_2(\overline{X}_2)$. Denote the merged firm's marginal cost at the combined pre-merger output by $\overline{MC_{12}} = MC_{12}(\overline{X}_1 + \overline{X}_2)$.

Using this framework, Farrell and Shapiro prove generally the important result that mergers generating no synergies raise price. Without synergies, the merged firm's ability to rationalize production between its existing operations (by equating the marginal cost of production in the two operations) is not sufficient to offset the incentive to raise price that results from combining the ownership interests of the two operations.

Farrell and Shapiro also ask about the magnitude of synergies necessary for a horizontal merger to lead to a reduction rather than an increase in price. This is an important question in practice because, as discussed in subsection 4.4.3, mergers tend to be judged based on their impact on consumers. Farrell and Shapiro provide a very general necessary and sufficient condition: a merger reduces price if and only if $\overline{MC_2} - \overline{MC_{12}} > \overline{P} - \overline{MC_1}$. That is, the merger will reduce price if and only if the marginal cost of the merged firm (at the pre-merger combined output) is less than the marginal cost of the more efficient firm (at its own pre-merger output) by an amount that exceeds the difference between the price and the marginal cost of the smaller, less efficient firm prior to the merger. This inequality can be expressed in proportion to the pre-merger price as

$$\frac{\overline{MC_2} - \overline{MC_{12}}}{\overline{P}} > \frac{\overline{P} - \overline{MC_1}}{\overline{P}} = \frac{S_1}{|\varepsilon_D|}, \tag{6}$$

where we have added the pre-merger relationship between firm 1's margin and its share. This is a very demanding condition in an industry with moderate to large pre-merger margins. For example, consider a Cournot industry in which the market elasticity of

demand at the pre-merger price is $\varepsilon_D = -1.0$, normalize the pre-merger price at $\overline{P} = 100$, and suppose that the pre-merger market shares of the two firms are 10% and 30%, so $S_1 = 0.1$ and $S_2 = 0.3$. Using the pre-merger Cournot equilibrium conditions, the pre-merger marginal costs of the two firms must be 90 and 70 respectively. The inequality above tells us that the merger will lower price if and only if the marginal cost of the merged firm, at the combined output of two merging firms, is less than 60.

Using these general results, Froeb and Werden (1998) provide calculations that relate the required magnitude of the synergies to the pre-merger shares of the merging firms. In the symmetric case, they show that the proportionate reduction in marginal cost necessary for price not to rise is equal to $S/(|\varepsilon_D| - S)$, where S is the pre-merger market share of each merging firm.

Analyzing the welfare impact of such mergers is more complex, in part because welfare effects depend heavily on the cost function of the merged entity in comparison with the cost functions of the two constituent firms, which captures any synergies resulting from the merger. However, Farrell and Shapiro (1990b) are able to obtain general results about the "external" effect of the merger, that is, the combined effect of the merger on consumers and rivals. If we are prepared to presume that a proposed merger raises the combined profits of the merging firms (for otherwise they would not choose to merge), then any merger that generates positive external effects must raise total welfare. For a range of demand and cost conditions, Farrell and Shapiro provide an upper bound on the combined share of the merging firms such that their merger must generate positive external effects. If the combined share of the merging firms is small, they will not find it profitable to restrict output much, if at all; when they do restrict output the larger firms are likely to expand, and shifting output toward larger firms actually boosts welfare, since the larger firms have lower pre-merger marginal costs. This approach has the significant virtue that it does not involve an inquiry into the efficiencies generated by the merger, which can be difficult to quantity and verify, as we discuss below.

Until now, we have examined the effects of mergers on price and welfare but have not related this analysis to the effect of the merger on industry concentration, a typical focus of horizontal merger enforcement policy (as reflected in the Merger Guidelines). Specifically, concern is typically thought to be greater, the higher is pre-merger concentration and the greater is the merger-induced increase in concentration, notably, as measured by the HHI. Farrell and Shapiro show, however, that increases in the HHI may well increase total welfare. In particular, they show that, starting from a Cournot equilibrium, an arbitrary small change in the outputs of all of the firms raises welfare if and only if $\frac{dX}{X} + \frac{1}{2}\frac{dH}{H} > 0$, where X, as before, is industry output. Naturally, an increase in output raises welfare, since price is above marginal cost for all of the firms. More surprisingly, for a given change in total output, welfare is higher the greater is the change in concentration. Why? Each firm's price-cost margin is proportional to its market share, so the larger firms have higher margins and thus lower marginal costs. As a result, shifting output toward them, which raises concentration, raises welfare as well. This observation tells us that an increase in concentration cannot serve as a proxy

for a decrease in total welfare when studying horizontal mergers.[98] (It should be noted in this regard that, ordinarily, when enforcement agencies and courts consider increases in concentration, this is viewed diagnostically and prospectively under the maintained assumption that the share of the merged firm will equal the combined pre-merger shares of the merging firms.)

The applicability of the Cournot model is limited to industries where competition is accurately modeled as a quantity-setting game, or perhaps as a capacity-setting game followed by pricing competition, with fairly homogeneous products, and where the predictions of the one-shot Cournot model (rather than a model of repeated Cournot) fit the industry reasonably well. The Cournot model is not suitable for industries with highly differentiated products, especially if capacity constraints are unimportant in the medium to long run. In those industries, a Bertrand model with differentiated products fits better. We now study mergers in that model.

4.1.2. Bertrand model with differentiated products

A very extensive literature has developed to explore the effects of horizontal mergers in models of Bertrand competition with differentiated products.[99] These models are extensively used in practice to estimate and simulate the effects of proposed mergers, particularly in markets with branded products, ranging from consumer goods such as breakfast cereal to computer software.

Deneckere and Davidson (1985) provide a nice entry point into this literature. In contrast to the results of Salant, Switzer, and Reynolds (1983), they find that mergers are always profitable and will always involve price increases. Prior to the merger, the price of each product was set to maximize the profits earned on that product, given the prices of all other products. Now consider what happens if the price of one of the merging products, say product 1, is raised slightly. This will lower the profits earned on product 1, but the first-order effect will be zero since the price of product 1 was already optimized. The higher price for product 1 will, however, increase sales of product 2, thus raising the profits of the merged firm (a positive externality that firm 1 ignored prior to the merger). The increase in profits from product 2 will be larger, the greater is the increase in sales of product 2 that results from the increase in the price of product 1 and the larger is the price-cost margin on product 2. What about changes in the prices set by the other firms? In Bertrand equilibrium, best-response curves slope upwards, so the other firms will find it optimal to raise their prices in response to the higher price for product 1 (and for product 2, the price of which it will also be profitable to increase). These higher prices for other products increase the demand for products 1 and 2, further adding to the profits of the merged firm, which prospectively makes the

[98] Farrell and Shapiro (1990a) show more generally how changes in the ownership of assets in Cournot oligopoly affect output, welfare, and the HHI.

[99] See Ivaldi et al. (2003b), Motta (2004, pp. 243–265), and especially Werden and Froeb (2007) for more extensive reviews of this literature. Baker and Bresnahan (1985) is an important early contribution.

merger even more attractive.[100] Note also that each non-merging firm welcomes the merger since it earns higher profits because, as explained, the merged firm charges higher prices for both of its products, which increases the demand for the rival products.

These ideas are very general: in models with differentiated products and Bertrand competition, mergers that involve no synergies are profitable for the merging firms, raise the prices charged by the merging firms, and raise the price and profits of the non-merging firms as well. Clearly, such mergers lower consumer surplus; they also tend to lower welfare. It is possible that such mergers raise welfare, however, if they involve significant synergies or if the merging firms are inefficient, so shifting output away from them and toward the other firms is efficient.

To apply these ideas in practice, where the emphasis tends to be on whether, and how much, a proposed merger will raise price, it is helpful to understand what economic variables tend to make the price effects of a merger between two suppliers of differentiated products large or small. We return to this issue below, where we discuss the sophisticated simulation methods now used to estimate the price effects of such mergers.

A good sense of the basic forces at work can be gleaned by comparing the prices in a Bertrand duopoly with two differentiated products, each sold by one firm, with the price charged by a single firm selling both products. Focusing on just two products is not as restrictive as it might appear: one can interpret the demand functions for the two products in this model as demand in a general oligopolistic market, taking as given the prices of all of the other products. In the absence of any efficiencies, the logic of Deneckere and Davidson (1985) tells us that the merged firm will have an incentive to raise its price, given the prices of the other firms, and that the optimal price for the merged firm, given those other prices, is less than the new Bertrand equilibrium price once one accounts for the price increases by the other firms. Therefore, the price increases calculated using a duopoly model will (somewhat) underestimate the price increases in the full oligopoly model.

We derived a formula in subsection 2.3.2 for the difference between the monopoly price and the Bertrand equilibrium price in a simple, symmetric Bertrand duopoly model with linear demand and constant marginal cost. Following Shapiro (1996), we showed that the percentage gap between the monopoly price and the Bertrand price is given by $\frac{P_M - P_B}{P_B} = \frac{\alpha}{2(1-\alpha)} \frac{P_B - MC}{P_B}$, where $\alpha \equiv \frac{dX_2}{dP_1} / |\frac{dX_1}{dP_1}|$ is the diversion ratio, that is, the fraction of the lost unit sales of product 1, when the price of product 1 is raised, that are captured as unit sales of product 2, as previously defined in subsection 2.3.2. If we define the pre-merger price-cost margin as $\overline{m} \equiv \frac{P_B - MC}{P_B}$, then, in this very simple model, the percentage price increase predicted from the merger of the two firms is $\frac{\alpha}{2(1-\alpha)} \overline{m}$.

[100] The logic in the Cournot case is different because best-response functions slope down. When the merged firm optimally reduces its output, the other firms expand output, which reduces the profits of the merged firm.

This price increase is proportional to the pre-merger price-cost margin, $\overline{m}$. This comports with intuition: since the profits gained on the sale of product 2, when the price of product 1 is raised, are proportional to the margin on product 2, the magnitude of the margin on product 2 is proportional to the incentive to increase the price of product 1 (and conversely). Therefore, *ceteris paribus*, mergers between firms selling differentiated products are likely to raise price more, the greater are the pre-merger margins on their products.

The price increase associated with the merger is also proportional to the factor $\frac{\alpha}{1-\alpha}$, which is increasing in the diversion ratio and which is zero if the diversion ratio is zero.[101] This, too, is intuitive: the greater is the diversion ratio, the greater is the share of the lost sales from product 1 that are captured by product 2 and thus internalized after the merger. Therefore, *ceteris paribus*, mergers between firms selling differentiated products are likely to raise price more, the closer is the degree of substitution between their products, as measured using the diversion ratio. Note that a high gross margin is consistent with a high diversion ratio; this pattern arises if the demand for product 1 is not very elastic and if a significant fraction of the (relatively few) sales lost when the price of product 1 rises are diverted to product 2.

As shown by Shapiro (1996), however, a rather different formula applies with constant-elasticity (rather than linear) demand. In this case, the percentage price increase predicted from the merger of the firms 1 and 2 is $\frac{\alpha\overline{m}}{1-\alpha-\overline{m}}$.[102] This ratio is larger than in the case of linear demand, and possibly much larger for plausible parameter values. To illustrate, suppose that the pre-merger gross margin is $\overline{m} = 0.35$, not an uncommon number for branded products, and that one-quarter of the sales lost when the price of product 1 is raised are captured by product 2 (and vice versa), so $\alpha = 0.25$. With these parameters, the post-merger price increase with linear demand is about 6%, while the post-merger price increase for constant-elasticity demand is nearly 22%.

Suppose that one observed the pre-merger margin of 35% and was able to estimate the diversion ratio of 25% between these two products. Both of these models—one with linear demand, one with constant elasticity of demand—can be parameterized to be consistent with these observations. Yet the two models give significantly different predictions for the price increase associated with a merger because the two demand systems diverge somewhat as prices depart from their pre-merger equilibrium levels. This should not be totally surprising: mergers are discrete events, and if nontrivial price changes are possible, their magnitude must in fact depend upon demand at prices distinctly different from the pre-merger prices.

All of this tells us that, in a merger involving differentiated products, making reliable predictions of unilateral price effects based on a model of Bertrand oligopoly requires an accurate structural model of the demand system, and that the shape of the demand system at prices some distance away from the pre-merger equilibrium affects

[101] We require $\alpha < 1$ or else the merged entity faces perfectly inelastic demand at all positive prices.

[102] We require $1 - \alpha - \overline{m} > 0$ so that the elasticity of demand facing the merged firm is greater than unity.

the post-merger price increase.[103] If a structural model can be estimated that fits industry demand, then, with luck, the post-merger equilibrium can be simulated using that model, thereby predicting the magnitude of post-merger price increases. This promising approach has been explored analytically and applied in practice in recent years, as we discuss in subsection 4.6.2. The logit model for differentiated products, championed by Werden and Froeb (2007), is especially tractable and has been used extensively to estimate the effects of horizontal mergers. In this model, each consumer picks one unit of a single brand from a set of choices that includes the differentiated products, $i = 1, \ldots, N$, along with the alternative of an outside good, which can simply be interpreted as picking none of these products. The consumer's utility from selecting brand $i = 1, \ldots, N$ is the sum of a "systematic" component associated with that brand, V_i, and an unobservable idiosyncratic component. Under suitable assumptions about the distribution of the idiosyncratic terms, the probability that a consumer will pick brand i is given by $\phi_i = e^{V_i} / \sum_{k=0}^{N} e^{V_k}$, where the index zero corresponds to the outside good. If we define $\Phi = \sum_{k=1}^{N} \phi_k$ as the probability that the consumer will pick one of the N brands (rather than the outside good), then firm i's market share is $S_i = \phi_i / \Phi$, so the market shares are proportional to the choice probabilities ϕ_i.

In the simple specification described in Werden and Froeb (2007), the systematic component of utility for brand $i = 1, \ldots, N$ is given by $V_i = \gamma_i - \beta P_i$, where γ_i reflects the underlying quality or average attractiveness of brand i, P_i is the price of brand i, and β is a constant that determines the degree of substitutability among the different products. For large values of β, the competing brands are very close substitutes, and price-cost margins are low. Differentiating the demand for brand i with respect to the price of brand i gives $d\phi_i / dP_i = -\beta\phi_i(1-\phi_i)$. Transforming this expression into elasticity form, the own-price elasticity for brand i is given by $-\beta P_i(1-\phi_i)$. Differentiating the demand for brand i with respect to the price of brand j gives $d\phi_i / dP_j = \beta\phi_i\phi_j$. Transforming this expression into elasticity form, the cross-price elasticity of demand for brand i with respect to the price of brand j is $\beta P_j \phi_j$. Therefore, the diversion ratio from brand j to brand i, when the price of brand j rises, is given by $\frac{dX_i/dP_j}{|dX_j/dP_j|} = \frac{\phi_i}{1-\phi_j}$.

This model has the attractive, but restrictive, property that the diversion ratio from brand j to brand i is proportional to firm i's market share. In this important sense, the logit model is the antithesis of spatial models in which some products are very close substitutes, others are distant substitutes, and proximity need not bear any particular relationship to popularity. The logit model is a good starting point in a situation where all of the brands compete against one another and it is not clear which are "close" to each other. Nested logit models can be used when additional information about proximity is available.

Werden and Froeb (2007) show that, in the Bertrand equilibrium with single-product firms, the gap between firm i's price and firm i's marginal cost is given by

[103] All of these ideas carry over to mergers between multi-product firms, but the pertinent calculations are more complex.

$P_i - MC_i = \frac{1}{\beta(1-\phi_i)}$. This expression tells us that the firms with more attractive products, and thus larger market shares, have higher markups, much like firms with lower costs and thus larger shares have higher margins in a Cournot equilibrium. In the Bertrand equilibrium that results after the merger of brands 1 and 2, the equilibrium gap between price and marginal cost for each of the merging brands is given by $P_1 - MC_1 = P_2 - MC_2 = \frac{1}{\beta(1-\phi_1-\phi_2)}$. (Remember, in interpreting this equation, that the market shares of the merging brands are not constants; they will fall as a result of the merger.) To illustrate, consider the situation in which product 2, say, is inherently more attractive, that is, in which $\gamma_2 > \gamma_1$. For simplicity, suppose that both products are produced at constant and equal marginal cost. Prior to the merger, product 2 would have a larger market share and a larger gap between price and marginal cost than would product 1. After the merger, the prices of both products would be higher than their pre-merger levels, and the gap between price and marginal cost for the two products will be equal, which in this case further implies that the prices will be equal; product 2 would have a larger market share than product 1. Therefore, the post-merger price increase will be larger for product 1 than for product 2. This fits with intuition: the incentive to raise the price of product 1 is greater since a relatively large fraction of its sales will be diverted to product 2, due to that product's popularity. Furthermore, the pre-merger gap between price and marginal cost for product 2 is larger than that for product 1, so any diverted sales are actually adding to the profits of the merged entity.

The symmetric logit model with constant marginal cost can readily generate predictions about the price effects of mergers, given an estimate of the elasticity of demand for the market as a whole and an estimate of the pre-merger gaps between prices and cost. As an example, Werden and Froeb (2007) report that with six (symmetric) firms, a market elasticity of demand of -0.5, a normalized pre-merger price of \$1, and a pre-merger gap between price and marginal cost of \$0.40, so marginal cost is \$0.60 (all this corresponding to $\beta \approx 2.9$), the merger of any two brands causes their prices to increase by about 6%. As they note, the logit model, with its lack of localization in competition, shows that a merger between two brands can easily raise price significantly even if the merging brands are not each other's next closest substitutes in any market-wide sense. Prices rise because, with only six firms, there are a nontrivial fraction of consumers for whom the merging brands are the first and second choices. In this model, the merged firm cannot identify and price discriminate against those consumers, so the merged firm raises price somewhat to all consumers.[104]

Until now, the analysis has focused on the price effects of mergers that involve no production synergies. The consideration of efficiencies is facilitated by a convenient feature of models of Bertrand competition with differentiated products: the magnitude of the efficiencies necessary for a merger to reduce rather than raise price depends only

[104] If price discrimination were possible, the merged firm would raise price much more to the identifiable customers who ranked product 1 and product 2 as their first and second choice, and not at all to other customers. Effectively, one can compute a new post-merger equilibrium for each identifiable customer or group.

upon the shape of the demand system (the diversion ratio between the merging products) at prices in the immediate neighborhood of the pre-merger equilibrium prices. A reduction in marginal cost of product 1 at the merged firm increases the gap between the firm's price and marginal cost on that product, giving the firm an incentive to lower its price. Price will in fact fall if this incentive is stronger than the incentive to raise the price of product 1 based on internalizing the diversion to product 2, now owned by the same firm. Since both of these effects are evaluated at the pre-merger prices, no information is required about the shape of the demand system at other prices.

Based on this logic, Werden (1996) derives an expression for the cost reductions necessary to prevent a merger from raising price.[105] In the symmetric case, where the two merging firms have equal market shares and gross margins prior to the merger, he shows that a merger will reduce price if and only if the cost reduction satisfies

$$\frac{MC_1 - MC_{12}}{MC_1} > \frac{\overline{m}}{1-\overline{m}}\frac{\alpha}{1-\alpha}, \tag{7}$$

where $\overline{m} \equiv \frac{\overline{P_1}-\overline{MC_1}}{\overline{P_1}}$ is again the pre-merger margin. This is a rather stringent condition in mergers between close rivals. Using our previous numerical example of $\overline{m} = 0.35$ and $\alpha = 0.25$, the merger must reduce marginal cost by about 18% to lead to a price reduction rather than a price increase. Note that reductions in fixed cost have no bearing on (short-run) price effects.

4.1.3. Bidding models

In the Bertrand model, each firm sets a price, and buyers make their purchasing decisions given these prices. Bertrand models are especially well-suited for markets with differentiated consumer products in which there are a large number of relatively passive consumers.[106] In many other settings, however, there are large buyers who behave strategically, designing their procurement procedures so they can obtain the best price from their suppliers. In these settings, competition typically takes the form of bidding to win the business of a single customer who has designed a procurement procedure.

Many purchasing situations fit this pattern, including procurement auctions. The precise manner in which competition takes place depends upon the auction rules established by the customer. Klemperer (2004) provides an excellent overview of the enormous literature on auctions. Werden and Froeb (2007) discuss merger analysis in a situation where a seller is auctioning off an item using an ascending oral auction and the bidders have private values for the item. (Precisely the same ideas would arise in a

[105] This is the analogue in a Bertrand model of the necessary and sufficient condition for a merger to reduce price in Cournot oligopoly derived by Farrell and Shapiro (1990b).

[106] Even in markets for branded consumer products, large buyers such as large retailers may play a significant role. These buyers may be more active and strategic in dealing with manufacturers, in part by setting up bidding contests among their would-be suppliers.

situation where a buyer is running a procurement auction and the bidders are suppliers who differ in their costs of serving the buyer.) This auction format is equivalent to a second-price sealed-bid auction; it is a dominant strategy for each bidder to bid up to its value, and the price ultimately paid, P, is equal to the second-highest valuation among the bidders.

In this context, consider a merger between bidder 1 and bidder 2, and label the two bidders so that bidder 1's valuation, B_1, is at least as large as bidder 2's valuation, B_2. Label bidder 3 as the one with the highest valuation, B_3, among the other bidders. A merger between bidder 1 and bidder 2 (which is equivalent here to collusion between these two bidders) will have no effect on the price paid for the item unless $B_2 > B_3$, that is, unless the two merging bidders have the two highest valuations on the item. If they do, price will fall from B_2 to B_3. Viewed statistically, merger effects depend on the joint distribution of the valuations of the bidders, including the merging bidders. Waehrer and Perry (2003) show how the price effect of a merger can be estimated for certain cumulative distributions of valuations.

4.2. *Oligopoly theory and coordinated effects*

Mergers also can pose a risk to competition by increasing the likelihood that a collusive outcome will prevail. Such coordinated-effects theories of harm from horizontal mergers are featured in the Merger Guidelines, which state in §2.1: "A merger may diminish competition by enabling the firms selling in the relevant market more likely, more successfully, or more completely to engage in coordinated interaction that harms consumers." Merger enforcement based on coordinated effects is more important, the more one believes that increased concentration contributes to coordinated outcomes and the less one believes that collusive behavior is readily deterred by antitrust law.

As discussed in section 3, collusion is generally thought to be easier to achieve and sustain when there are fewer suppliers in the industry. Therefore, at the simplest level, reducing the number of competitors by one tends to increase the likelihood of collusion. This idea underlies what is referred to as the "structural presumption"— that increases in concentration lead to less competitive interactions—that has long played a central role in antitrust. The heyday of the structural presumption corresponded with a time when industrial organization economists devoted substantial efforts to validating empirically the core idea of the structure-conduct-performance paradigm: markets that are highly concentrated tend to have higher prices and higher profits, and thus tend to serve consumers less well, than do markets with more competitive structures, *ceteris paribus*.

Demsetz (1973) mounted a strong attack on those who claimed that a positive cross-sectional relationship between concentration and profits was indicative of market failure or the need for an interventionist antitrust policy. Demsetz pointed out that a positive correlation would also arise if some firms were more efficient than their rivals, and if the more efficient firms had large market shares. Market concentration would then result from the presence of large, efficient firms. Under this hypothesis, small firms in concentrated market would earn normal profits, with the large, efficient firm earning

profits due to Ricardian rents. If margins are associated with a firm's market share, not overall market concentration, this may well reflect the greater efficiency of larger firms, at least in the short run. The implications for merger policy are profound: if a large firm seeks to buy a smaller rival, the resulting increase in concentration might go along with lower prices and consumer benefits if the large and efficient firm is able to improve the efficiency with which the assets of the smaller acquired firm are used. Bork (1978) is also well known for attacking the presumptions under merger policies of the 1960s and 1970s.

Reviewing the enormous literature on the cross-sectional industry relationships among concentration, prices, margins, and profits in order to distinguish among these competing hypotheses is beyond the scope of this chapter. Schmalensee (1989) and Salinger (1990) are good starting places for readers interested in learning more. Pautler (2003) provides a more recent summary of the literature, which has made progress in distinguishing effects on a firm's profits that are related to market concentration, the firm's market share, or the firm's identity (looking across multiple markets). Overall, economists have grown less confident over the past several decades in stating that there is a systematic relationship between market concentration and market performance, at least over the range of market structures in which there are more than two or three firms. Even so, the cautionary statement made by Salinger (1990, p. 287) bears repeating today:

> First, despite the well-known problems with this literature, it continues to affect antitrust policy. The inappropriate inferences used to justify an active antitrust policy have given way to equally incorrect inferences that have been used to justify a relaxed merger policy. Second, the alternative to cross-industry studies is to study specific industries. . . . [I]t is important to realize that it was the failure of studies of individual industries to yield general insights that made cross-industry studies popular.

Whatever one thinks of this literature, one should bear in mind that these cross-industry studies do not directly measure the effects of horizontal mergers, which we take up in subsection 4.3. The primary variation studied is across industries, not within an industry over time. Furthermore, through the early 1980s, highly concentrating horizontal mergers would simply not have been allowed. So, to the extent that one sees efficient larger firms in certain industries, through at least the early 1980s these firms mostly arose through internal growth, non-horizontal mergers, or horizontal mergers involving firms with relatively small market shares, not through highly concentrating horizontal mergers.

The key question regarding coordinated effects in merger analysis is whether a given merger will significantly increase the likelihood that a collusive outcome will arise.[107]

[107] Concern would also arise if the merger makes collusion more effective, for example, by raising the price at which collusion can be maintained to a level closer to the monopoly price or be reducing the frequency and duration of price wars. For simplicity, in our discussion below we use the shorthand of talking about the likelihood that a collusive outcome will arise.

In section 3, we explored in considerable detail how various industry conditions affect the likelihood of effective collusion. All of that theory and evidence can be brought to bear when considering coordinated effects in horizontal mergers. While we lack methods, such as those we just discussed regarding unilateral effects, to quantify these coordinated effects, we know quite a lot about how a change in market structure resulting from a merger will affect the likelihood of effective collusion. In principle, then, one can trade off the increased costs from potential collusion against any efficiencies associated with a merger.

Some highly relevant and robust lessons emerged from the analysis of collusion in section 3. A horizontal merger between two significant suppliers, by reducing the number of players by one, can significantly increase the likelihood that the remaining firms will be able to reach a collusive agreement. One possibility is that a merger may establish a clear market leader who can play the role of price leader, serving the function of establishing and adjusting collusive prices, with the other firms following. Perhaps most important, when a firm that would have been reluctant to join in a collusive scheme (which, as previously noted, is sometimes termed a maverick) is acquired by another supplier who is larger or otherwise more inclined to participate, collusion can be greatly facilitated. Beyond these points, a merger reduces the number of bilateral links between firms in a market, which is some measure of the difficulty of reaching an agreement. With N suppliers, the number of such links is $N(N-1)/2$. A 5-to-4 merger reduces the number of links from 10 to 6; a 4-to-3 merger reduces the number of links from 6 down to 3.

For similar reasons, horizontal mergers also can make it easier to sustain a collusive outcome. A firm with a larger market share tends to have less to gain from cheating and more to lose if a price war erupts than do smaller firms. As a result, the merger of two smaller firms may increase the price at which collusion can be sustained. In general, a merger that significantly increases concentration will tend to make cheating on the collusive price less attractive, at least for the merging parties.

These observations are surely important for merger enforcement policy, even if our knowledge about the relationship between industry conditions and the likelihood of collusion does not give us a specific quantitative procedure to weigh the increased danger of collusion in, say, a 4-to-3 merger against efficiencies promised by that merger. However, a paradox of proof (different from the one that we noted in subsection 3.4.2) can present some problems when one seeks to apply collusion theory to horizontal mergers. To illustrate with an overly sharp example, suppose that one concludes in a given industry that effective collusion is quite unlikely if there are five or more firms, possible but not likely if there are four firms, and quite likely if there are three or fewer firms. Concerns about coordinated effects would therefore be minimal for any merger that left at least five firms in the industry. A merger from 5 to 4 firms would be a cause for concern, as would be a merger from 4 to 3 firms. But even more concentrating mergers, from 3 to 2 firms, and perhaps even a merger to monopoly, would cause fewer concerns: collusion is hypothesized to be likely with or without these mergers. While this is surely too strong a conclusion—even with only two firms, there probably is a nontrivial chance

that collusion will break down—this logic at least undermines the standard presumption that mergers become more worrisome as the number of firms declines. This additional paradox is avoided only if one believes that the probability of successful collusion is not just declining in the number of firms but also is a convex function of the number of firms.

There is relatively little formal theory exploring the implications for merger policy of the relationship between collusion and market concentration, apart from the papers already discussed in section 3. But several of the them are especially pertinent for evaluating coordinated effects in horizontal mergers. Notably, Compte, Jenny, and Rey (2002) and Vasconcelos (2005) ask how the distribution of capacities affects the ability of the firms to sustain collusion in price-setting and quantity-setting supergames, respectively. Davidson and Deneckere (1984) point out that reverting to the static Nash equilibrium typically is a less severe punishment when there are fewer firms (in a quantity-setting supergame or a price-setting supergame with capacity constraints), making for a complex relationship between market concentration and the likelihood of collusion.

Kovacic et al. (2006) propose an interesting new way to quantify the dangers associated with coordinated effects in a situation where a number of suppliers are bidding for the customer's patronage. They propose measuring the effects of incremental collusion, that is, collusion that only involves two firms, before and after the proposed merger. They show how this calculation can be performed in a particular bidding model. While a large number of calculations are necessary to implement their method, these calculations are all well rooted in oligopoly theory, and in fact use the results already discussed in the analysis of unilateral effects.

Baker (2002) has emphasized the important role of maverick firms in disrupting or preventing collusion and thus the particular dangers that arise when a merger eliminates such a firm (an idea embraced in the Merger Guidelines as well). Collusion theory indicates that reaching an agreement and sustaining an agreement may be difficult if one of the firms expects to gain significant market share in the absence of collusion. Therefore, firms with strategies, products, or costs that are distinct from those of their rivals, and firms that are optimistic and growing rapidly, perhaps because they recently entered the market, are obvious candidates to be mavericks. Accordingly, Baker advocates an approach to merger enforcement policy that goes beyond the measurement of increases in market concentration by emphasizing the identification of mavericks. He argues that placing the focus on identifying maverick firms will reduce judicial errors by allowing the enforcement agencies and the courts to identify more accurately those mergers that are likely to have coordinated anticompetitive effects for any given level of and change in market concentration. He also notes that a merger may actually create a new maverick.

4.3. Empirical evidence on the effects of horizontal mergers

Given the large number of mergers that are consummated every year, including many horizontal mergers, one might think that there would be extensive, definitive evidence

regarding the effects of these mergers. Under what circumstances have horizontal mergers been found to raise or lower prices, or more generally to benefit or harm consumers? And what has their impact been on the profits of the merging parties and on the profits of their rivals?

Sadly, there is no such clear and definitive body of evidence. To some extent, this reflects a lack of data: even in those cases where one can accurately measure the prices charged before and after a merger, it may be hard to attribute price changes to the merger rather than to other changes in industry conditions. Also, the effects of a merger may arise in non-price dimensions such as product quality, customer service, or innovation. Furthermore, if merger enforcement policy is working well, the mergers most likely to have large adverse price effects are never proposed or are blocked on antitrust grounds. We do not mean to suggest that it is impossible to identify the effects of horizontal mergers; but nor is it easy. See FTC (2002) for some recent evidence.

Pautler (2003) offers an extensive review of empirical work on the effects of mergers and acquisitions. Readers interested in exploring this literature in greater detail should turn to his paper, which contains a treasure trove of information on the subject. Whinston (2006, pp. 110–127) also provides a valuable discussion of the evidence. We examine here several distinct methods for identifying and measuring the effects of mergers. In evaluating this evidence, one should bear in mind that over the past 25 years, only about 2% to 4% of the mergers reported every year under the Hart-Scott-Rodino Act were considered to raise sufficient antitrust issues to warrant a second request from the FTC or the DOJ, so data on the effects of all mergers may not reflect the effects of the major horizontal mergers that are most likely to be scrutinized by antitrust authorities.[108]

4.3.1. Stock market prices

One way to measure the effects of mergers is to study the stock market performance of the merging firms. Usually, this is done using an event study around the time of the announcement of the merger. This approach has been extensively explored in the finance literature. Andrade, Mitchell, and Stafford (2001) provide an excellent introduction. The advantage of this approach is that it relies on detailed and accurate stock market data. However, by its nature, this approach cannot distinguish between favorable stock market returns based on efficiencies versus market power.[109] In addition, this approach

[108] Pautler (2003) provides data on FTC and DOJ second requests. See FTC and DOJ (2003) and DOJ (2006) for more detailed recent data on merger challenges. Leary (2002) also provides some data on merger enforcement activities and merger activity. Baker and Shapiro (in press) update these data and comment on the interpretation of enforcement data.

[109] In principle, a merger that would lead to synergies and lower prices would depress the stock market value of rivals, while an anticompetitive merger that would lead to higher prices through unilateral or coordinated effects would boost the stock market value of rivals. Pautler (2003) reviews studies that attempt to measure the impact of horizontal mergers on the stock price of rivals. Such effects are more difficult to measure

measures the expectations of investors about merger effects, not the actual effects of mergers. Furthermore, this literature is not focused on horizontal mergers. Thus, the finance literature is best seen as addressing a more general question: do mergers and acquisitions produce wealth for shareholders or do they reflect managerial hubris? Finally, event studies do not readily disentangle predicted effects of the merger and other information that may be signaled by the announcement.

Andrade, Mitchell, and Stafford (2001) report abnormal negative returns for acquiring firms, based on 1864 deals from the 1990s: 1.0% during a three-day window around the announcement and 3.9% during a longer window from 20 days prior to the announcement through closing of the deal. However, target firms showed a 16% abnormal positive return during the three-day window. The combined firms gained about 1.5% over the short or longer window. They also report several studies that found negative abnormal returns over the three to five years following the completion of mergers, stating (p. 112): "In fact, some authors find that the long-term negative drift in acquiring firm stock prices overwhelms the positive combined stock price reaction at announcement, making the net wealth effect negative." However, Andrade et al. are skeptical of these results, disputing the reliability of these longer-term studies, in part since it is hard to know what the "normal" return should be over these longer periods of time.

In the end, Andrade et al. state (p. 117): "We are inclined to defend the traditional view that mergers improve efficiency and that the gains to shareholders at merger announcement accurately reflect improved expectations of future cash flow performance. But the conclusion must be defended from several recent challenges." One of these challenges arises from the fact that the source of the stock market gains to the combined firms from mergers has not been identified. In the case of horizontal mergers, at least, those gains could well come from enhanced market power. Another challenge arises because acquiring firms do not appear to benefit from mergers, which at the least is an uncomfortable fact for those who believe in a reasonably efficient stock market. In fact, there is some evidence that many mergers involve managerial hubris or empire building. Bargeron et al. (2007) find that public firms pay a 55% higher premium to targets than do private acquirers. Harford and Li (2007) find that in mergers that leave acquiring firm shareholders worse off, bidders' CEOs are better off 75% of the time. This issue will be important below when we consider merger synergies: if there truly are unique synergies resulting from the merger, why do acquiring firms fail to capture any of these gains from trade?

4.3.2. Accounting measures of firm performance

A second method for measuring the effects of mergers is to study accounting data for the firms involved to look for changes in various measures, such as rates of return,

reliably than are effects on the stock market value of the merging parties, especially if the rivals are diversified companies with a relatively small share of their revenues coming from the sale of products in markets where the merging firms are significant horizontal rivals.

cash flows, or profit margins. Ravenscraft and Scherer (1987, 1989), using widely cited FTC Line of Business Data, reach rather negative conclusions: many of the mergers and acquisitions they study were unsuccessful, leading to a decline in the post-merger profitability of the acquired line of business. Their study supports the view of excessive managerial zeal about acquisitions. However, they mostly examine conglomerate mergers, not horizontal mergers, so much of their evidence is not directly relevant to horizontal merger control policy. Also, they find that horizontal mergers tended to be more profitable than conglomerate mergers (although, again, this result does not distinguish market power from the possibility of greater synergies in horizontal mergers). Healy, Palepu, and Ruback (1992) examine post-merger operating performance for the fifty largest mergers that took place from 1979 to 1984. They find that the merged firms exhibited improved operating performance, as measured by operating cash flows, relative to their industry peers. They attribute these gains to increased operating efficiency. Along similar lines, Lichtenberg and Siegel (1987) and McGuckin and Nguyen (1995) find plant level productivity gains associated with mergers in manufacturing industries, using the Census Bureau's Longitudinal Establishment Data for 1972–1981. This was not a period, however, when highly concentrating horizontal mergers were permitted by antitrust enforcers.

4.3.3. Case studies

A third approach is to study specific mergers, tracking the firms or industries involved, looking at such measures as prices, output, product quality, or R&D intensity. In principle, one can also try to measure the impact of a merger on rivals or customers. Kaplan (2000) provides a useful collection of case studies of mergers in a diverse set of industries, including hospitals, tires, banks, pharmaceutical drugs, airlines, and oil field services. The cases studied were not selected specifically to shed light on major horizontal mergers. These studies illustrate the great variety of fact patterns that arise in merger analysis, the important role of mergers as a means by which industry participants adjust to changing market conditions (making it especially hard to distinguish the effects of mergers from other changes taking place in the industry, especially once one recognizes that firms self-select to participate in mergers), and the risks as well as opportunities associated with mergers.

For antitrust purposes, it is most useful to study horizontal mergers that raised serious antitrust concerns when proposed but ultimately went forward. This approach has the virtue of focusing attention on the very small fraction of all mergers that are most relevant for assessing merger control policy.

Airline mergers have received a great deal of attention, in no small part because good data on fares are available and one can use fares on other routes as a good benchmark when measuring the effects of mergers on fares. Borenstein (1990), Werden, Joskow, and Johnson (1991), and Peters (2003) study two airline mergers from the mid-1980s that were approved by the Department of Transportation over the objections of the DOJ: the merger of Northwest Airlines with Republic Airlines, and the merger of Trans

World Airlines (TWA) with Ozark Airlines. These mergers raised significant antitrust issues because they combined directly competing hubs: Northwest and Republic both had hubs at Minneapolis, and TWA and Ozark both had hubs at St. Louis. Borenstein (1990) found significant fare increases following the Northwest/Republic merger but not following the TWA/Ozark merger. Werden, Joskow, and Johnson (1991) found that the Northwest/Republic merger raised fares by about 5% and the TWA/Ozark merger raised fares by about 1.5%, and that both mergers led to significant service reductions. Kim and Singal (1993) examine fourteen airline mergers from the mid-1980s. They compare price changes on the routes served by the merging firms with price changes on other routes of the same distance and conclude that any efficiency gains in mergers between rival airlines were more than offset by enhanced market power, leading to fares that averaged 10% higher after six to nine months. Fare increases were especially large for mergers involving airlines in bankruptcy, which had unusually low (perhaps unsustainably low) pre-merger fares.

The banking industry is another industry in which good price data are available and many horizontal mergers have occurred, making it possible to measure the price effects of horizontal mergers. Prager and Hannan (1998) study the effects of major horizontal mergers in the U.S. banking industry during the early 1990s. They look at changes in interest rates paid on deposits for several types of deposit accounts, using monthly data. They define "substantial horizontal mergers" as those that increase the HHI by more than 200 points to a post-merger value greater than 1800. They find that substantial horizontal mergers reduce the deposit interest rates offered by the merging banks.

The price and quality effects of hospital industry mergers have been examined in a number of studies, as described in Pautler (2003). For example, Vita and Sacher (2001) find large price increases, not reflecting increases in service quality, following a hospital merger in Santa Cruz, California.

Recent papers look at other industries as well. Pesendorfer (2003) studies the effect of horizontal mergers in the paper industry on capacity choices. Hastings (2004) looks at pricing in the retail gasoline market in Southern California.

One natural way to gain information to inform horizontal merger policy would be for the antitrust enforcement agencies to perform retrospective studies on the deals that they have investigated closely but ultimately allowed to proceed without significant divestitures. Neither the FTC nor the DOJ has officially reported results from any such study, at least in recent years.[110] Barton and Sherman (1984) do report price increases from a highly concentrating merger that was challenged by the FTC several years after it was consummated.[111] In addition, the U.K. Office of Fair Trading, in conjunction with the Department of Trade and Industry and the U.K. Competition Commission, sponsored a study of ten mergers that took place during 1990–2002. These were mergers that the Office of Fair Trading had reviewed and found to raise sufficient competition issues

[110] FTC (1999) reports on a study designed to determine the efficacy of the divestitures it had negotiated.
[111] The acquiring company was Xidex, and the products involved were types of duplicating microfilm.

that they were worthy of referral to the Competition Commission but that the Competition Commission had subsequently cleared. See Office of Fair Trading (2005). Based on interviews with customers of the merging firms, this study did not find a significant lessening of competition in eight of the ten cases studied. In the other two cases, a short-term loss of competition was found to have been corrected by subsequent entry into the market.

4.4. Antitrust law on horizontal mergers

This subsection outlines current U.S. antitrust law on horizontal mergers. We start with a brief statutory background and an explanation of procedures, emphasizing pre-merger notification and analysis by enforcement agencies. We then discuss the substantive law regarding requisite anticompetitive effects and whether merger synergies may be offered to defend otherwise anticompetitive mergers.

Throughout the discussion, it is useful to keep in mind the relationship between merger law, on one hand, and the law concerning price-fixing and monopolization, on the other hand. Because collusion is difficult to detect and prosecute (and, depending on the means of collusion, is of uncertain illegality), as discussed in section 3, it makes sense to some degree for merger policy to adopt a prophylactic approach toward mergers that threaten greater cooperation among firms. Likewise, because the law on monopolization does not regulate price-setting once a merger has been validated and imposes only modest constraints on exclusionary practices, as will be discussed in section 5, there is also reason to be wary of approving a merger that threatens unilateral effects or exclusionary conduct.

4.4.1. Background and procedure

As elsewhere, our discussion will focus on antitrust law and procedures in the United States; there is a growing but incomplete convergence in how horizontal mergers are treated across jurisdictions.[112] Relevant U.S. law has three primary, overlapping provisions: Sherman Act Section 1's prohibition on any "contract, combination..., or conspiracy in restraint of trade" (the focus in section 3 on collusion); Clayton Act Section 7's prohibition on acquisitions of stock or assets whose effect "may be substantially to lessen competition, or to tend to create a monopoly"; and the Federal Trade Commission

[112] In 2004, the European Union promulgated new horizontal merger guidelines that in many respects are similar to the preexisting Horizontal Merger Guidelines in the United States. European Union (2004b). Prior to that, although few mergers had been blocked, the enforcement stance of the Commission is generally regarded to have been stricter than that in the United States. The European Court of First Instance's reversal in 2002 of three Commission attempts to block mergers is seen as the catalyst for the recent reform. Other notable administrative changes include the appointment of a chief competition economist. See, for example, Dabbah (2004). It is too early to tell just how much practice under the new regime will differ in fact from that in the United States.

Act Section 5's prohibition on any "unfair method of competition."[113] In spite of diverse histories and statutory language, a largely unified approach to enforcement of these provisions has emerged. Notably, the DOJ and FTC (1992) in their most recent Horizontal Merger Guidelines have promulgated a single policy statement applicable regardless of the statute involved. Commentators and courts have largely taken a similar approach.[114]

Most challenges to mergers are brought by one of the two federal agencies, although mergers may also be challenged by states and private parties.[115] Since 1976, the federal procedure has taken its current form, which is similar to procedures in many other jurisdictions.[116] Firms intending to merge are required to file specified information with the pertinent agencies. Each deal is cleared to either the FTC or the DOJ. In mergers for which there is any serious prospect of a challenge, the parties usually submit substantial supplemental material. They hire a team of lawyers and economic experts (often associated with consulting firms) that typically have substantial experience in merger filings; indeed, they may have handled numerous prior mergers in the same or related industries. This team gathers and analyzes information and produces an often-elaborate study document defending the merger with regard to competitive effects and anticipated efficiencies. The goal typically is to persuade the agencies to approve the merger, and to do so promptly.

An important aspect of the procedure concerns the effects of agency delay—which arises when the agency feels that it needs additional information or must undertake more substantial independent investigation and analysis—or of an ultimate agency challenge. Even if the parties anticipate eventual approval, whether from the agency or after litigation in court, the prospect of delay will kill many deals and impose substantial costs on others. Keeping financing in line, making interim investment decisions in plant and equipment, deciding on strategic matters such as launching new products or terminating old ones, maintaining customer loyalty in the presence of uncertainty about product

[113] Observe that none of the statutes is limited to mergers per se; other forms of combination, notably including acquisitions of some or all of another firm's assets, are included. (Thus, for example, if the only two products in a market are patented, the acquisition of one of the patents by the owner of the other would be analyzed similarly to a horizontal merger.) EU regulations have a similar reach.

[114] See, for example, Areeda, Hovenkamp, and Solow (2006, vol. 4, pp. 43–44).

[115] One might think that competitors would frequently challenge mergers. However, under the doctrine of "antitrust injury," this is not ordinarily possible: competitors tend to be injured by pro-competitive mergers that lead to lower prices (deemed not the sort of injury that the antitrust laws were enacted to prevent) but helped by anticompetitive mergers (recall from subsection 4.1, for example, that unilateral effects tend to benefit non-merging parties). Relatedly, when agencies are investigating proposed mergers, they are less likely to give weight to the views of competitors on overall effects (for fear of manipulation) and more commonly seek the reactions of large purchasers (for example, health insurers, in the case of hospital mergers). In this regard, Coate and Ulrick (2005) find that the probability that the FTC takes action against mergers is higher, *ceteris paribus*, when there are customer complaints about the merger. In recent commentary, the agencies affirm that "Consumers typically are the best source, and in some cases they may be the only source, of critical information . . . ." FTC and DOJ (2006, pp. 9–10).

[116] Indeed, there has been some explicit international cooperation, motivated by the fact that many substantial mergers are subject to the competition regulation of multiple national and international jurisdictions.

support, retaining talented employees who may fear job loss, and so forth may present significant challenges to merging firms, especially those being acquired, that do not know if or when their deal may be approved. Accordingly, great energy is devoted to obtaining a quick and successful conclusion to the antitrust agency's deliberations. If the agency challenges the deal and no prompt settlement is reached, the merging parties must either abandon their transaction or confront considerable further delay in the process of litigating the matter in federal court.

To give some sense of the level of merger review activity, during the years 2001–2005 for the DOJ there were about 1000 to 2400 pre-merger notifications annually, of which 70 to 106 resulted in decisions to investigate further, and 2 to 7 led to cases being filed, depending on the year.[117] As already suggested, however, these latter statistics can be misleading because some mergers will be dropped along the way either because the parties are insufficiently confident of success or simply because they cannot tolerate the anticipated delays. Also, no doubt, some potential mergers are deterred; the more predictable are the agencies, due in part to the Merger Guidelines and years of experience under them, the more one would expect there to be few proposed mergers with a high likelihood of being challenged.

Another important outcome is settlement, most frequently through the parties spinning off plants, other operations, or lines of business in areas of significant competitive overlap.[118] That is, some mergers may be found to pose a serious competitive threat but only in certain geographic markets or only with respect to some of the many products the firms produce. In such cases, appropriate divestiture of pertinent assets will ordinarily satisfy the enforcement agencies.

For challenges that do proceed to court, the agencies often attempt to obtain a preliminary injunction requiring the merging firms to continue to operate independently pending a final outcome, and in the end successful challenges produce permanent injunctions against the merger (or subsequent negotiations leading to asset divestitures). This approach contrasts dramatically with the course of proceedings in earlier years, before the pre-merger notification regime was in place. Then, mergers were promptly consummated, and final decrees against challenged mergers ordinarily took many years, sometimes more than a decade, at which point the two firms were often sufficiently integrated (plants closed, brands discontinued, new joint operations well underway) to make practical divestiture difficult or impossible.

[117] DOJ (2006). The FTC's merger enforcement activity is comparable to that of the DOJ. In the European Union, from 1990 until May 2002 (and thus before the promulgation of the new 2004 regulations and guidelines), 86% of notified mergers were approved unconditionally, 5% were approved subject to undertakings (such as spin-offs) by the end of the one-month initial investigative period, an additional 3% after further investigation, and 1% (18) were prohibited. (Another 1% were withdrawn during in-depth investigations; various others were found to be outside the Commission's jurisdiction.) See Walle de Ghelcke and Gerven (2004, §5.01).

[118] During the past decade, the combined number of transactions that were restructured or abandoned after a formal challenge was announced but before a case was filed in court usually exceeded the number of cases filed. Furthermore, these statistics include only terminations that followed the issuance of a formal challenge.

The remainder of this section focuses on the substance of the legal restriction on horizontal mergers, first examining the core inquiry into anticompetitive effects and then considering the role of efficiencies in justifying mergers that would otherwise be proscribed. As will be seen, the approach toward both issues has evolved a great deal over time. Furthermore, throughout this evolution these two issues have not been entirely independent. In particular, the central question of the likelihood and extent of anticompetitive effects required to condemn horizontal mergers—the threshold for a successful challenge—seems to be answered in a manner that substantially reflects underlying views about the typical probability and magnitude of merger synergies.

4.4.2. Anticompetitive effects

In the 1960s (in the wake of the strengthening of Clayton Act Section 7 in 1950), the U.S. Supreme Court, following the lead of the federal enforcement authorities, adopted a restrictive view toward horizontal mergers. The Court condemned a number of mergers where the parties' combined market shares were under 10%, for example, in *Brown Shoe* and *Von's*.[119] The first government merger guidelines, promulgated in 1968, adopted similarly stringent thresholds for challenging mergers. Likewise, they endorsed the structural presumption that concentration implies anticompetitive effects, as articulated by the Supreme Court in *Philadelphia Bank*: "a merger which produces a firm controlling an undue percentage share of the relevant market, and results in a significant increase in the concentration of firms in that market is so inherently likely to lessen competition substantially that it must be enjoined in the absence of evidence clearly showing that the merger is not likely to have such anticompetitive effects."[120] A shift was signaled by the 1974 decision in *General Dynamics*.[121] The specific holding—that the *prima facie* case established by market share statistics could be rebutted by showing that the figures gave a misleading depiction of competitive effects—was not itself truly novel (*Brown Shoe* had suggested as much). However, the acts of subjecting the government's case to heightened scrutiny and ultimately rejecting it were taken as a signal of a new direction.

[119] Brown Shoe Co. v. United States, 370 U.S. 294 (1962); United States v. Von's Grocery Co., 384 U.S. 270 (1966). Indeed, in *Von's*, after pointing out that the combined share was 7.5% of the Los Angeles market for grocery stores, that the number of independent stores had fallen from 5365 to 3590, that half of the top 20 chains had acquired stores from smaller firms, and similar facts, the Court proclaimed: "These facts alone are enough to cause us to conclude . . . that the Von's–Shopping Bag merger did violate §7." 384 U.S. at 274. The dissent criticized the majority for attempting to "roll back the supermarket revolution" and asserted that "[t]he sole consistency that I can find is that under §7, the Government always wins." 384 U.S. at 288, 301.

[120] United States v. Philadelphia National Bank, 374 U.S. 321, 363 (1963). The Court cited prominent economic and legal authorities in support of this view, although the market share levels in *Philadelphia Bank* and proposed by most of the commentators were substantially higher (20% or more) than the levels deemed sufficient in many of the other cases of the period and in the 1968 guidelines.

[121] United States v. General Dynamics Corp., 415 U.S. 486 (1974).

Since the mid-1970s, there have not been further merger opinions by the Supreme Court.[122] Nevertheless, a confluence of three factors has made clear that the law has moved substantially: changing views toward competition and the effectiveness of market forces (both broadly and in the academy), a change in the composition of the Supreme Court and in the nature of its opinions on other antitrust subjects, and a new direction from the government as embodied in the 1982 merger guidelines. The current (1992) Merger Guidelines are a successor to the 1982 version, which differed in many key respects from those issued in 1968. Details of the current methodology for evaluating horizontal mergers will be examined throughout the remainder of this section.[123] Perhaps the most notable change, however, was in the thresholds for challenge: they were notably higher, sufficiently so that a number of the famous cases of the 1960s (that the government won) would not have been brought had the new guidelines been in effect.

The presumptive thresholds in the Merger Guidelines (once the market is defined; see subsection 4.5) are as follows. If the post-merger HHI is below 1000, the market is regarded as unconcentrated and ordinarily no further analysis will be undertaken. If the post-merger HHI is between 1000 and 1800, significant concerns are raised if and only if the merger increases the HHI by more than 100, in which case further analysis is undertaken.[124] And if the post-merger HHI exceeds 1800, significant concerns are deemed to exist when the merger raises the HHI by more than 50.[125] This air of certainty is misleading. In actual application of the Merger Guidelines, it has become apparent that, in certain industries, the *de facto* thresholds are much higher. For example, most of hundreds of hospital mergers subsequent to the 1982 guidelines (which had the same thresholds as those just described) have gone unchallenged even though post-merger HHIs and HHI increases were greatly above 1800 and 50, respectively, presumably reflecting a view of typical efficient scale in this industry.[126] Even so, the lawyers and economists who specialize in merger practice are generally aware of such

[122] This apparent anomaly is largely explained by two procedural changes: the new requirement of premerger notification reduced the flow of questionable mergers into the courts, and the elimination of special rules permitting automatic appeal in some cases directly to the Supreme Court greatly reduced the proportion of antitrust cases reaching that court.

[123] For further elaboration on how the guidelines are implemented in practice, see the enforcement agencies' commentary, FTC and DOJ (2006).

[124] Further analysis involves consideration of entry, efficiencies, and the possibility that one of the firms may be failing.

[125] The new 2004 EU guidelines (European Union, 2004a) are strikingly similar: 2000 replaces 1800, and the increases must be 250 and 150 rather than 100 and 50, respectively. As noted in the text and notes to follow, however, actual practice under the U.S. Merger Guidelines indicates the use of higher *de facto* thresholds (to an extent that varies by industry). Likewise, the additional factors considered in the new EU guidelines are largely the same as in the United States—notably, they include efficiencies as a defense—and they even appear in the same order.

[126] See also FTC and DOJ (2003), which gives statistics on the post-merger HHI and change in HHI for merger challenges in a number of industries. The tables strongly suggest that the thresholds for challenges vary greatly by industry. Coate and Ulrick's (2005) analysis of merger enforcement at the FTC finds that, for

patterns, which themselves probably reflect successful persuasion in prior merger filings in particular industries.[127]

Even though the Merger Guidelines are formally just a public statement by the federal agencies of how they intend to proceed internally, they have to a substantial degree dictated parties' practices in litigation and courts' analyses of mergers regarding the methodology for assessing market definition, anticompetitive effects, and other factors as well as the thresholds for condemnation. The courts undoubtedly welcome the guidance. Moreover, as noted, even though the most on-point Supreme Court opinions (from the 1960s) are much stricter, there has been a sufficient shift in understandings and in behavior of the Supreme Court on other issues that lower courts have followed the new approach rather than adhering to older precedents. The most direct indicator of this fact is that, in this more recent era, the government loses a good proportion of the cases it brings even though the government's threshold for challenging mergers is much higher than in the past (when they won nearly every case).

4.4.3. Efficiencies

Merger synergies play an important role in competition policy toward horizontal mergers, and one that has changed substantially over time. In this subsection, we first describe that role and its modern development and then consider how it relates to the goals of competition policy more generally.

Although efficiencies are usually discussed as a possible defense, advanced by the merging parties to justify a merger that might otherwise be condemned as anticompetitive, efficiencies have long had another, more significant influence on merger policy: in setting the threshold for antitrust scrutiny. Consider trivial combinations, say, when two individuals form a partnership or two small stores join forces. These combinations as well as some substantially larger mergers have never been subject to challenge. But they could have been. First, as noted in the discussion in subsection 3.5.2, such combinations literally involve price-fixing, going forward, which is automatically condemned, supposedly without proof of market power. Yet productive combinations, under the rule of reason, are permissible, presumably because they often create synergies (despite the fact that, once formed into a single entity, prices will be determined jointly rather than independently). Second, even when the number of firms is large, mergers in Cournot or

given HHIs and other factors, enforcement actions are more likely in the oil, grocery, and chemical industries. In addition, they find that, holding both the HHI and change in HHI constant, the probability of enforcement rises as the number of leading rivals falls from 5 to 4 and from 4 to 3.

[127] Leary (2002) emphasizes that, since the Merger Guidelines have now been followed for years, current merger policy exhibits a good deal of stability. Furthermore, in order "to provide greater transparency and foster deeper understanding regarding antitrust law enforcement," the DOJ and FTC issued a document, "Commentary on the Horizontal Merger Guidelines," in 2006. FTC and DOJ (2006). The agencies also held a merger enforcement workshop in 2004. FTC and DOJ (2004).

Bertrand oligopoly tend to raise prices (slightly). Nevertheless, nontrivial anticompetitive effects must be demonstrated before a merger will be challenged.[128]

Viewed more broadly, setting the threshold of anticompetitive effects significantly above zero may be rationalized by the view that mergers typically generate some synergies, so they should not be prohibited unless the reduction in competition is sufficiently great. In the 1960s and 1970s when U.S. anti-merger policy was strict, many fairly large mergers were nevertheless routinely permitted. Currently, as described in subsection 4.4.2, the thresholds are much higher; moreover, they are raised further in industries where synergies are thought to be unusually large relative to the size of the market (hospitals, for example).

Therefore, it seems appropriate to understand an efficiencies defense to a merger whose suspected anticompetitive effects exceed the threshold as implicitly involving a claim that the merger synergies are not merely substantial but are large enough to notably exceed the level ordinarily presumed to exist. After all, they must be enough to justify the merger in light of what would otherwise be substantial, not merely trivial, anticompetitive effects. This framing of the question may help explain why courts and enforcement agencies are cautious in accepting efficiency defenses (in addition to the obvious reason that merging parties have every incentive to assert the existence of synergies when there are few and the merger is in fact anticompetitive).

In the 1960s, the U.S. Supreme Court exhibited a somewhat schizophrenic but ultimately hostile attitude toward merger synergies. Most famously, in *Brown Shoe*, the Court viewed the efficiencies resulting from the vertical aspects of the merger as problematic because they would give the combined entity an advantage against competitors. On one hand, the Court stated: "It is competition, not competitors, which the Act protects."[129] However, the passage continues: "But we cannot fail to recognize Congress' desire to promote competition through the protection of viable, small, locally owned businesses. Congress appreciated that occasional higher costs and prices might result from the maintenance of fragmented industries and markets. It resolved these competing considerations in favor of decentralization. We must give effect to that decision."

Over time, the first statement—"competition, not competitors"—has continued to be among the most-quoted passages from all Supreme Court antitrust opinions, by courts and commentators alike, whereas the latter clarification (contradiction) is not usually appended anymore. Lower courts in the last couple of decades (recall, there have been no recent Supreme Court pronouncements on mergers) have varied in their approaches, some expressing uncertainty about an efficiencies defense but most accepting it, at least in principle.[130]

[128] Although the language of the Clayton Act demands a substantial effect, that of the Sherman Act does not. (Historically, the Clayton Act, passed in 1914 to remedy perceived weakness in the Sherman Act and amended in 1950 to strengthen it further, has been viewed as stricter than the Sherman Act. As noted in subsection 4.4.1, however, the enforcement agencies and, increasingly, the courts apply a more unitary approach.)

[129] Brown Shoe Co. v. United States, 370 U.S. 294, 344 (1962).

[130] See, for example, the cases cited in Areeda, Hovenkamp, and Solow (2006, vol. 4A, pp. 31–32 n. 17).

The Merger Guidelines are now unequivocal. In 1997, amendments to the 1992 version added a new section on efficiencies. They state: "Indeed, the primary benefit of mergers to the economy is their potential to generate such efficiencies. Efficiencies generated through merger can enhance the merged firm's ability and incentive to compete, which may result in lower prices, improved quality, enhanced service, or new products." Two main conditions are imposed: only merger-specific efficiencies are recognized, and those efficiencies must be sufficiently verifiable. Cognizable efficiencies also must be sufficiently large to prevent the merger from being anticompetitive.[131]

Which efficiencies are merger specific? Under the agencies' standard approach, economies of scale from, say, building larger plants, are likely to be accepted. Efficiencies from some other functions, like combining payroll operations, are less likely to be credited because of the option of contracting out for such services, achieving benefits of scale short of merger. And if two hospitals demonstrate a need to share modern imaging equipment due to high fixed costs, they may be permitted to form a joint venture limited to that purpose but not allowed on this ground to go forward with an otherwise anticompetitive merger of all of their operations.

Farrell and Shapiro (2001) explore which efficiencies should qualify as being merger specific, recognizing that in the presence of economies of scale firms can grow internally to reduce their average costs. They distinguish between efficiencies based solely on scale economies and efficiencies that reflect the combination of specific, non-tradable assets owned by the merging parties, which they consider true merger-specific efficiencies, or synergies. They argue that many claimed efficiencies are not in fact merger specific.

There is a deeper problem underlying many disputes about whether efficiencies are merger specific. Often, the parties will claim that the merger is necessary whereas enforcers will question why the purported benefit cannot be achieved through some more limited form of contractual arrangement. Under what circumstances there exist benefits from combining activities under the direction of a single firm that cannot be achieved through contracting is, of course, a core issue in the theory of the firm, one whose exploration was launched by Coase (1937), extended by Williamson (1975, 1985), and explored in subsequent work by Grossman and Hart (1986), Holmstrom and Tirole (1989), Hart and Moore (1990), and Hart (1995), among many others. Since the findings of this literature are often subtle, depending on factors that may not be readily and reliably ascertained in the course of an antitrust dispute, there will be an inevitable area of uncertainty. The level of theoretical sophistication reflected in modern analysis of contracts and the firm has yet to make significant inroads in merger analysis. In practice, the agencies tend to look at the forms of collaboration short of merger that are actually used in the industry to determine whether certain claimed efficiencies are merger specific.

In the end, enforcers do and should have a healthy skepticism about self-serving efficiency claims made by competitors seeking to merge. As we noted in subsection 4.3,

[131] As noted previously, the 2004 EU guidelines are quite similar, including a recognition of an efficiencies defense and the particulars of how it may be established in a given case.

acquiring firms often seem to overestimate the benefits of an acquisition, and the overall record regarding merger efficiencies is mixed at best. It also tends to be very difficult for the enforcement agencies or the courts to assess whether claimed efficiencies are indeed likely to arise. Kolasky and Dick (2003) examine the treatment of efficiencies in merger cases, arguing that the enforcement agencies and the courts have made substantial progress over the past twenty years incorporating efficiencies into merger analysis.[132]

Suppose that synergies can be quantified, or that in a particular case the enforcement agency or the court has done its best in determining the matter. The important remaining question is how efficiencies and anticompetitive effects are to be balanced. Should the decision depend on total economic welfare—the sum of consumer and producer surplus—the standard normative economic approach? Should it instead turn solely on consumer surplus—whether prices rise or fall—that is, whether the efficiencies are sufficiently passed on to consumers to at least offset any price increase that would otherwise ensue due to anticompetitive effects? Or should some other standard be used, which seems to be the approach in earlier cases like *Brown Shoe*?

This set of questions brings to mind the discussion in subsection 3.5.2 on the meaning of the rule of reason.[133] Recall that the *Chicago Board of Trade* case and subsequent decisions define reasonableness in terms of what promotes versus suppresses competition. Yet there remains ambiguity concerning the meaning of competition.

Under a process-oriented view, which has support in many modern cases, one might be concerned with protecting rivalry *per se*, which might imply a strict merger policy. That view seems consonant with the 1960s merger decisions and language such as that quoted above from *Brown Shoe* (the latter part, referring to the merits of protecting competitors and preserving decentralization) or that in *Philadelphia Bank* referring to the preservation of a "traditionally competitive economy."[134] Under this approach, not only would one condemn a merger from 2 firms to 1, 3 firms to 2, and 4 firms to 3, but also, it might seem, from N firms to $N - 1$, no matter how large was N. But as discussed previously, even when this view of merger policy was dominant, smaller mergers were permitted. Although efficiencies were not recognized as a defense and were sometimes viewed as an evil in larger mergers, the threshold for challenge was high enough to allow countless mergers to go unchallenged. Furthermore, modern cases such as *Broadcast Music* that interpret the rule of reason and the bulk of lower court cases on mergers, as well as the Merger Guidelines, accept more explicitly that efficiency counts posi-

[132] See also Pitofsky (1999) and Muris (1999) for different views on the role of efficiencies in merger analysis.

[133] The rule of reason, note, arose in interpretation of Sherman Act Section 1, which is also one of the statutes applicable to horizontal mergers. Thus, as a formal legal matter, one might say that the rule of reason is the Sherman Act standard for horizontal mergers. However, courts have generally considered mergers separately, and as noted, increasingly without regard to which statute is invoked in a particular case. Nevertheless, given the growing convergence in approaches under all of the antitrust statutes, one should expect some congruence between courts' analyses of horizontal mergers and of what are denominated as rule of reason cases.

[134] United States v. Philadelphia National Bank, 374 U.S. 321, 371 (1963).

tively, and in particular that better serving customers, notably, through lower prices, is desirable.[135]

Thus, at present the main contest seems to be between consumer welfare and total welfare, that is, whether efficiencies should be credited when they increase producer surplus rather than being passed on to consumers. The influence of merger synergies on price depends on both the manner in which the firms' cost function is altered by the merger and on the nature of firms' interaction. At one extreme, if a merger produces no (or negligible) cost reductions but reduces the number of competitors in a Cournot or Bertrand oligopoly, and N is not very large, prices will rise nontrivially and the reduction in consumer welfare will be approximately equal to the fall in total welfare. No tradeoff needs to be considered. To take another possibility, suppose that two firms merge to monopoly and that all the savings are in fixed costs; then prices will rise (unless there was perfect collusion previously) because fixed costs do not ordinarily affect pricing decisions. However, if the increase in deadweight loss is not that great but the savings in fixed costs is large, then consumer welfare may fall while total welfare rises. For a further contrast, consider a merger of two firms in a setting with many firms and assume that the cost saving involves a reduction in marginal cost; then the merged firm may price sufficiently more aggressively to bring prices down, in which case both consumer surplus and total welfare would rise. Viewed more broadly, the analysis in subsections 4.1 and 4.2 above (the latter of which drew on section 3 on collusion) elaborates conditions under which prices may be expected to rise or fall and how those conditions depend on the manner in which a merger may affect firms' costs.

Accordingly, although many mergers raise both consumer and total surplus or reduce them both, there will exist a notable subset of mergers that increase total surplus but reduce consumer surplus. The resulting need to choose between consumer welfare and total welfare as guides to merger policy was presented sharply by Williamson's (1968) classic discussion of the tradeoff between market power and efficiencies and, in particular, his demonstration in a basic case that even modest gains in productive efficiency could exceed the losses in allocative efficiency from price increases.[136] Recent contributions to the debate about the proper objective of antitrust policy include Farrell and Katz (2006), Heyer (2006), Kolasky and Dick (2003), and Salop (2005).

The precise language of the various statutes—which as prior discussions indicate are not taken literally or interpreted independently—gives conflicting guidance in de-

[135] Other jurisdictions, with different laws and histories, might give weight to other objectives in addition to or instead of economic welfare (whether consumer welfare or total welfare). For example, antitrust policy in the European Union has traditionally placed some weight on the value of integration into a single market.

[136] In a market where prices are initially at a competitive level, the benefits from productive efficiency are a rectangle (assuming a uniform downward shift in the marginal cost curve) whereas deadweight loss is the familiar triangle, which for small price increases will be a much smaller quantity. However, if price is nontrivially above marginal cost before a merger, as might be expected if pre-merger concentration is high, the incremental deadweight loss from price increases will be larger. Also, cost savings from mergers, when they exist, can take many different forms, so the benefit need not be indicated by a rectangle (that is, quantity times a uniform change in marginal cost).

termining the underlying principle.[137] Clayton Act Section 7's prohibition on mergers that may substantially lessen competition or create monopoly might be interpreted to prohibit any mergers that reduce rivalry or at least those leading to high concentration, and certainly to a single firm serving the market. Yet it is possible that even a merger to monopoly could raise both total welfare and consumer welfare. The Sherman Act's prohibition on restraints of trade is interpreted under the rubric of the rule of reason, which, as has been explained, contains substantial ambiguity even after being translated into the promote/suppress competition test and refined through modern applications. The FTC Act's prohibition of "unfair" competition is vague and question-begging on its face.

That said, the modern trend in the United States seems to be toward a consumer welfare standard when considering the efficiencies defense, although the legal authorities have not elaborately rationalized this view, specifically by defending it against the alternative of total welfare. The Merger Guidelines' discussion of efficiencies at one point refers to those that "likely would be sufficient to reverse the merger's potential to harm consumers in the relevant market, for example, by preventing price increases in that market." Likewise, a number of lower courts considering the issue have indicated the need for the merging firms to demonstrate that efficiencies would be passed along to consumers.[138]

From an economic point of view, however, it would seem that in principle total welfare should be the standard.[139] Producers have owners who are people, just like final consumers. One might nevertheless favor consumers on distributive grounds because owners are on average richer than consumers (although obviously the groups overlap substantially). However, it is usually more efficient to achieve distributive objectives directly, through the income tax and transfer system.[140] In addition, one should add that,

[137] The legislative histories, however, are clearer. When the Sherman Act was enacted in 1890, it is unimaginable that the legislators appreciated modern concepts of efficiency and deadweight loss, which were unknown even to most economists at that time. Legislators did seem to care about high prices and also about the protection of small businesses. As discussed by the Supreme Court in *Brown Shoe*, the Clayton Act amendments in 1950 that led to the version of Section 7 that is close to its current form were motivated by general concerns (largely of a political and social sort) about concentration of power, most of which seem disconnected from either consumer or total economic welfare. As indicated by the text in this section and in section 3, however, those views have not significantly influenced courts' decisions in recent decades.

[138] See Areeda, Hovenkamp, and Solow (2006, vol. 4A, p. 40 n. 1).

[139] A total welfare standard is also consistent with maximizing worldwide welfare. When there are multiple jurisdictions with authority over mergers with international spillovers, different jurisdictions might find it in their national self-interest to adopt different standards. For example, a country that is weighted toward consumers in sectors most influenced by antitrust policy might want to be stricter and use a consumer welfare standard, whereas a country with stronger producer/owner representation among its citizens might prefer total welfare or even a focus on productive efficiency. See Guzman (1998). Some have suggested that a mix of national bias and protectionism—rather than (or in addition to) differences in underlying approaches—may explain the (few) instances (the most notable being the proposed but ultimately abandoned merger of G.E. and Honeywell and the ultimately approved merger of Boeing and McDonnell-Douglas, subject to some conditions) in which the U.S. and EU antitrust authorities have taken different views of mergers or other practices.

[140] See, for example, Kaplow and Shavell (1994) and Kaplow (2004).

in antitrust, the protected class is often producers. This is patently so when the merging parties sell intermediate goods (although savings to purchasing firms that operate in a competitive market will ultimately be passed on to subsequent consumers). Furthermore, antitrust law treats buying cartels and horizontal mergers that create monopsony power similarly to the way it treats price-fixing and the creation of market power by sellers.[141] Yet lower input prices can, to a degree depending on market structure and other factors, translate into lower prices to final consumers. If only consumer welfare mattered, increases in buyer power through horizontal mergers and otherwise might be praised, not condemned.

Pragmatic considerations may, however, provide justification for cabining efficiency defenses in various respects, including perhaps a focus limited to consumer welfare. For example, adopting a consumer welfare standard may induce firms to undertake deals that obtain potential synergies while causing less harm to competition, leading to even higher total welfare than would a total welfare standard. In any event, the previously-noted difficulties of ascertaining merger-specific efficiencies in individual mergers, *ex ante*, counsel caution, which is already reflected in the Merger Guidelines and also seems evident in court decisions. Indeed, this may explain why average efficiencies heavily influence the thresholds for anticompetitive effects, both generally and in specific industries, while at the same time efficiencies are often given short shrift in examining particular cases, absent an unusually strong demonstration. But once sufficiently persuasive proof of atypically great efficiencies is offered, it is not clear that a requirement of pass-through to consumers makes the agencies' or courts' task easier rather than harder. On one hand, determining pass-through requires resolution of an additional, challenging issue. On the other hand, in some instances the best evidence of a merger's effects will be from observed pricing behavior that may reflect a combination of anticompetitive effects and efficiencies. For example, in *Staples*, evidence of higher prices in more concentrated markets presumably reflected the combined impact of concentration and whatever pass-through of efficiencies may have been occurring. Note that reductions in marginal cost are generally passed through to some degree to consumers, even by a monopolist, although the pass-through rate can be sensitive to the shape of the demand curve and the nature of oligopolistic interaction.[142]

There are also longer-run concerns relating to dynamic efficiency. Some degree of competition has often been thought conducive to firms' running a tight ship, better serving customers, and being more innovative. The relationship between competition and

[141] For example, the Merger Guidelines state: "The exercise of market power by buyers ('monopsony power') has adverse effects comparable to those associated with the exercise of market power by sellers. In order to assess potential monopsony concerns, the Agency will apply an analytical framework analogous to the framework of these Guidelines."

[142] Bulow and Pfleiderer (1983) derive an expression for the pass-through rate for a monopolist. Shapiro (1989) discusses pass-through rates in various oligopoly models.

innovation has proven difficult to establish as a general theoretical matter.[143] In any event, if the process of competition itself—which seems to have been favored by the drafters of antitrust legislation and earlier courts and commentators—is of some value, but this value is difficult to measure, then it makes sense to tilt the balance against concentration. This might be done by making the threshold for challenge lower—requiring less demonstration of anticompetitive effects—or through other means, such as being less generous in considering efficiencies in justifying otherwise problematic mergers. Greater stinginess might be accomplished by raising proof burdens or by imposing additional requirements, such as by requiring savings to be passed along to consumers. The optimal manner of incorporating these sometimes subtle and typically unpredictable dynamic concerns into horizontal merger policy is not obvious.

4.5. Market analysis under the Horizontal Merger Guidelines

The Horizontal Merger Guidelines are utilized to evaluate specific proposed mergers, requiring one to move from abstractions and general theories to decisions about a concrete case. To motivate this problem, consider the proposed merger of US Airways and Delta Air Lines. In November 2006, US Airways made an unsolicited offer to acquire Delta, which at the time was in bankruptcy. Predicting the effects of this merger on airline passengers is highly complex. First, one must identify the routes on which US Airways and Delta are significant direct rivals. Then one seeks to determine whether there will be a unilateral or coordinated price increase on those routes. Diversion ratios and gross margins are certainly relevant to this inquiry, but many other factors enter into the picture as well. Will competition from other carriers, including low-cost carriers, prevent the merged entity from profitably raising fares? Given the complex fare structures that are used in the airline industry, will the merger have different effects on fares for different classes of customers, such as leisure versus business travelers? How do frequent-flier programs affect the analysis? Will the merger generate substantial efficiencies based on running a larger network of flights at a single airline? In evaluating these efficiencies, how does one factor in the role of airline alliances, a less complete form of collaboration than a merger? If efficiencies lead to lower fares on some routes, but if the reduced competition leads to higher fares on other routes, how does one balance these diverse effects on different sets of consumers? Since one is looking ahead a year or more, is the industry changing in significant ways, such as through the growing role of regional jets and low-cost carriers, that make it appropriate to discount historical experience when making predictions? Lastly, how strong a competitor will Delta be if it is not acquired by US Airways, especially given the frequency of bankruptcies in this industry?

[143] Competition generally enhances innovative incentives, as firms seek to gain ground on their rivals by introducing new and improved products. But a competitive market structure can cause problems for innovation if firms have difficulty appropriating the returns from their innovative efforts, for example, due to rapid imitation by rivals.

We will consider only some of the core, recurring issues.[144] The first step in the analytical framework provided by the Merger Guidelines is defining the relevant market, typically a relevant product market along with a relevant geographic market. Then, as previously described in subsection 4.4.2, the government determines whether there is a likely competitive risk based on the post-merger HHI and the increase in HHI due to the merger. If there is, it considers entry and other forms of supply response (as well as efficiencies, already discussed in subsection 4.4.3, and the possibility that one of the firms is failing). In this subsection, we describe this approach and relate it to the preceding economic analysis and further work.

4.5.1. Market definition: general approach and product market definition

In evaluating methods used to define the relevant market, one should ask whether the resulting measures of concentration are reasonably probative of anticompetitive effects in instances in which concentration is high and notably increased by the merger. This question is particularly pressing because, as we explained in subsection 2.4.2.2, the common approach of defining markets and looking to market shares is just one of many, not necessarily the most reliable, and rather indirect and incomplete—indeed, sufficiently so that it does not play a central role in economists' analysis of market power.

It is useful to begin by revisiting the precise relevance of market definition and market share. A single-product firm's pricing decision is governed by equation (1),

$$\frac{P - MC}{P} = \frac{1}{|\varepsilon_F|},$$

which is to say that the firm's margin is inversely related to the magnitude of its own, firm-specific elasticity of demand. In the model of a single, dominant firm pricing with a competitive fringe, this firm-specific elasticity is given by equation (2),

$$|\varepsilon_F| = \frac{|\varepsilon_D| + (1 - S)\varepsilon_R}{S}.$$

Thus, we recall that the firm's market share, S, is relevant for two reasons. First, a higher share means that there are fewer competitors, so for a given elasticity of supply response, the total response will be smaller. (Hence the $1 - S$ in the numerator.) Supply response will be considered further in subsection 4.5.3. Second, a higher share indicates that the firm captures a greater proportion of the industry profits due to a price increase.

Expression (2) also indicates, as we emphasized previously, that the firm-specific elasticity of demand depends on the market elasticity of demand, which raises complex problems for an approach that first defines a market and then looks to market share. If the market elasticity of demand were the same in every properly defined market,

[144] Some issues that we do not cover, such as whether Delta is appropriately considered a "failing firm," are treated extensively in the Merger Guidelines. Most others are noted in passing or would be admissible in light of various catch-all phrasings inviting consideration of any pertinent factors.

then (setting aside issues concerning rivals' supply response) shares would have clear implications for market power. However, this is certainly not the case (and, if it were, no one has ever suggested what that common market elasticity is). Generically, an approach that defines the relevant market (whatever that might mean) and then looks only at market share can be highly misleading. As we explained, a high share does not imply substantial market power if market demand is highly elastic, and a low share does not imply a lack of significant market power if market demand is sufficiently inelastic.

At best, markets can be defined so as to minimize these concerns, although as we explained earlier, one often needs to know the right answer—that is, how much market power exists—in order to know which market definition is best. A major factor that contributes to this difficulty is the all-or-nothing nature of market definition: products (or regions) are either "in" or "out." Moreover, we discussed how even good substitutes, in the sense of having a high cross-elasticity of demand with the firm's product, may impose little restraint, notably, if they are a small share of consumers' expenditures, whereas a group of mediocre substitutes might, taken together, impose substantial restraint.

Nevertheless, the Merger Guidelines in fact utilize an approach that relies heavily on defining markets, and the outcome of litigated merger cases often turns on how the relevant market is defined. Accordingly, it is important to consider how this task is presently accomplished and how it might be improved. For the case of product markets (the analysis of geographic markets is analogous), the Merger Guidelines (§1.11) specify the following procedure:

> Absent price discrimination, the Agency will delineate the product market to be a product or group of products such that a hypothetical profit-maximizing firm that was the only present and future seller of those products ("monopolist") likely would impose at least a "small but significant and nontransitory" increase in price [SSNIP]. . . . Specifically, the Agency will begin with each product (narrowly defined) produced or sold by each merging firm and ask what would happen if a hypothetical monopolist of that product imposed at least [an SSNIP], but the terms of sale of all other products remained constant. If, in response to the price increase, the reduction in sales of the product would be large enough that a hypothetical monopolist would not find it profitable to impose such an increase in price, then the Agency will add to the product group the product that is the next-best substitute for the merging firm's product. . . . The price increase question is then asked for a hypothetical monopolist controlling the expanded product group. . . . This process will continue until a group of products is identified such that a hypothetical monopolist over that group of products would profitably impose at least [an SSNIP], including the price of a product of one of the merging firms. . . . In attempting to determine objectively the effect of [an SSNIP], the Agency, in most contexts, will use a price increase of five percent lasting for the foreseeable future.[145]

[145] When price discrimination is feasible and profitable, the Merger Guidelines (§1.12) allow for separate "price discrimination markets" "consisting of a particular use or uses by groups of buyers of the product"

This approach focuses on the ability of a hypothetical monopolist to raise prices, but the question at hand is a merger, and typically not a merger to monopoly. The behavior of the hypothetical monopolist seems most relevant when considering the possibility of coordinated effects. In the theory of collusion that we discussed in section 3, we implicitly assumed that we knew the group of firms in the market that might collude. If, regarding the merger under examination, that group of firms corresponds to the firms in the market just defined, then the Merger Guidelines approach will (supposing for the moment that it works) have indicated that successful collusion by the firms in the industry would indeed have a significant anticompetitive effect. If not—if, say, the market properly includes firms producing products very different from those of the merging firms, where heterogeneity or other factors are such that collusion is unlikely to be feasible—then the analysis would suggest that coordinated effects are probably not a concern. The second question with coordinated effects is, assuming that collusion would be profitable, how much does the present merger increase the likelihood of successful collusion? We examined this question in subsection 4.2, above, where we discussed various reasons that higher concentration and greater increases in concentration bolster collusion but also noted some reservations (namely, that if collusion is sufficiently likely short of monopoly, it is possible that further increases in concentration would not have much incremental anticompetitive effect).

The role of market definition is less clear, however, under a theory of unilateral effects, where the profitability of a price increase for the merged entity depends on the demand system and on such factors as the gross margins and the diversion ratios that measure the proximity of the merging firms' products, not on drawing lines between products that are in or out of the relevant market. However, in the particular case in which there is a logit demand structure, the elasticity of demand for the inside products as a group is relevant to calculating unilateral effects. If the merged entity would have only a small share of the relevant market, calibrated and defined using the SSNIP test, unilateral competitive effects in excess of the amount used to define the SSNIP are relatively unlikely, but such effects may well arise if the firms have a large combined share in the relevant market. Defining markets in this way thus allows the government to meet its presumption under a unilateral effects theory if the combined share of the merging firm exceeds some threshold.[146] Also, in industries where products are fairly homogeneous and capacities are important, such as in some chemical and energy markets, using

for which an SSNIP would be profitable. For example, in a railroad merger, the hypothetical monopoly rail carrier on a given route may be able to price discriminate between shippers who have the ability to ship by water versus those whose next best alternative to rail service is trucking. See Varian (1989) on price discrimination generally and Stole (2007) on its relationship to imperfect competition. In some circumstances, the hypothetical monopolist may be able to engage in price discrimination even if such discrimination is not observed prior to the merger because it is undermined by competition among the suppliers in the proposed relevant market. However, Hausman, Leonard, and Vellturo (1996) point out that price discrimination can be unprofitable if the methods used by the hypothetical monopolist to price discriminate are imperfect.

[146] The Merger Guidelines provide that, "Where market concentration data fall outside the safe harbor regions of Section 1.5, the merging firms have a combined market share of at least thirty-five percent, and

the Merger Guidelines to define a relevant product and geographic market, and making inferences about unilateral effects based on concentration measures, fits reasonably well with the theory of Cournot equilibrium.

Now consider how, as a practical matter, one might implement the Merger Guidelines' approach to defining markets. For simplicity, consider a symmetric situation in which the pre-merger price of each product is P, the pre-merger output of each product is X, and the marginal cost of producing each product is a constant, MC, so the pre-merger gross margin is $m \equiv (P - MC)/P$. In this case, the SSNIP test reduces to a very simple formula. Suppose that the hypothetical monopolist raises the price of one of the products by a factor G, to $P(1 + G)$. (The same calculation could be done for a price increase applied to all of the products.) If the price increase would cause no decline in unit sales, it obviously would be profitable, whereas if it would cause unit sales to drop to zero, it would not be profitable. Since profits are continuous in sales, there exists an intermediate level of sales for which the price increase would break even. The largest percentage reduction in sales such that this price increase is just barely profitable is referred to as the "critical loss," L. By definition, $(P - C)X = (P(1+G) - MC)X(1 - L)$. Solving for L gives $L = G/(G + m)$. This expression comports with intuition: a larger loss of sales is tolerable if the price increase is greater, but the acceptable loss of sales is smaller, the larger is the initial profit margin. For illustrative purposes, suppose that the magnitude of the SSNIP is 10%, so we are considering $G = 0.1$, and that the pre-merger margin is 30%; then $m = 0.3$ and we get $L = 0.25$.

These calculations are directly relevant to the SSNIP test in the Merger Guidelines. Consider a cluster of products that is being tested to see if the group of products is sufficiently inclusive to form a relevant market. Suppose that the pre-merger gross margin on each product is the same and, furthermore, that one is asking about a uniform price increase for all of these products and all customers. That hypothesized SSNIP will be profitable, and hence the products will form a relevant market under the SSNIP test, if and only if the actual loss of sales is less than the critical loss. Thus, in the foregoing example above, the hypothetical monopolist would find it profitable to impose a 10% price increase so long as the resulting actual loss would be less than 25% of the initial level of sales. Using simple arithmetic, we have thus translated the SSNIP test into a very well-defined economic question regarding the (arc) elasticity of demand facing the hypothetical monopolist. (Keep in mind, however, that this analysis is applicable to a uniform price increase in all products in the market, not to whether two merging firms would find it profitable to unilaterally increase the prices on one or both of their products.)

Katz and Shapiro (2003) and O'Brien and Wickelgren (2003) show how one can go beyond the simple arithmetic underlying critical loss to use pre-merger equilibrium

where data on product attributes and relative product appeal show that a significant share of purchasers of one merging firm's product regard the other as their second choice, then market share data may be relied upon to demonstrate that there is a significant share of sales in the market accounted for by consumers who would be adversely affected by the merger."

relationships to sharpen the SSNIP test. Their work is motivated by the observation that the critical loss falls if the margin is higher. In the example just given, if the gross margin rises from 0.3 to 0.4, the critical loss falls from 25% to 20%. Based on this observation, merging parties whose products are sold at large gross margins have been known to argue for the inclusion of many products in the relevant market, based on the fact that the critical loss is small. However, these articles show that such logic is incomplete and can be highly misleading. The reason is that the existence of high pre-merger margins itself indicates that the pre-merger firm-specific elasticities of demand must be relatively small, and this of course has direct implications for the elasticity of demand facing the hypothetical monopolist in the post-merger market.

To see how the logic works, consider imposing the percentage price increase G on one product in the symmetric case just discussed. Since the firm-specific elasticity of demand for that product is the inverse of the margin, the percentage of sales lost for this product will be approximately G/m when G is small. Katz and Shapiro define the aggregate diversion ratio D for a given product as the fraction of the overall sales lost by that product that are captured by (diverted to) any of the other products in the candidate product market.[147] Therefore, the hypothetical monopolist who raises the price of this product effectively only loses a fraction $(1 - D)$ of the sales that are lost by that particular product. As a consequence, the actual loss of sales for the hypothetical monopolist is only $A = (1 - D)(G/m)$. The price increase is profitable if and only if the actual loss is less than the critical loss of $L = G/(G + m)$. With a few steps of algebra, it can be shown that $A < L$ if and only if $D > L$.[148]

Where applicable, this formula tells us that a group of products will form a relevant market when the aggregate diversion ratio is larger than the critical loss. To illustrate, suppose that the pre-merger gross margins are 40% and that one is considering a 10% SSNIP, so the critical loss is 20%. Now imagine that the price of one product in the candidate group of products is raised. If more than 20% of the lost sales are diverted to other products in this group, rather than outside the group, the group forms a relevant market. Katz and Shapiro (2003) stress that this test can lead to relatively narrow markets, especially in cases where the pre-merger margins are large. O'Brien and Wickelgren (2003) argue that critical loss analysis has often been done incorrectly, in a way that is inconsistent with the pre-merger equilibrium conditions, which they illustrate with two

[147] In the symmetric case with N products, the aggregate diversion ratio is equal to $N - 1$ times the diversion ratio between the product in question and any one of the other products. More generally, the aggregate diversion ratio is the sum of the diversion ratios for all of the other products being considered.

[148] Katz and Shapiro (2003) show how the calculations change if the pre-merger margins differ for the different products in the candidate market. In such cases, the Merger Guidelines ask whether the hypothetical monopolist will raise the price of *any* of the products sold by the merging firms. It may be profitable to raise the price of a product with an especially low margin because diversion of sales to other products will actually raise the profits of the hypothetical monopolist, offsetting sales that are not diverted to other products in the candidate group.

litigated merger cases, *FTC v. Tenet Healthcare Corp.*, a hospital merger, and *FTC v. Swedish Match*, a merger involving loose leaf chewing tobacco.[149]

4.5.2. Geographic market definition

The Merger Guidelines approach to geographic market definition formally parallels its approach to product market definition. To illustrate how the analysis would proceed, consider BP's acquisition of Arco. These two companies were the leading producers of Alaskan North Slope crude oil, the vast majority of which was sold to refineries on the U.S. West Coast. These refineries required crude oil to make refined products, notably gasoline, so the relevant product was clearly crude oil. To determine the geographic market for the supply of crude oil to U.S. West Coast refineries under the Merger Guidelines, one starts with the merging firms' production locations, Alaska. One asks whether a hypothetical monopolist over Alaskan crude oil could profitably impose an SSNIP, taking as given the price of crude oil supplied from other locations to the West Coast refineries, which in this case included California and a variety of countries from which crude oil was shipped to the U.S. West Coast by tanker. If substitution to California crude oil would defeat a price increase imposed only on Alaskan North Slope crude oil, then California must be added to Alaska and the exercise repeated for foreign sources. In this case, since many West Coast refineries had demonstrated the ability to shift to imported crude oil in response to small price changes, the relevant market was arguably worldwide.[150]

For the purposes of geographic market definition, it is common to look at the patterns of imports and exports across a given geographic boundary. However, care must be taken, for otherwise this method can be misleading.[151] A prominent way of examining imports and exports is the method advanced by Elzinga and Hogarty (1973), which

[149] FTC v. Tenet Healthcare Corp., 17 F. Supp. 2d 937 (E.D. Mo. 1998), rev'd, 186 F. 3d 1045 (8th Cir. 1999); FTC v. Swedish Match, 131 F. Supp. 2d 151 (D.D.C. 2000).

[150] See Bulow and Shapiro (2004) and Hayes, Shapiro, and Town (2007) for further analysis of geographic market definition in this case.

[151] One might wonder how it can be widely recognized that nontrivial product substitution is consistent with the existence of significant market power whereas geographic market substitution (imports and exports) is so often viewed as strong presumptive, if not conclusive, proof of the lack of market power. The common reasoning seems to imagine a case of homogeneous products where the only distinction involves transportation costs, which are identical for all consumers at a given location. Think about supplies of fungible intermediate goods to firms. If a fraction of local firms is already importing its supply, all local producers must be indifferent between local purchases and imports at current prices, so an increase in local prices would cause substitution limited only by the extent to which foreign supply net of foreign demand slopes upward. But this reasoning is inapplicable when products are not homogeneous or consumers' locations or transportation costs vary (the latter of which is often true when the consumers rather than the goods must travel). Then, the relevant analogy is to markets with differentiated products, as will become apparent in the discussion to follow in the text. See, for example, Kaplow (1982).

has been widely used in recent years in hospital merger cases.[152] Basically, one starts with a narrow geographic market (typically a modest radius around the locale of two merging hospitals) and then progressively expands the candidate geographic market to include more distant hospitals until two conditions hold: (1) some large fraction of the merging hospitals' business comes from the candidate geographic market, and (2) some large fraction of the individuals resident in the candidate geographic market use hospital services within that region.

While such measures of imports and exports can be informative for the hypothetical-monopolist SSNIP test, and for estimating the competitive effects of mergers, they cannot substitute for evidence regarding the ability and willingness of consumers to substitute products supplied outside versus inside a candidate geographic market in response to price changes.[153] To see the problem, consider the case in which the two merging hospitals are nearby, with no other hospitals in the immediate area. Unless the fraction θ of the customers served by these hospitals that comes from the local area is very high, the proposed approach would draw the market more broadly.[154] How does this reasoning relate to whether an SSNIP could profitably be imposed by the merged entity, which is in fact a monopolistic supplier in the local area? Suppose that the elasticity of demand from local customers is ε_L and the elasticity of demand from customers who must travel further is ε_T. If the local customers can be identified and discriminated against by the merged entity, the fraction of customers coming from outside the local area, $1-\theta$, is of little or no relevance to an SSNIP targeted at local customers, so there is no reason to believe that the test gives a meaningful answer. How does this test perform when price discrimination is not possible, and when more distant customers exhibit more elastic demand? The elasticity of demand facing the hypothetical monopolist is given by $\theta\varepsilon_L + (1-\theta)\varepsilon_T$. With $\varepsilon_T > \varepsilon_L$, this is decreasing in θ, that is, the elasticity is greater, the more patients come from outside the local area. But this hardly answers the

[152] See Frech, Langenfeld, and McCluer (2004) for an extensive discussion of how these tests have been used in hospital merger cases. Capps et al. (2002) emphasize problems with using patient-flow data to define markets in hospital mergers and suggest an alternative approach based on estimating a logit model of hospital demand. They find that mergers that would be permitted by relying on patient-flow data may easily fail according to the SSNIP criterion under all three approaches that they examine; indeed, such data were largely uncorrelated with SSNIP except in extreme cases.

[153] Whinston (2006, pp. 92–93) comments on the "serious flaws" that can arise by defining relevant markets based on transshipment patterns.

[154] In practice, this fraction will depend on the type of service. For example, markets for emergency care or maternity care may well be more local than markets for elective surgery. The geographic market definition exercise must, in principle, be performed for each separate relevant product, at least if these products are priced separately, although sometimes the focus is on the average for all services, which is systematically misleading. (In our example, it may be that a high fraction who have unusual or extreme conditions travel a good distance, to major medical centers, whereas travel would not be worthwhile for most other medical needs.) Another practical difficulty is that consumers' locations are usually measured from the zip code that appears in admission records, but some individuals live part of the year elsewhere or they may work far from their residence, so for them a more distant facility may be less inconvenient (or may be the only one available under their employer's health plan).

SSNIP question, which requires information about the pre-merger margins, the critical loss, and the actual loss, which depends on $\theta\varepsilon_L + (1-\theta)\varepsilon_T$. In fact, the methods of Katz and Shapiro (2003) and O'Brien and Wickelgren (2003) discussed above tell us that the local area is a local market if and only if the diversion ratio from one hospital to the other is greater than the critical loss. These measures, in turn, bear no necessary relationship to measures of the fraction of demand coming from local residents or the fraction of services local residents obtain outside the proposed narrow market. That is, a local geographic area may well constitute a relevant geographic market, under the SSNIP test, even if a significant number of patients served by the local hospitals come from outside the area and even if a significant number of patients who live locally use more distant hospitals.

Looking at import and export data can also readily lead to the opposite error: incorrectly concluding that a local region is a relevant market. Recall our example in subsection 2.4.2.2 of a trucking company that served 100% of a particular route, yet a slight price increase would lead nearby trucking companies to offer their services. To consider a more substantial example, return to the BP/Arco merger and ask whether Alaska and California constitute a relevant geographic market for crude oil. As long as transportation costs are not essentially zero and product differentiation is minimal, one would expect oil supplies to flow to the nearest refineries. Suppose that the supply of crude oil from Alaska and California is sufficient to supply 95% of the needs of refineries on the U.S. West Coast and that more than 95% of crude oil from Alaska and California is used at West Coast refineries. The Elzinga-Hogarty tests would indicate that Alaska and California qualify as a relevant geographic market. But this result would be incorrect if there is an elastic supply of crude oil from other locations at the current price, or at a price just above the current price but less than the price after the SSNIP, which would be true if the cost of importing oil from more distant locations were not significantly greater than that of transporting oil from Alaska. In fact, there is strong evidence that the supply of imported crude oil to the U.S. West Coast is highly elastic, with prices set by transportation arbitrage conditions.

4.5.3. Rivals' supply response

As we noted (reminded) in subsection 4.5.1, the ability to raise price depends not only on market share—and on the market elasticity of demand—but also on rivals' supply response, the subject of subsection 2.4.2.3. The Merger Guidelines explicitly take this dimension into account. They distinguish between "uncommitted entry" and "committed entry."

Uncommitted entrants are firms that would likely enter the market in response to an SSNIP in less than one year and without the expenditure of significant sunk costs. They are counted as market participants. That is, in computing market shares, their capacity is taken into account.[155] For example, a firm that currently makes few sales in the relevant

[155] There are additional subtleties involved in computing firms' market shares. In some cases, especially where products are differentiated, market shares are based on sales. In other cases, capacities are used. The

market would be assigned a large share if that firm would greatly expand its sales in response to an SSNIP. By defining market share in this manner, the supply responses of all market participants to an SSNIP are incorporated into the analysis. This approach helps to adjust for the fact that the Merger Guidelines define the relevant product and geographic markets solely on the basis of demand-side considerations. If the government's *prima facie* case is established using these shares in markets thus defined, there is a presumption that current competitors' ability to expand is not sufficient to alleviate concerns associated with the merger.

However, it is also possible that the prospect of additional, committed entry, which might take as long as two years and require the expenditure of costs that then would be sunk, will either deter any post-merger price increase or ensure that it is short-lived. In order for the anticompetitive effects of a highly concentrating merger to be negated by the prospect of entry, the Merger Guidelines require that such entry be timely, likely, and sufficient to counteract the competitive effects of concern. Werden and Froeb (1998) explain how entry may well not be profitable if the pre-merger equilibrium reflects the sunk costs associated with entry. They also show that, if entry does occur, it can make mergers unprofitable unless they generate synergies. Baker (2003b) discusses the role of entry in recent merger cases.

4.6. Predicting the effects of mergers

In practice, myriad factors can come into play when predicting the competitive effects of mergers. Every industry has its own unique attributes, be they scale economies, the role of advertising and reputation, networks effects, consumer switching costs, product compatibility, technological change, regulations, intellectual property rights, or trade barriers. Virtually any topic in industrial organization economics can come into play in merger analysis. We confine our attention here to the two leading techniques used to predict the price effects of mergers, with the caveat that to use these methods reliably in any given case one must have a thorough awareness of industry-specific factors.

4.6.1. Direct evidence from natural experiments

In some cases, a reduced-form approach is possible, under which empirical evidence is presented showing directly, through "natural experiments," that the merger will lead to higher prices. The *Staples* case is an excellent example of the successful use of this approach by the government.[156] In this case, the FTC successfully challenged the

general principle is that "Market shares will be calculated using the best indicator of firms' future competitive significance." For example, to the extent that a firm's capacity is committed—perhaps under a long-term contract, or perhaps to other highly profitable uses, including those within a vertically integrated firm—and thus unavailable in response to an SSNIP, that capacity will not be included in measuring the firm's market share.

[156] Federal Trade Commission v. Staples, Inc., 970 F. Supp. 1066 (D.D.C. 1997).

proposed merger between Staples and Office Depot, two of the three leading office superstore chains (the other being OfficeMax). As explained by Baker (1999), the FTC presented econometric evidence showing that the prices charged by Staples at stores facing competition from nearby Office Depot stores were lower than the prices charged by Staples at stores without nearby Office Depot stores. The FTC offered further expert testimony showing that these price differences were not caused by other factors that happened to be correlated with the presence or absence of nearby Office Depot stores.

Economists might say that the FTC's reduced-form approach, by going directly to the likely effects of the merger, reduced or even eliminated the need to define the relevant product market. In practice, however, this evidence was used in no small part to establish that "the sale of consumable office supplies through office superstores" was a relevant product market, rather than the broader market for retail sales of office supplies, as alleged by the merging parties.[157]

To take another example, in the Union Pacific/Southern Pacific railroad merger, the DOJ opposed the deal, arguing that freight rates would rise as the number of carriers on many routes declined from 3 to 2 or even 2 to 1. The DOJ supported its claims by presenting cross-sectional regression models showing how freight rates varied with the number of carriers on a route, along with other factors. Based on these reduced-form estimates, freight rate increases of about 20% were estimated on routes going from 2 carriers to 1, and freight rate increases of about 10% were estimated on routes going from 3 to 2 carriers.[158]

More generally, reduced-form methods ask about the relationship between market structure (such as market concentration or the presence of certain companies) and prices or other measures of market performance, without specifying a structural model of the market. These methods require variation in the data on market concentration or competition, and to obtain reliable results one must be careful to correct for other factors that may influence price. Baker and Rubinfeld (1999) provide a broad discussion of the use of reduced-form estimates to predict the effects of horizontal mergers.

4.6.2. Merger simulation

A small industry has arisen in recent years that uses simulation methods to estimate the effects of mergers in markets involving differentiated products. Merger simulation has most commonly been employed to study mergers involving consumer products, for which highly disaggregated retail scanner data on prices and sales are often available. This approach to merger simulation is described in detail and surveyed by Werden and Froeb (2007). Epstein and Rubinfeld (2001, 2004) also provide a very useful discussion of the merger simulation methodology.

[157] See Dalkir and Warren-Boulton (2004) for further economic analysis of the *Staples* case.

[158] See Kwoka and White (2004) for a balanced presentation of this major merger case. They cite (pp. 40–41) industry evidence from 2001, several years after the merger, that shippers on the 2-to-1 routes were paying a 20% to 30% premium, consistent with the DOJ's estimates.

There are two steps to simulating mergers. First, a demand system for the differentiated products must be specified and estimated. Werden, Froeb, and Scheffman (2004) emphasize that the weight given to merger simulations should depend on how well the specified model fits the industry, based on historical evidence.[159] Estimating a demand system for differentiated products can be highly complex and require a great deal of detailed data. A number of methods have been developed to limit the number of parameters that must be estimated. One is to build a model in which demand for the various differentiated products depends on their underlying characteristics, an approach pioneered by Berry (1994) and Berry, Levinsohn, and Pakes (1995), with application to the automobile industry. Nevo (2000a, 2001) applies similar methods in the ready-to-eat cereal industry, and Nevo (2000b) provides a practitioners' guide. Another approach is explored in Hausman, Leonard, and Zona (1994), who employ a multi-stage budgeting procedure, under which products in a market are sorted into sub-groups based on their characteristics and demand is then estimated using this additional structure. Werden and Froeb (1994) use a logit model, which imposes a great deal of structure but requires the estimation of relatively few parameters. Epstein and Rubinfeld (2001) advocate use of the "Proportionality-Calibrated Almost Ideal Demand System" (PCAIDS) model for calibrated demand simulation, which they consider superior to other calibrated-demand models, including the logit model.

Second, the post-merger equilibrium is simulated using the structural model that was fitted to pre-merger data. This involves solving the post-merger equilibrium conditions using the parameters estimated based on pre-merger conditions. Note that this approach assumes that the same model of oligopoly, a Nash equilibrium in prices, applies before and after the merger, so the method is only capable of estimating unilateral effects, not coordinated effects. Peters (2003) evaluates the performance of these methods using data from the U.S. airline industry.

5. Monopolization

Whereas section 3 addressed collusion—when a group of competitors act in the manner of a single firm—and section 4 examined horizontal mergers—when competitors join to form a single firm—here we analyze how competition policy limits the behavior of a preexisting single firm. As mentioned in subsection 2.5, the offense of monopolization under U.S. antitrust law has two requirements: monopoly power and exclusionary practices.[160] Accordingly, we begin by elaborating on the element of monopoly power from an economic perspective, drawing on our broader discussion of market power in section 2 and its application to horizontal mergers in subsection 4.5. Then we consider

[159] For example, one can check to see if the estimated firm-specific elasticities of demand are consistent with the Lerner Index and separately observed measures of marginal cost.

[160] EU competition policy regulates abuse of a dominant position, which is analogous to the anti-monopolization provision in U.S. law and will be mentioned below, mostly in notes in subsection 5.2.

antitrust law on monopolization, with regard to both monopoly power and the exclusionary conduct requirement, viewed generally. Finally, we examine the economics and law as applied to certain important practices, predatory pricing and exclusive dealing, in this relatively controversial realm of competition policy.[161] Myriad additional practices, some the subject of substantial literatures, are not considered here, although some of the principles adduced in our discussions of predatory pricing and exclusive dealing are pertinent.[162]

5.1. Monopoly power: economic approach

5.1.1. Rationale for monopoly power requirement

The monopoly power requirement under U.S. monopolization doctrine is that, as a prerequisite to liability, there must exist a significant degree of market power—how much is required will be considered in subsection 5.2.1. The rationale for market power thresholds has been addressed previously. In subsection 2.5, we identified the screening function, that is, the reduction of false positives. The value of a significant market power screen is particularly important with regard to monopolization for two interrelated reasons: single-firm behavior (which is obviously ubiquitous) is being regulated, and it often is ambiguous whether that behavior is anticompetitive (and most is not). These points deserve some elaboration.

At the heart of a market economy is the principle that firms have free rein to compete aggressively to win business and earn profits, possibly vanquishing their rivals in the process. If one firm does gain a dominant position, that is the firm's just reward for best serving the interests of consumers. Imposing liability on companies that compete most effectively, perhaps to the point of driving their rivals out of business, would contravene

[161] Crandall and Winston (2003) argue that significant consumer benefits did not result from the remedies ordered in a number of the most visible government enforcement actions under Sherman Act Section 2 (the monopolization provision) during the twentieth century. Even if these results are accepted, their approach does not tell us about the deterrence benefits that result from inducing changes in the behavior of monopolists. In a companion piece, Baker (2003a) reviews evidence of anticompetitive outcomes before the enactment of the Sherman Act, during the 1890–1910 period when enforcement of Section 2 was often ineffectual, and from other countries without comparable laws.

[162] The practice of tying, under which a firm requires customers who are purchasing product A also to purchase product B, has received a great deal of attention in the economics literature and in the law. Nalebuff (2003) provides a clear and accessible explanation of the complex economic issues that arise when evaluating the effects of tying as well as multi-product discounts, also known as bundling. Tirole (2005) provides a practitioner-oriented introduction to some of the issues that arise in the area of tying. Whinston (1990) examines the strategic use of tying to foreclose competitors in an imperfectly competitive market for the tied good. Refusals to deal, under which a vertically integrated firm refuses to sell its upstream input to its downstream rivals, are another practice that has been studied extensively. Rey and Tirole (2007) examine a range of strategies of what they call "vertical foreclosure," including the denial to rivals of access to a bottleneck input, as well as "horizontal foreclosure," including bundling and tying. Katz (1989) surveys the literature on vertical contractual practices.

the fundamental workings of a market economy. Furthermore, the argument goes, government intervention is unnecessary because even the most successful companies must continually face the gales of creative destruction, as new and innovative rivals challenge their positions. (And, in exceptional cases of natural monopoly, some form of industry-specific regulation is the answer, not broad-based limits on competitive practices.)

Even most who accept strong forms of this laissez-faire view would embrace rules against collusion and horizontal mergers to monopoly. A much greater danger, however, is raised when various practices of individual firms are challenged by the government—and, in some jurisdictions like the United States, by private plaintiffs (often unsuccessful rivals). Here, even those highly skeptical of extreme laissez-faire recognize the potentially high costs of litigation, of erroneous condemnation of benign or affirmatively beneficial practices, and perhaps most importantly of chilling routine competitive behavior. As noted, the risks are especially great because it is often difficult to distinguish exclusionary from pro-competitive conduct, a subject to which we will return in subsections 5.2–5.4.

Fortunately, for most firms in most industries, the danger of socially costly anticompetitive behavior is negligible because the firms lack significant market power or any serious prospect of acquiring it even using the challenged practices. Accordingly, for such firms, there is likely to be little benefit from examining in detail the effects of their conduct, whereas substantial costs of administration, mistaken prohibition, and inhibition of competitive vigor can be avoided by in essence granting them immunity. To give a concrete example, imposing a monopoly power screen in the area of predatory pricing avoids potentially enormous costs that could arise if every firm contemplating an aggressive low-price strategy had to fear a possible predatory pricing challenge from its rivals.

The monopoly power requirement is similar to the rule examined in subsection 4.4.2 that horizontal mergers must exceed some threshold level of anticompetitive effects as a prerequisite to liability. With mergers, some level of synergies are presumed to exist in typical cases, so scrutiny is only triggered when anticompetitive effects are nontrivial. With monopolization, the threshold is ordinarily understood to be much higher, not so much because single-firm practices are ordinarily far more valuable than mergers, but rather because of the more substantial problem of false positives and related chilling effects.

Another point bearing on the value of a monopoly power requirement concerns the magnitude of incremental harm if prices do rise. If there is no technical market power as defined in section 2 (that is, if price is at marginal cost), the marginal loss in total surplus as price begins to rise is zero. The greater is the extent of initial market power—that is, the higher the initial margin—the greater is the marginal distortion from further price increases. Note, however, that this observation is pertinent under a total welfare standard rather than a consumer welfare standard, on which see subsection 4.4.3. After all, the marginal reduction in consumer surplus is highest when price rises from the point at which it equals marginal cost; the higher is the initial price, the smaller is the

incremental reduction in consumer surplus. (The incremental reduction per unit of price increase is just the quantity demanded, which falls with price.)

Application of a monopoly power test raises a number of issues. First, how high should the market power requirement be? Unfortunately, this question is difficult to answer because it is so hard to measure most of the costs and benefits of a higher threshold, notably, the level of chilling effects and the relative proportions of beneficial and undesirable practices among those that are deterred. As we suggested in subsection 2.5, it seems optimal for the threshold to depend on the practice: for those obviously undesirable, little or no threshold seems necessary; for those more questionable, some intermediate standard; and for those sufficiently likely to be beneficial, an extremely high one—at some point being tantamount to deeming the practice legal *per se*. Again, to compare with horizontal mergers, many economies that may be realized through mergers can, although perhaps more slowly and at somewhat greater cost, be obtained through internal growth. But if innovation or aggressive, competitive pricing by individual firms is deterred, alternative outlets seem unavailable.

Another factor bearing on the height of the monopoly power threshold concerns the cost and potential for error in the market power inquiry itself. As discussed in subsection 2.4 (and elaborated in subsections 4.5 and 4.6 with regard to horizontal mergers), there are numerous means of assessing a firm's market power that vary greatly across markets in their feasibility and reliability.[163] If a practice seems fairly clearly evil, it neither saves enforcement costs nor significantly reduces false positives to impose much (or any) market power requirement.

We also discussed how conduct itself may be highly probative of market power in cases in which the conduct would not be rational in its absence. Requiring proof of power without taking into account such conduct makes little sense, and if the conduct is, logically, used to infer sufficient power, then the monopoly power requirement is not serving as an independent threshold test. From an economic viewpoint, this is an appropriate result. The danger arises when the practice is more uncertain regarding whether it is desirable, is in fact being employed, or its use necessarily implies the existence of significant market power. (These points are usefully reconsidered in the case of predatory pricing, taken up in subsection 5.3.)

5.1.2. Application to challenged practices

Subsection 5.1.1 addresses the purpose of a market power threshold in monopolization cases. We have not yet, however, revisited the important question that we raised in subsection 2.5 about whether the requisite monopoly power is that which exists but for the challenged practices or in light of them, and what if any is the relevance of the difference between these two levels of market power. For horizontal mergers, recall from

[163] Indeed, much of what economics has to offer antitrust law concerns the assessment of market power. Because of our extensive treatment of the subject in the previous sections, we do not take up the matter further here.

subsection 4.4.2 that the Merger Guidelines (in the United States, and similarly in the European Union) impose a two-part test that requires both that the post-merger HHI be sufficiently high and that the increase in HHI due to the merger be at least of a certain magnitude (which itself depends on the former measure). The analogue for monopolization would be to require that there exist monopoly power with the challenged practices and also that the practices contribute appreciably to that power. We defer discussion of the current state of the law to subsection 5.2.1; here, we consider what economic analysis has to say about these issues.

At first glance, it might appear that only the increment should matter, for what is challenged is a set of practices, not the means by which the firm had previously gained its position. Thus, if a firm with a valid and powerful patent engages in a practice that slightly increases its monopoly power, ordinarily all that would be enjoined would be the illegitimate practice; likewise, fines or damages would be based on the addition to power, not the whole of profits legitimately attributable to the patent itself. Accordingly, it seems that the key question is not the disembodied query "How much economic power does the defendant have?" but rather "Will the challenged practices harm competition?" or "Will the challenged practices enhance the defendant's market power?" That well-defined economic question often can be usefully recast as: "Will the challenged practices significantly remove or relax constraints on the defendant's pricing?"

Our analysis in subsection 5.1.1, however, suggests that this should not be the sole inquiry regarding market power. The screening function and the related problems of false positives and chilling effects indicate that we probably should not freely allow challenges in industries where there is little market power or against particular firms with little power (although if the increment were sufficiently large, say, all the way to a monopoly, that would be another matter). Also, if prices are near marginal cost, price increases cause little deadweight loss.[164]

Taken together, for many practices (setting aside those that are unambiguously undesirable), it may make sense both to insist that the firm possess some significant level of market power and that the challenged practices contribute importantly to it. Regarding the former, it often would not much matter whether the overall level of market power was measured with or without the challenged practices, unless they had a very large impact. In such cases, it probably makes sense to consider power with the practices, although if power is quite low without them and the practices themselves are ambiguous, one may be skeptical of their effects and much of the screening function would be lost if extravagant claims were permitted against firms that had little market power. This assessment, however, imagines a prospective challenge against practices that have not yet had their (alleged) impact. If, instead, the practices have had time to take effect and the result is substantial market power, it hardly seems sensible to excuse the defendant that asserts its power would be small without the practices, for that would be an admission of large anticompetitive effects.

[164] Recall from section 2, however, that when there are significant fixed costs, equilibria will be characterized by high margins, so the marginal welfare cost of a price increase would be nontrivial.

Consider further the notion that the challenged practices themselves should be shown to contribute significantly to the firm's market power. Just as in the case with horizontal mergers, if there were no possible benefits that might accompany the practices, it would seem that any increment (perhaps beyond a *de minimis* level) should be condemned. There are two important reservations to this conclusion in the monopolization context. First, as with horizontal mergers, some practices that may have anticompetitive effects may also promote efficiency. Consider, for example, the possible tradeoffs that may be involved with exclusive dealing, as we discuss in subsection 5.4. Then, one would need to balance the two, and how that balance should be conducted would importantly depend on whether the standard is limited to consumer surplus or is defined in terms of total economic welfare, as we discussed in subsection 4.4.3. Second, there is again the problem of uncertainty and false positives. If we are uncertain about whether challenged practices are undesirable at all, then depending on our Bayesian prior about the likelihood of different effects and the evidence before us, it may well be that the appropriate loss function is minimized by requiring that apparent anticompetitive effects be above some magnitude before condemning the practices.

Of course, applying this additional test regarding the increment to market power is not without cost because such an inquiry requires additional information about the conduct at issue. When practices are ambiguous—which is when screens tend to be most helpful—it will often be uncertain whether there is any anticompetitive effect, so one might wonder how its impact might be quantified. An answer is that sometimes it can be quantified conditionally; that is, one can assume that the practices have some specified type of effect and then attempt to quantify what that effect would be.

To apply these suggestions more concretely, it is useful to return to the framework introduced in section 2 on market power and consider an example. In subsection 2.4.2.2, when discussing substitutes, we derived equation (5), showing that the elasticity of demand for a given product (product 1 of the N products) is equal to one plus the sum of the cross-elasticities of all the other products with that product, each weighted by that product's share of expenditures: $|\varepsilon_{11}| = 1 + \frac{1}{s_1}\sum_{i=2}^{N} s_i\varepsilon_{i1}$. According to equation (1), the profit-maximizing markup for a single firm producing product 1 is given by $m_1 = 1/|\varepsilon_{11}|$. The expression for the elasticity of demand captures the familiar idea that the firm's market power is increased if the ability of consumers to shift to certain substitutes is reduced. Practices that reduce a number of these cross-elasticities ε_{i1} reduce the magnitude of the elasticity of demand for the product in question, $|\varepsilon_{11}|$, giving the firm greater market power and thereby raising the firm's profit-maximizing price.

Suppose that the practices at issue are alleged to reduce the attractiveness of some of the substitutes to the firm's product. For example, predatory pricing might eliminate one or more of the substitute products from the market. In principle, a full inquiry would allow us to determine that the practices reduce the cross-elasticity of demand with substitute i at preexisting prices from ε_{i1} to $\hat{\varepsilon}_{i1} \leq \varepsilon_{i1}$. The full inquiry thus would tell us that the practices reduce the magnitude of the elasticity of demand facing the firm

at preexisting prices from $|\varepsilon_{11}|$ to $|\hat{\varepsilon}_{11}| = 1 + \frac{1}{s_1}\sum_{i=2}^{N} s_i \hat{\varepsilon}_{i1}$. Since $|\hat{\varepsilon}_{11}| < |\varepsilon_{11}|$, the firm's profit-maximizing price will rise as a consequence of the practices.[165]

Now return to our questions about levels of market power versus changes in power due to the challenged practices. Regarding the monopoly power threshold, if the magnitude of the elasticity of demand facing the firm without the challenged practices, $|\varepsilon_{11}|$, is sufficiently high, no liability can arise if the test looks to preexisting power. As noted, the rationale would be that a firm without significant market power initially cannot profitably create such power by engaging in the practices at issue, leading us to be skeptical about the alleged anticompetitiveness of the practice. Under this approach, if $|\varepsilon_{11}|$ is sufficiently high, one need not look at $|\hat{\varepsilon}_{11}|$. Alternatively, if the question is power with the practice—which is more natural to employ if the practice has existed for awhile—the question would be whether $|\hat{\varepsilon}_{11}|$ is sufficiently high.

We also considered the relevance of the extent to which the practices at issue may plausibly enhance the firm's market power, which focuses on the difference between $|\hat{\varepsilon}_{11}|$ and $|\varepsilon_{11}|$, not on the level of either in isolation. To see how this difference might be measured, suppose that the challenged practices only affect a certain subset of substitutes, J. At worst, the practice would eliminate those substitutes as alternatives available to consumers, which is equivalent to setting $\varepsilon_{1j} = 0$ for all $j \in J$. Without these substitutes, the magnitude of the firm's elasticity of demand at preexisting prices would fall to $|\hat{\varepsilon}_{11}| = 1 + \frac{1}{s_1}\sum_{i \notin J} s_i \varepsilon_{i1}$; of course $|\hat{\varepsilon}_{11}| < |\varepsilon_{11}|$. If the gap $|\varepsilon_{11}| - |\hat{\varepsilon}_{11}|$ is small, so the challenged practices only modestly reduce the magnitude of the firm's elasticity of demand at preexisting prices, then these practices, even if effective, cannot lead to a significant increase in market power.[166]

5.2. *Legal approach to monopolization*

As we noted in subsection 2.5 on market power, under U.S. Sherman Act Section 2, the Supreme Court has stated that "[t]he offense of monopoly ... has two elements: (1) the possession of monopoly power in the relevant market and (2) the willful acquisition or maintenance of that power as distinguished from growth or development as a consequence of a superior product, business acumen, or historic accident."[167] Here we consider each in turn.

[165] Consumers not only would be harmed by the firm's price increase but they would suffer a reduction in utility from using the substitute products as a consequence of the challenged practice. When we study exclusive dealing in subsection 5.4, we explain how a firm's conduct can make substitute products less attractive and whether there are offsetting benefits to consumers. Efficiencies could be included in this analysis by allowing the practices at issue to reduce the firm's marginal cost as well as its elasticity of demand, so the firm's price might fall even as its markup rises.

[166] This approach is analogous to the hypothetical-monopolist test that is used for horizontal mergers (see subsection 4.5.1) where we imagine that the hypothetical monopolist controls all of the products affected by the challenged practices.

[167] United States v. Grinnell Corp., 384 U.S. 563, 570–71 (1966). We focus on the antitrust violation of monopolization; attempted monopolization, which is noted briefly below, is similar but places less weight

5.2.1. Monopoly power

The central legal question is how much market power is denoted by "monopoly power." The just-quoted authoritative statement from the Supreme Court does not answer this question. Unfortunately, even those cases that offer quantitative statements are far less illuminating than meets the eye.[168]

Most famous is the pronouncement in *Alcoa* that a ninety percent share (in the market for aluminum) "is enough to constitute a monopoly; it is doubtful whether sixty or sixty-four percent would be enough; and certainly thirty-three per cent is not."[169] The difficulty in interpreting this statement is that two distinct issues are conflated: how much market power was thought to exist in that case (a fact question distinctive to that industry under the then-existing conditions), and how much market power is deemed sufficient to constitute monopoly power (a legal/policy question, the answer to which may be entirely independent of the particular case or, if not, its dependence requires specification that was not offered). For example, might the court have thought that monopoly power consists of the ability to sustain a margin of at least 20%, that a 90% share conferred the power to price 35% above cost, a 33% share only 10% above cost, and a 60–64% share somewhere near 20%? Or might it have thought that monopoly power required only the ability to sustain a 10% margin, but that the power implied by each of the stated shares was only half as high?

This ambiguity is fundamental because in future cases, not in the aluminum industry (under the conditions prevailing at the time of *Alcoa*), a given share, whether 33%, 90%, or some other figure, may convey much more or significantly less power than did a similar share in *Alcoa*. But we know neither how much power over price *Alcoa* required nor how much power was thought to exist for any given share in that industry. Hence, even if both parties' experts in a subsequent case agree that, say, the sustainable margin was 16%, there is no way to tell from *Alcoa* which side wins on the element of monopoly power.

The same opacity characterizes all statements that a given market share is or is not adequate under any market power test—that is, unless one accepts that a given share in a properly defined market conveys the same market power, regardless of the market.[170]

on the defendant's current (versus prospective) monopoly position. We also note that in the European Union Article 82 prohibits "Any abuse by one or more undertakings of a dominant position," which requires an analogue to both monopoly power (dominant position) and exclusionary conduct (abuse). As we will note at various points below, however, the interpretations have differed, although the divergence seems to be shrinking over time.

[168] For prior discussion and questions, see Areeda, Kaplow, and Edlin (2004, sec. 3.B).

[169] United States v. Aluminum Co. of America, 148 F.2d 416, 424 (2d Cir. 1945). Although this was a Court of Appeals case, it was decided by a prominent panel of judges, the opinion was written by the famous Judge Learned Hand, the court was acting in lieu of the Supreme Court, and the opinion was blessed in subsequent Supreme Court cases.

[170] In the European Union, there is an even greater tendency to rely on market share in cases alleging abuse of a dominant position, although if its merger law is any indication of a general trend, there seems to be

But this supposition is emphatically false. Instead, as developed in subsection 2.2 and reemphasized in subsection 4.5.1 on market definition in horizontal merger cases, in the basic case a firm's price-cost margin is the inverse of the absolute value of the firm-specific elasticity of demand, $|\varepsilon_F|$, which in turn is given by equation (2),

$$|\varepsilon_F| = \frac{|\varepsilon_D| + (1 - S)\varepsilon_R}{S}.$$

Thus, in addition to the market share S, both the market elasticity of demand, ε_D, and the elasticity of supply response, ε_R, are important, and we have seen that it is quite possible for a high share to be associated with low market power and a modest share to be associated with substantial market power.

Furthermore, as discussed at some length in subsection 2.4.2.2, there is no means of defining markets in such a way as to circumvent this problem. Assuming some hypothetical, benchmark market in which a stipulated level of market power is associated with each market share, it would only be by chance that there would be a readily available market definition in any given case that would yield a share that indicates the correct amount of market power. This inability relates to the all-or-nothing nature of market definition—products or regions are either "in" or "out"—and the fact that the market elasticity of demand is, as noted in subsection 5.1.2, one plus the (revenue-share) weighted sum of the cross-elasticities of all products, rather than being determined solely (or even primarily) by the cross-elasticities of one or two products. Note also that, even if one could match every actual market to such a hypothetical benchmark market, we have no way of knowing how the aluminum industry in *Alcoa* or other industries and markets in other antitrust cases relate to that imaginary market.[171]

Before proceeding, it should be noted that the foregoing problem does not rule out the utility of the standard practice of determining market power by defining a relevant market and then measuring the firm's share in that market. As we discussed, consistent with the formula for a dominant firm's elasticity of demand, these steps need to be supplemented. At best, courts tend to do this indirectly. Thus, they attempt to define the relevant market such that the market elasticity of demand is not so great that high market shares in that market are consistent with negligible market power. Additionally, they check for the presence of some entry barriers, an aspect of possible supply substitution. Even if this is done well, however, and even if there is no better way to determine market power in a given case, one still needs to know how much market power is required, which is the question at hand.

a tendency to move toward a more economic approach to the assessment of market power, as we note in subsection 4.4.2. In any event, EU cases suggest that a 50% market share may well be enough, and for some practices even lower shares might be accepted.

[171] As should be clear, the problem we identify is not unique to *Alcoa*. Consider any case that states, say, that a 50% market share is required. If that statement is to be associated with the market power that exists in the case at hand, then our discussion of *Alcoa* applies directly. If it is to relate to the level of power that exists in a "typical" market, we need to know what that market is and how much power is implied.

Another problem with the failure to state the underlying market power threshold explicitly is that there is no way to relate the many other means of measuring market power that we examined in subsection 2.4 to pronouncements about market share. In this regard, one can contrast the approach under the Horizontal Merger Guidelines under which "the Agency, in most contexts, will use a price increase of five percent lasting for the foreseeable future" to define the SSNIP. See subsection 4.5.1. Under such an explicit approach, it is possible to use the sorts of empirical techniques described in subsection 4.6 to ascertain market power, such as in the *Staples* case where the government presented and the court was persuaded by evidence of price differences across regions with different numbers of competitors. As already noted, even if both sides were to agree on the level of market power in a monopolization case, we cannot tell from existing statements whether monopoly power would be deemed to exist. And since no quantitative threshold has been stated, there is no way of determining whether it reflects an appropriate balance of screening benefits, litigation costs, and so forth.

Consider next the relationship between the monopoly power requirement and the challenged practices. One issue is whether exclusionary conduct may be used as a basis for inferring monopoly power. Courts seem to contemplate that nearly any relevant evidence will be admitted but nevertheless are reluctant to find a violation unless the monopoly power requirement is established through proof of a relevant market in which substantial power exists. On one hand, it is rational to insist on proof of power—given the purpose of the monopoly power screen, to avoid false positives and related chilling effects—in cases when there is ambiguity about the practices under scrutiny. As we mentioned in note 27 in subsection 2.4.3 on inferring power from conduct, the Supreme Court's insistence in *Spectrum Sports* that a plaintiff in an attempted monopolization case must independently prove the requisite power may reflect the fact that the challenged act involved terminating the plaintiff distributor in favor of another, a practice that hardly evidences such power (although it seems to have convinced a jury).[172] Requiring independent demonstration of a predator's ability to recoup losses, to be discussed in subsection 5.3.3, also seems to reflect skepticism about whether truly predatory pricing can confidently be identified. As already noted, however, if a practice is unambiguously exclusionary, there is reason to infer some market power, and there is little reason to impose a strong filter.[173]

We also addressed whether it makes economic sense to focus on extant market power (whether with or in the absence of the challenged practices) rather than on how much practices contribute to that power. The monopoly power requirement itself seems focused on extant power. However, in showing that practices are exclusionary—on which more in subsection 5.2.2—it is required that they have contributed to that power. This

[172] Spectrum Sports v. McQuillan, 506 U.S. 447 (1993).

[173] It is independently problematic that courts often seem to insist on defining a relevant market, which we have seen is only an aspect of one means of inferring market power—and not always the best means—although we also have noted that more recently courts have tended to accept more direct evidence when it is offered and found to be persuasive.

demand is viewed as part of the determination of the second monopolization element rather than as part of the first. To illustrate the relevance of increments to market power, in the U.S. government's settlement with Microsoft in the first wave of enforcement activity in the mid-1990s, the DOJ agreed to modest conduct remedies (which some commentators and intervenors viewed as too lax) precisely because DOJ insisted that only a moderate portion of Microsoft's market power in operating systems was attributable to the challenged conduct.[174] The issue also arises in claims of attempted monopolization. To demonstrate the required dangerous probability of success, it is necessary to show that the practices will contribute appreciably to market power since the defendant in such cases is not yet imagined to have monopoly power.

In subsection 5.1.2, we also briefly discussed whether monopoly power is to be gauged with or without the challenged practices. In many cases, the practices have been in place sufficiently long that the status quo plausibly reflects their effects (if any). This setting has sometimes led to confusion, most notoriously in the *du Pont* (*Cellophane*) case.[175] In defining a broad market—not just cellophane but also other flexible wrapping materials—the Supreme Court was heavily moved by the fact that many customers already used various alternatives to cellophane. As we explained in subsection 2.5, however, such a view implicitly asks whether the firm could profitably raise prices significantly above present levels. The answer to that question will almost certainly be negative even if power is great, for if higher prices were profitable, they would already be observed. Furthermore, if the firm-specific elasticity of demand is less than one, it is necessarily profitable to raise price, and to keep doing so until one hits a region of the demand curve that is more elastic. (The margin, given in expression (1), equals the inverse of the magnitude of the firm's demand elasticity; since the margin is $m \equiv (P-MC)/P$, a finite P implies that the magnitude of the elasticity must exceed 1.) Since supply substitution was limited in *Cellophane*, it must have been that price was in an elastic region of the demand curve for cellophane; hence, profit-maximization implies that price must have been high enough that significant substitution occurred. This problem, the so-called "*Cellophane* fallacy," reflects that, although courts have long viewed market power in terms of the ability to elevate price, they have only gradually incorporated economic analysis that bears on how the degree of market power is determined.[176] We also note that, to a substantial extent, the fault does not lie with the courts but rather with litigants who have not made full, cogent use of economic teachings. In

[174] United States v. Microsoft Corp., 56 F.3d 1448 (D.C. Cir. 1995).

[175] United States v. E.I. du Pont de Nemours & Co., 351 U.S. 377 (1956).

[176] Consider also our discussion in subsection 4.5.2 of geographic markets concerning patient flows in hospital merger cases and imports more broadly. For example, with homogeneous goods, common production costs, and positive transportation costs, the presence of imports implies that local producers' mark-up equals transportation costs, which may be significant; the absence of any imports implies that the mark-up is less than transportation costs, suggesting that the elasticity of local demand or rivals' supply exercises more of a pricing restraint. Hence, it would be a mistake to infer that the existence of imports—a readily observable form of substitution—implies less market power than would be implied by their absence.

cases such as *Staples* in which courts have been offered direct and persuasive evidence and analysis, they often seem ready to accept it, even if their decision is still articulated using more traditional rubrics (in that instance, by defining a narrow market upon being presented evidence of sufficient market power).[177]

5.2.2. *Exclusionary practices*

Most economic analysis of exclusionary practices focuses on particular types of conduct, such as predatory pricing and exclusive dealing, which we take up in subsections 5.3 and 5.4, respectively. Before considering specific applications, however, it is useful to consider the law's general formulation of the second element of a monopolization claim, with an eye to how it might be given an economic interpretation. Unfortunately, this aspect of monopolization law is also rather obscure.

The sort of authoritative statement with which subsection 5.2 began illustrates the problem. The quoted Supreme Court language refers to "the willful acquisition or maintenance of [monopoly] power as distinguished from growth or development as a consequence of a superior product, business acumen, or historic accident."[178] The latter portion of this clause does clearly indicate that the term "exclusionary practices," often used to capture this requirement, cannot be taken on its face, for superior products and business acumen tend to exclude inferior competitors from the market.[179] Thus, the test refers only to some subset of practices that exclude. But what subset? The quotation speaks of power acquired or maintained willfully, but that limitation is most unhelpful, for it suggests that only accidental or mistaken behavior is exonerated.

Another common formulation distinguishes between competition that is on the "merits" and that which is not. The reference to merits, however, is patently question-begging.[180] The use of the term competition is suggestive of the rule of reason's test of whether arrangements promote or suppress competition. As we discussed in subsection 3.5.2, that test seems often (but not always) to suggest a process orientation toward the meaning of competition rather than a focus on the results of competition in terms of

[177] Another issue is the extent to which monopolization policy is guided by factors other than considerations of economic welfare (in total or that of consumers alone), a topic considered further in our discussion of the law on exclusionary practices that follows. This consideration is probably of greater contemporary importance under EU law on the abuse of a dominant position. See Hawk (1988).

[178] As elsewhere, our focus will be on U.S. antitrust law; the EU prohibition on the abuse of a dominant position suffers similar ambiguity. Additional specifications under EU law suggest a broader scope that, if taken literally, might include the ordinary monopoly behavior of elevating price and reducing output. Although the EU's prohibition does seem to be interpreted more broadly than is Sherman Act Section 2, it is not given such breadth.

[179] The term is also under-inclusive. A horizontal merger by a firm with monopoly power may be considered to be an act of monopolization, but it would not ordinarily be characterized as exclusionary (although one could state that the formerly independent firm has thereby been excluded from the market).

[180] Similarly problematic are statements regarding the EU "abuse" requirement that it refers to other than "normal" competition.

economic welfare. Given that the rule of reason is employed under Sherman Act Section 1 and that monopolization is prohibited by Section 2 of the same Act, as well as the growing convergence of interpretation in the United States under all of its antitrust statutes, this possible connection between these two locutions is worth keeping in mind.

Commentators have recognized the ambiguity of the exclusionary practices requirement (both in the U.S. law of monopolization and the most closely corresponding EU law on abuse of a dominant position), and they have proposed a variety of ways to give meaning to the second element of monopolization charges.[181] These alternatives are often of a more explicitly economic nature. Suggestions include a focus on consumer welfare, an inquiry into whether a firm has sacrificed short-term profits, an examination of whether a practice makes no economic sense but for its effect of excluding a rival, and an assessment of whether more efficient rivals are or can be excluded.

To obtain an overview of some of the choices involved in picking a general test, it is useful to consider briefly the core question of what price a monopolist is permitted to charge (anticipating somewhat our discussion of predatory pricing in subsection 5.3). To begin, it is well established that a monopolist is generally permitted to charge the classically defined monopoly price. Since this price tends to minimize both consumer and total surplus (relative to lower prices, that is), this broad permission seems inconsistent with (indeed, in contradiction to) an economic welfare standard.

There are two primary justifications for allowing monopoly pricing, and they provide insights into how one should think about the exclusionary practices requirement. First, monopoly profits often reward socially valuable *ex ante* investments, such as in innovation, cost-cutting, or generally running a tight ship. Thus, a dynamic view of welfare is adopted. Note, however, that this view is embraced broadly; defendants are not required to prove in particular cases of high pricing that their profits are efficient *ex post* rewards for prior good behavior. Second, price regulation is not thought to be in the institutional competence of courts or generalist antitrust regulators. To be sure, in certain cases (notably, natural monopoly), comprehensive price (and other) regulation is employed, but under the supervision of specialized agencies. These points are mutually reinforcing in that it is thought to be inefficient for inexpert institutions to intrusively interfere with single-firm behavior at the risk of dynamic efficiency. But the points are also in tension, for if the pertinent calculus is a complex, dynamic one, and the agencies and courts charged with the antitrust task have limited skills, how are they to identify which practices should be prohibited?

Having set aside the notion that a pure, general welfare test (whether total or consumer welfare) is to govern, some of the other proposed tests can be placed in context. Inquiring solely into whether there is a short-run profit sacrifice is obviously problematic because any investment, whether in innovation, new plant, or even routine employee training, involves sacrificing short-run profits for long-term gain. Asking whether a

[181] See, for example, Areeda and Hovenkamp (2002, vols. 3, 3A), Elhauge (2003a), Melamed (2006), Popofsky (2006), Posner (2001, pp. 194–95), Salop (2006), Vickers (2005), and Werden (2006).

practice is profitable but for its exclusionary effects is similarly deficient, for better products and superior service are profitable in significant part due to their tendency to capture business from rivals. (A firm researching a new cure would anticipate no revenue if no consumers would switch from existing treatments.) In addition, not all anticompetitive practices involve a short-term sacrifice.[182] Banning only practices that keep out more efficient rivals might be consistent with maximizing productive efficiency but gives no weight to consumer surplus and accordingly would often permit practices that reduce total surplus as well.[183] The contrast is nicely posed by a case we will consider in subsection 5.3: if a firm's pricing strategy drives out (or deters) slightly less efficient entrants, with the result that prices are substantially higher and consumer and total welfare significantly lower, should a violation be deemed to occur? Better results may be possible by refining or mixing these tests, such as by insisting that a practice be profitable but for the possible additional price increment one can charge if a rival has been excluded from the market rather than hypothetically remaining and continuing to offer its product.

No simple solution is readily apparent. Any general test—or, in its absence, particular tests for particular practices—should probably be grounded in concerns for long-run economic welfare and a recognition of courts' and agencies' limited capacities. As we discussed earlier, the rule of reason's focus on whether arrangements promote or suppress competition—often applied with a process rather than outcome orientation—is rationalized in part by the view that, although we care about the results of competition, it may often be easier for antitrust authorities to assess the process. Accordingly, we mentioned that courts often respond to arguments for exceptions—assertions that challenged behavior, despite being anticompetitive, is socially valuable—by stating that they should be addressed to the legislature (or, in some jurisdictions, to the enforcement agency that is authorized to promulgate exceptions).[184] In addition, the rule of reason is sometimes interpreted to imply more specialized rules, such as the *per se* prohibition on price-fixing. Likewise, for monopolization, it may readily be optimal to employ different, more specific tests for certain practices, wherein those tests as well as the general rule applicable when no such test has been developed are derived from broader guiding principles.[185]

[182] We offer an example below in subsection 5.4.3 when discussing exclusive dealing.

[183] One also notes the inconsistency with the tendency to focus solely on consumer surplus in the horizontal merger context, as discussed in subsection 4.4.3.

[184] In the United States, where cases are often decided by lay juries who one supposes may find conflicting expert testimony to be confusing, there is added pressure to limit which cases can reach juries for fear of excessive false positives. Many court decisions in recent decades seem in part motivated by this consideration. Yet this approach raises problems because there is a single antitrust law that is equally applicable to decisions by the FTC, a specialized agency. Given that the FTC is independently authorized to enforce its own statute, courts could draw distinctions that give the FTC more flexibility, but they have tended not to do so. In other countries, where juries are not used and decisions in the first instance are typically made by competition authorities rather than general courts, the appropriate legal rules may well differ.

[185] And it indeed seems to be the case that rules on different exclusionary practices vary, as illustrated by our discussions of the legal tests for predatory pricing and exclusive dealing in subsections 5.3.3 and 5.4.3, below.

Competition policy is not seen as comprehensive regulation but as merely offering what may be viewed as rules of the competitive game. Antitrust law does not dictate players' specific moves, but certain types of behavior are prohibited. When such prohibitions can be stated in simple, general terms—like prohibitions on naked price-fixing—this strategy works well. When firms' behavior is more complex and subtle, often involving dynamic considerations—a common state of affairs in monopolization cases—the task is more daunting. The challenge is especially great because, as we emphasized in our discussion of the monopoly power requirement, the law seeks to avoid excessive administrative costs, false positives, and perhaps most important the chilling of socially valuable business activity, objectives that are not easily achieved when the law is highly uncertain. Indeed, it is precisely the difficulty in defining and identifying exclusionary conduct that is seen as justifying the monopoly power screen that we examined in subsection 5.2.1.[186]

5.3. Predatory pricing

Predatory pricing is one of the most storied areas of antitrust law. Indeed, the Sherman Act resulted in no small part from concerns about predatory practices, and predatory practices were central to the 1911 *Standard Oil* case, which gave form to Section 2.[187] After a century of debate, the antitrust treatment of predatory pricing still elicits strong reactions. Supporters of tough limitations on predatory pricing believe that they are necessary to prevent large, powerful firms from using their market positions and financial strength to deter entry and to drive existing smaller, weaker rivals from the market, thus fortifying their monopoly power. Skeptics argue with equal vigor that price cutting is the essence of competition, that imposing antitrust liability on a firm for setting its prices too low should only been done with great caution if ever, and that successful predation, if it happens at all, is extremely rare. Here we consider the relevant economic theory, empirical evidence, and appropriate legal test.

[186] The discussion in the present subsection has emphasized difficulties in defining the standard, whereas in practice difficulties in proving what actually happened are often far greater. These problems will be addressed somewhat in the discussions to follow of predatory pricing and exclusive dealing. In attempting to differentiate efficient from anticompetitive behavior (however defined), most attention will be devoted to evidence bearing on the consequences of the practices under scrutiny. Another channel of proof, which is promising but also fraught with pitfalls, involves examining the defendant's internal decision making, which is sometimes done under the rubric of inquiries into intent. The promise is that many complex strategies with anticompetitive or efficiency-enhancing effects (or both) cannot be analyzed and implemented in a large firm without extensive communications that seem, in modern times, difficult to undertake without leaving paper and/or electronic trails. The pitfall is that, especially with decision making by lay juries, aggressive rhetoric ("we will crush the competition") that is logically quite consistent with efficient behavior (for example, trimming costs, improving quality, marketing, and service) can, taken out of context, be mistaken for evidence of anticompetitive designs.

[187] Standard Oil Co. of New Jersey v. United States, 221 U.S. 1 (1911).

5.3.1. Economic theory

McGee (1958) initiated the Chicago School attack on traditional concerns about predatory pricing by strongly challenging whether Standard Oil had in fact engaged in predatory pricing, as was commonly believed at the time. McGee also argued, as a theoretical matter, that predatory pricing would only be an optimal strategy under very stringent conditions that are rarely if ever met. First, he emphasized that it will typically be more profitable for a monopolist to acquire its rivals than to drive them out of business. However, this argument fails because such horizontal mergers would likely violate the antitrust laws (see section 4) and, even if permitted, the monopolist might be able to acquire its smaller rivals on more favorable terms if it can establish a reputation as a predator (a possibility that McGee discounts). Second, McGee pointed out that a monopolist might well lose more money than its smaller rivals during a period of predation because it is selling more units. This is an important point, but it does not eliminate the possibility of profitable predation, especially predation that involves price discrimination, with price cuts targeted at the prey's customers, or predation that establishes a reputation. Third, McGee noted that driving a rival from the market might do little to increase the predator's market power if the prey's productive assets remain intact and available to a new entrant. Lastly, McGee asked why the prey could not credibly survive the predation by drawing on internal funds or borrowing as needed.

Subsequently, there has arisen a burgeoning literature on the economics of predation, which has been surveyed by Ordover and Saloner (1989). We briefly note a few highlights. The early theory of predatory pricing was based on the superior financial resources of the incumbent monopolist in comparison with a smaller rival/entrant. Under this "deep pocket" theory, as formalized by Telser (1966) and later Benoit (1984), the smaller firm would earn positive profits, if not for the predation, but the smaller firm had a limited ability to sustain losses before it must exit the market. By depleting the rival's financial resources, or credibly threatening to do so, the predator can induce exit or deter entry. This theory assumes, however, that the prey cannot obtain the financial resources necessary to sustain itself despite the prospect of positive profits that await if it can survive the predation. This assumption would not be justified if a firm with substantial financial resources could enter the market in question or if the smaller firm could obtain a large line of credit to finance its operations. Either prospect would deter predation, and the financial resources would not in fact need to be drawn upon. Yet economic theorists have also explained how asymmetric information between potential lenders and the firm that seeks to borrow funds may interfere with this possibility. Bolton and Scharfstein (1990) show how deep-pocket predation can occur even if the prey and its lenders are sophisticated.

Much theoretical work on whether predation can be economically rational focuses on asymmetric information and uncertain time horizons in attempting to formalize the intuition that predators can develop a reputation that will not only drive the current prey from the market but deter others from entering. To frame the problem, consider the following simple game between a potential entrant and an incumbent firm. First,

the entrant decides whether to incur the sunk costs necessary to enter the market. Next, the incumbent decides whether to accommodate entry or engage in predation. Suppose that the entrant will make positive profits if the incumbent accommodates entry but will lose money if the incumbent engages in predation. Following the Chicago School critique, suppose further that, viewing the problem as a one-shot game, predation is unprofitable—that is, the threat to predate is not credible—for whatever reason; perhaps the entrant will fight for a long period of time before exiting, or perhaps accommodation is quite profitable for the incumbent, so the opportunity cost of predation is high. The only subgame-perfect equilibrium is for the potential entrant to enter the market and for the incumbent to accommodate entry. Furthermore, if this game is repeated in a finite sequence of distinct markets involving the same incumbent, the unique subgame-perfect equilibrium is for the potential entrant in each of these markets to enter and for the incumbent to accommodate every time. Following Selten (1978), this (formerly) counter-intuitive backward-induction argument is known as the "chain-store paradox."

A body of subsequent work has shown that the chain-store paradox may well dissolve when any of a number of realistic dimensions is added. Milgrom and Roberts (1982, appendix A) show that the paradox is an artifact of the known, finite number of potential entrants: predation becomes credible so long as there always remains a sufficient probability that future potential entrants will arrive. And even with a finite number of periods, predation based on reputation is rather easily supported using game theory. Milgrom and Roberts (1982) and Kreps and Wilson (1982), in highly influential work, demonstrate the power of reputation and signaling to support credible predation. These papers rely heavily on the presence of asymmetric information, the essence of predation based on reputation being that the predator is signaling its willingness to engage in predation (for example, predatory behavior may signal low marginal cost). Scharfstein (1984) and Fudenberg and Tirole (1986) show how predation can also work by disrupting the ability of the entrant to determine whether remaining in the market will be profitable.

Collectively, these papers establish that predation can, in theory, be profitable for an established monopolist in a variety of plausible circumstances. This body of sophisticated theoretical work, however, cannot resolve the debate about whether predatory pricing is in fact a widespread threat to competition that must be met with vigorous antitrust enforcement or instead constitutes a phantom practice that rarely if ever occurs.

5.3.2. Empirical evidence

The empirical evidence on predation has been hotly disputed since at least the 1950s. Regarding the landmark *Standard Oil* case, McGee (1958, p. 168) writes: "Judging from the Record, Standard Oil did not use predatory price discrimination to drive out competing refiners, nor did its pricing policies have that effect." Ever since then, Chicago School proponents have complemented their theoretical attack on the economic logic underlying predatory pricing with the empirical claim that predatory pricing is either

extremely rare or nonexistent.[188] In this tradition, Koller (1971) examines 26 cases, ranging from 1907 through 1965, in which he was able to obtain a substantial trial record and the defendant was found guilty of engaging in predatory pricing. By his count, only seven of these cases involved below-cost pricing with predatory intent; four of these seven cases involved predation to eliminate a rival, and three involved predation to acquire a rival or improve market discipline. Koller considers the predation to have been successful in only four cases.

The response to McGee and Koller began with Yamey (1972), who argued that predatory practices may not be nearly as rare as McGee suggests and provided an example of predation in the China-to-England ocean shipping business around 1890. Zerbe and Cooper (1982) update and expand on Koller's study. Based on their examination of 40 predatory-pricing cases from 1940 through 1981, they recommend a modified version of the Areeda and Turner (1975) rule (see subsection 5.3.3) to prevent predatory pricing, finding that it performs much better than a rule of *per se* legality. In this tradition, a significant empirical literature identifying instances of successful predation has emerged over the past twenty years. Burns (1986) presents evidence that predation by the tobacco trust enabled it to acquire its rivals—those who were targets of the predation and others based on reputational effects—on more favorable terms. Ordover and Saloner's (1989, p. 545) survey directly challenges McGee's conclusions; citing Standard Oil documents, they state: "There is little doubt, however, that Standard Oil at least attempted to use pricing as a weapon to drive its rivals out." Weiman and Levin (1994) argue that the Southern Bell company engaged in predation to protect and build its telephone monopoly. Morton (1997) finds related evidence of deep-pocket predation in merchant shipping. Genesove and Mullin (2006) find evidence of predatory pricing in the U.S. sugar refining industry before World War I. Bolton, Brodley, and Riordan (2000) assemble and discuss the body of empirical evidence of predatory pricing.[189]

In the end, whether this evidence is sufficient to conclude that strong rules against predatory pricing are needed to protect competition is difficult to say. The rarity of predatory pricing convictions in the United States may simply indicate that the law is working well to deter this practice. On this score, it is interesting to note that many of the documented instances of predatory pricing are either prior to 1911, when the *Standard Oil* decision condemned predation, or from outside the United States.

[188] As described by Baker (1994), the Chicago School view of predatory pricing is that it is akin either to a white tiger, an extremely rare creature, or to a unicorn, a complete myth. He calls theories of predation based on reputation effects an example of "Post-Chicago Economics," which gives greater weight to market imperfections, such as those based on incomplete information, than does the Chicago School.

[189] Isaac and Smith (1985) report results from a laboratory experiment designed to see if predatory pricing would arise. In their experiment, a subject controlling a large firm competed against another subject controlling a smaller firm. The firms produced with economies of scale, with the larger firm being more efficient and having superior financial resources. They also included sunk entry costs that would need to be incurred again in the event of exit. Despite these conditions, arguably favorable to predation, the subjects did not employ predatory pricing. On the other hand, Jung, Kagel, and Levin (1994) find frequent predation in their experiment, in which a single monopolist plays a sequence of eight periods against a series of different entrants.

5.3.3. Legal test

The contemporary discussion of rules to control predatory pricing can be dated to a highly influential paper by Areeda and Turner (1975). They expressed concern that the treatment of predatory pricing in the cases and in the literature did not clearly and correctly delineate which practices should be illegal and that fears of predatory pricing were overblown. Areeda and Turner were sharply critical of the case law, stating (p. 699): "Courts in predatory pricing cases have generally turned to such empty formulae as 'below cost' pricing, ruinous competition, or predatory intent in adjudicating liability. These standards provide little, if any, basis for analyzing the predatory pricing offense." Areeda and Turner proposed a test for predation based on whether prices were below average variable cost. A cost-based test gradually won favor in the courts, most explicitly through the Supreme Court's endorsement in *Brooke Group*, which insisted on "some measure of incremental cost" although without choosing a particular measure.[190] A cost-based test is also employed in the European Union, but the approach is less strict.[191]

A vigorous debate ensued over workable rules to control predatory pricing without generating a large number of false positives. Scherer (1976) criticized Areeda and Turner's analysis, Williamson (1977) and Baumol (1979) offered alternative tests, and further critical commentary was offered by Joskow and Klevorick (1979) and Ordover and Saloner (1989). More recently, there has been a further round of proposals and criticisms, including Bolton, Brodley, and Riordan (2000), Edlin (2002), and Elhauge (2003b). See also Areeda and Hovenkamp (2002, vol. 3).

To illuminate this controversy, it is useful to relate the question of the appropriate legal test for predatory pricing to two of our previous discussions. First, our survey of the economic theory in subsection 5.3.1, in explaining how predatory strategies could credibly deter entry or in some instances drive out rivals, made no reference to any cost-based test. The literature neither suggests that pricing below some particular concept of cost is a necessary condition nor that it is a sufficient one. (Indeed, many proposed alternatives are motivated by a belief that stringent cost-based tests are under-inclusive, that is, that they would exonerate much predatory behavior.) Also it should be kept in mind that the literature does not for the most part specifically indicate the effects of predatory pricing on consumer surplus or total welfare, which would be necessary to translate its results

[190] Brooke Group Ltd. v. Brown & Williamson Tobacco Corp., 509 U.S. 209, 223 (1993). Our discussion in the text implicitly refers to monopolization claims under Sherman Act Section 2; claims are also possible under the Robinson-Patman Act, but *Brooke Group* greatly reduced the differences in the Acts' requirements (moving the Robinson-Patman Act test closer to the Sherman Act monopolization test).

[191] In addition to declaring prices below average variable cost by a dominant firm that eliminates competitors as abusive, prices above average variable cost but below average total cost that eliminate competitors might also be reached. Tetra Pak International SA v. Commission, Case C-333/94P [1996] ECR I-5951, ¶41; ECS/Akzo II, Decision of the Commission, December 14, 1985, 1985 OJ L 374/1; Hilti, Decision of the Commission, December 22, 1987, 1988 OJ L 65/19. And EU authorities have also left open the possibility that targeted price reductions above total cost might be reached.

into a normative rule, but rather is focused on the circumstances in which predatory strategies are credible. In addition, the literature is not addressed to benign or beneficial price reductions in a variety of settings and thus does not indicate whether insisting that pricing be below a certain cost measure is a good way to avoid penalizing or chilling desirable behavior. Many believe that a cost-based test, perhaps one that uses average or marginal variable cost, would be a reasonable and administrable manner of identifying dangerous conduct while immunizing other conduct. It must be admitted, however, that this view reflects more a set of hunches than any precise combination of formal analysis and empirical evidence.

Second, different views on predation standards to a large extent track different perspectives on the proper antitrust standard toward exclusionary conduct generally (see subsection 5.2.2), many of which were developed with predatory pricing in mind. For example, tests that condemn price cuts if and only if they are below marginal or variable cost are often defended because they reward productive efficiency, on the ground that in certain simple settings only less efficient rivals or entrants would be kept out of the market. This argument is consistent with defining exclusionary practices as those that reduce productive efficiency. As noted, however, a price reduction above this cost standard might still drive out slightly less efficient competitors whose presence might raise both consumer and total welfare. This sort of case motivates many of the proposed alternative predation rules, such as determining whether a defendant has made a short-run profit sacrifice or would not have engaged in a practice but for its effect of excluding rivals.

Another argument favoring a narrow prohibition on predatory pricing advanced by courts and commentators alike is that price cuts are a move in the right direction. This view does reflect a concern with consumer and total welfare. Of course, it is also a static view. Predatory pricing is problematic precisely because of a concern that higher prices will follow, a move in the wrong direction. The important truth underlying this argument is that most price cutting in the economy is desirable, and thus it is a reminder that false positives and related chilling effects are particularly costly when contemplating punishment of low prices.

The overall balance between false positives and false negatives (and corresponding *ex ante* effects of each) depends in significant part on how one assesses the empirical evidence and the quality of the system of adjudication. If predatory pricing really is rare, as some Supreme Court pronouncements (based on partial and dated evidence) suggest, then the optimal test should reflect a disproportionate concern with false positives.[192]

[192] In *Brooke Group*, the Court merely recalled its remark in *Matsushita* on "the general implausibility of predatory pricing." 509 U.S. at 227. *Matsushita*, a 1986 decision, in turn cited analytical arguments by legal commentators (but none of the literature mentioned in subsection 5.3.1 on modern economic analysis of predation) and the empirical papers by McGee and Koller but not that of others, such as Zerbe and Cooper, who study more cases. Matsushita Electric Industrial Co., Ltd. v. Zenith Radio Corp., 475 U.S. 574, 588–90 (1986). As we have noted elsewhere, however, one should be generous in the assessment of court decisions when much of the relevant economic analysis and evidence may not have been presented to them by the litigants.

The brief survey in subsection 5.3.2 does not support this view, but it remains the case that beneficial price cuts will vastly outnumber predatory ones, so heavy attention to false positives is nevertheless sensible. Of course, this is an important motivation for the monopoly power screen, as we discussed in subsections 5.1 and 5.2.1.

Another important factor that we noted regarding exclusionary practices generally is the sophistication of decision-making. In the United States, predatory pricing is usually assessed by lay juries, who one might imagine would find conflicting expert testimony to be confusing and who might be sympathetic to a small firm driven out of business by a monopolist. Leading court opinions in the United States seem quite concerned about this matter and accordingly are reluctant to allow cases to proceed unless certain hurdles are overcome. Arguably, less caution is required if only the government (and not also disgruntled competitors) may initiate suit and if more expert agencies are responsible for applying tests for predation.

Additional complexity lurks beneath seemingly simple cost-based tests. Determining firms' costs in most settings is notoriously challenging, as we discussed in subsection 2.4.1.1 on the difficulty of measuring market power by observing the difference between price and marginal cost. One puzzle concerns the allocation of common costs. Even with a marginal cost test, common costs are often influenced at the margin, the only problem being that the extent of this phenomenon is difficult to measure. For example, how does one measure the cost of an additional employee? After including salary and fringes, one must think about secretaries and other support staff, rent and utilities on the space taken by the employee and others (keeping in mind that there are opportunity costs, as the space may well not otherwise remain vacant), and various central functions (time spent by the human resources department in searching for and hiring the employee, support from the computer department, and so forth). For nontrivial changes in output, which often are involved if a firm significantly lowers price in response to entry, these are just some of the complications in determining the cost of additional employees directly involved, say, in production, not to mention other costs. Note that the practice of ignoring myriad indirect costs tends systematically to underestimate marginal or variable cost and by a sufficiently great magnitude to make tests requiring prices to be below those costs highly permissive in many settings.

Other factors cut in the opposite direction. For example, there may be learning by doing, in which case a proper analysis of current marginal cost suggests that a lower figure should be employed because account must be taken of how present output reduces the cost of future production. See Arrow (1962) and Spence (1981). Another difficulty concerns complementarity: selling more automobiles may increase future revenue from spare parts, increasing flights between destinations A and B may increase traffic on other routes to and from points A and B on the same airline, or added sales may improve familiarity with a brand shared by other products. None of these factors is easy to measure.

Yet another set of complications involves capacity and other forms of investment. In *American Airlines*, American placed additional flights on routes served by the en-

trant.[193] The appellate court focused on whether fares covered marginal costs on those flights. However, many of those passengers would otherwise have taken other American flights already serving the route, so true incremental revenue was less than appeared to be the case. Moreover, one supposes there was a substantial opportunity cost of diverting planes to the route in that, wherever they had previously been deployed, they probably contributed net revenue sufficient to justify American carrying the additional capacity.

Another possibility is that American might have maintained excess capacity precisely so that it could be deployed in response to new entry (and to the extent capacity is observable by prospective entrants, it would tend to discourage entry). More broadly, firms can invest in capacity to deter entry by making it easier to reduce price quickly, and they might make investments that lower future marginal costs, which may never be recovered directly but would allow them to charge lower prices in response to subsequent entry, which boosts credibility and also helps to avoid running afoul of predation tests based on variable costs. See Spence (1977) and Dixit (1979, 1980).

These latest examples begin to blur the distinction between predatory pricing and exclusionary practices more broadly. Ordover and Willig (1981) bridged this gap by offering examples of what they described as predatory product innovation, such as investments designed to make a product more attractive than a rival's product (while setting a price gap below the incremental value) or to make the products of a dominant incumbent firm incompatible with those of rivals or prospective entrants. Likewise, one can think of high expenditures on R&D designed to come up with a patented product that has low marginal cost. In all of these cases, ignoring prior investments gives a misleading picture of firms' possibly predatory behavior. On the other hand, subjecting complex investment, research, and product design decisions to intensive antitrust scrutiny may be quite dangerous. How is one to distinguish the firm that makes substantial expenditures in a new product or service that it hopes will have a huge ultimate payoff (think of Amazon.com, eBay, or Google) from the firm making a predatory investment? Hindsight is often twenty-twenty, but risky legitimate investments often fail, so it will usually be difficult to distinguish the cases even after the fact.

Returning to the law on predatory pricing proper, U.S. law (but not that in the European Union) has added an additional requirement, that a party alleging predation prove that it is likely that the defendant will ultimately recoup its interim losses.[194] On one hand, this requirement is certainly logical, for if recoupment is implausible, there will be no (or less than complete) ultimate loss of consumer or total welfare, and more importantly the inability to recoup casts doubt on whether predation has in fact occurred.[195]

[193] United States v. AMR Corp., 335 F.3d 1109 (10th Cir. 2003). For an analysis, see Edlin and Farrell (2004).

[194] For U.S. law, the requirement is announced in Brooke Group Ltd. v. Brown & Williamson Tobacco Corp., 509 U.S. 209 (1993). For EU law, see Tetra Pak International SA v. Commission, Case C-333/94P [1996] ECR I-5951.

[195] The discussion in the text assumes that the pertinent cost test is one under which there are losses that need to be recouped.

On the other hand, the requirement seems redundant. First, the monopoly power screen exists to distinguish cases in which anticompetitive conduct is plausible from those in which it is not. A firm would be unlikely to recoup its losses in situations in which there are close substitutes for its products or in which entry, supply substitution, and the like would impose significant constraints on price increases. But it is precisely these factors that negate the existence of monopoly power.[196] Second, if one is reasonably confident that predation has in fact occurred, that very fact gives rise to a logical inference of recoupment. After all, the firm would not sustain losses out of charity, so its own analysis suggests that recoupment is likely (more precisely, that on an expected basis, it will occur). It seems implausible that courts' or agencies' assessments in the context of litigation would be more accurate than those of the firm with its own funds on the line.

The main explanation for this seeming puzzle is the concern for false positives combined with uncertainty about the other legal elements for a monopolization claim. Doubts about the proof of predation are certainly understandable in light of the above discussion, but this uncertainty was the justification for a strong monopoly power screen, and it is harder to understand how the recoupment requirement supplements rather than repeats this screen. The central problem is to improve the ability to distinguish true predation from legitimate price-cutting. Yet alternative explanations for below-cost pricing—namely, promotion of new products (which can involve periods of sustained losses, which the above Internet examples illustrate)—also presuppose recoupment. That is, a firm will only be willing to suffer substantial losses in promoting an innovative product if the quasi-rents from subsequent above-marginal-cost pricing (discounted for time and probability) are greater. If there are close substitutes or ready prospects of entry (imitation), the costly campaign would not be undertaken.

In reflecting on the recoupment requirement, it is notable that it was announced in the *Brooke Group* case, which involved an unfavorable setting in many respects. The case was brought by a competitor, which raises suspicions. The alleged recoupment was to be through oligopoly pricing, about which the Court expressed skepticism.[197] Finally, the finding of liability was made by a lay jury. In all, the Court's strong concern about false positives may have been warranted, but the logic of an independent recoupment requirement in addressing this concern remains unclear.

[196] The degree to which this statement is true depends on how the monopoly power requirement is interpreted. If, as we suggest in subsection 5.1.2 and throughout, it is interpreted with regard to the danger posed by the practice under consideration, then the recoupment requirement indeed seems fully redundant. Put another way, the recoupment requirement might be understood as a warning to assess monopoly power less in a vacuum and more in light of the challenged practice.

[197] This skepticism may not have been warranted, for oligopoly pricing is hardly rare—and had previously been documented in the industry in *Brooke Group*, tobacco—and, as we discussed in section 3, punishments through price cuts are an important means of sustaining collusion.

5.4. Exclusive dealing

We now turn to exclusive dealing, an important form of non-price conduct that a monopolist might use to fortify its dominant position. Exclusive dealing involves a supplier's conditioning its sales to customers on their agreeing not to purchase from its rivals.[198] In addition to being important in its own right, analysis of exclusive dealing illuminates the economics of a variety of other practices that we do not take up explicitly. We begin by considering possible anticompetitive effects, then briefly examine efficiencies, and finally discuss the legal test, illustrating the principles with a number of cases.

5.4.1. Anticompetitive effects

We focus here on exclusive dealing imposed by an upstream monopolist, M, on downstream customers. In the scenario most common in antitrust cases, M imposes exclusivity on wholesalers or retailers (rather than on final consumers). Clearly, any customer who has agreed to deal exclusively with M cannot purchase from M's rivals. In this mechanical sense, those rivals are excluded from selling to customers. But this would be formally true to some extent even without exclusivity because every unit bought from M is a unit that otherwise may have been purchased from M's rival. To consider anticompetitive effects, it is necessary to introduce strategic considerations. Our analysis of these draws heavily on Whinston (2006, Chapter 4), a major contributor to the recent game-theoretic literature in this area, who provides a masterful treatment of exclusive dealing and other exclusionary vertical contracts.

Anticompetitive exclusion most plausibly arises when M requires its dealers to purchase only from itself, these dealers constitute a large proportion of the market, and profitable entry or continued survival requires the rival to achieve a scale greater than is possible if sales must be limited to dealers not subject to exclusive-dealing contracts. A leading criticism of the possibility of anticompetitive exclusive dealing (and, as we shall discuss, of other allegedly anticompetitive contractual practices) comes from the Chicago School. In essence, the argument is that the dealers are harmed by anticompetitive exclusion because, if successful, the dealers will then be confronted by a monopoly;

[198] We generally will write in terms of explicit exclusive dealing but note in passing that a variety of seemingly distinct contractual arrangements, without explicit exclusivity, can have very similar economic effects. Consider, for example, a quantity discount in the form of a two-part tariff with a large fixed fee and a per-unit price that equals marginal cost. This pricing structure reduces the customer's incentive to purchase from alternative suppliers relative to the case in which the supplier charges a uniform price above marginal cost. Consider also discounts to buyers that purchase a large fraction of their needs from the incumbent supplier, such as was present in LePage's, Inc. v. 3M, 324 F.3d 141 (3d Cir. 2003). (In the limit, if the price is prohibitively high for a buyer who purchases less than 100% of its needs, the contract is economically equivalent to an exclusive dealing arrangement, but similar effects might be achieved far before this limit is reached.) A further variation on exclusive dealing arises when a buyer requires its suppliers not sell to its rivals. This was the fact pattern in Toys "R" Us, Inc. v. Federal Trade Commission, 221 F.3d 928 (7th Cir. 2000). Because the economic principles and analysis are similar, we only discuss the case in which it is the seller that imposes exclusivity on its customers.

accordingly, it will be against their interests to enter into arrangements resulting in anticompetitive exclusion. As a corollary, if practices like exclusive dealing are nevertheless observed, it must be because they generate efficiencies rather than produce anticompetitive effects.[199] These views (and other criticisms of claims of anticompetitive exclusion) were initially launched in the 1950s by Director and Levi (1956) and were followed by a wave of related commentary, the most elaborate being Bork (1978).[200]

Over the course of this subsection, we will consider many aspects of this argument. We begin with the factor that is probably most important in antitrust challenges to exclusive dealing: the presence of multiple (often very large numbers of) buyers, which leads to a free-rider problem in attempts to foil M's anticompetitive design.[201] This point is elaborated informally in Kaplow (1985, pp. 531–36) and elsewhere, and it has been developed formally in subsequent work by Rasmusen, Ramseyer, and Wiley (1991), Innes and Sexton (1994), and Segal and Whinston (2000b).

To make this idea explicit, suppose that a prospective entrant E must attract some critical mass of buyers to cover its fixed costs, and imagine that M attempts to enter into exclusive agreements with more than enough buyers so that E cannot profitably enter by serving those remaining.[202] None of these buyers would be pivotal, that is, none of them alone can induce E to enter by refraining from agreeing to exclusivity with M. Therefore, each would in fact agree to exclusivity in exchange for a very small additional payment. This argument supports an equilibrium in which all buyers agree to exclusivity, in exchange for an arbitrarily small transfer, and E is excluded from

[199] As noted by Farrell (2005), however, one could equally well conclude that exclusivity must generate some anticompetitive effects not captured in the simple model advanced by the Chicago School. Farrell (p. 468) characterizes the Chicago School argument on exclusive dealing as "a Rorschach test, and the inference often drawn from it is mere spin."

[200] For further references, discussion of objections in addition to the argument emphasized in the text, and critical commentary, see, for example, Kaplow (1985). To give some flavor of the Chicago School critique, consider Bork's (1978, pp. 306–07) remarks on Standard Fashion Co. v. Magrane-Houston Co., 258 U.S. 346 (1922): "Standard can extract in the prices it charges retailers all that the uniqueness of its line is worth. It cannot charge the retailer that full worth in money and then charge it again in exclusivity the retailer does not wish to grant. To suppose that it can is to commit the error of double counting. ... If Standard finds it worthwhile to purchase exclusivity from some retailers, the reason is not the barring of entry but some more sensible goal, such as obtaining the special selling effort of the outlet." Bork goes on to say (p. 309): "A seller who wants exclusivity must give the buyer something for it. If he gives a lower price, the reason must be that the seller expects the arrangement to create efficiencies that justify the lower price. If he were to give the lower price simply to harm his rivals, he would be engaging in deliberate predation by price cutting, and that, as we have seen in Chapter 7, would be foolish and self-defeating behavior on his part."

[201] Note that the Chicago School argument and the multiple buyers/free-rider response are applicable to a wide range of exclusionary practices, including predatory pricing (where the argument has also been raised, although less frequently).

[202] How many buyers E needs to serve depends on a number of factors, including M's and E's cost functions, the nature of demand for their products, and the mode of competitive interaction (in formal models, the extensive form of the game being studied). The basic argument in the text, however, depends only on this number being positive and not on how it is determined.

the market.[203] Furthermore, this result follows even if E is more efficient than M. The key element that makes this equilibrium possible is a lack of coordination among the buyers: individual buyers, or even groups of buyers too small to offer the entrant enough business to enter profitably, cannot gain by refusing to sign exclusive contracts with M.[204] But these contracts do harm all of the buyers and cause inefficiency. Enabling competition from E is a public good, and M can induce buyers to free ride, undermining that competition.

Subsequent work has explored the robustness of this exclusionary equilibrium. Whinston (2006) points out that this outcome might seem fragile since there arguably is another equilibrium in which none of the buyers agree to exclusivity. If M is constrained to make nondiscriminatory offers to the various buyers, and if the buyers can coordinate to the extent of selecting their Pareto-preferred equilibrium, none will agree to exclusivity in exchange for a *de minimis* payment. However, Segal and Whinston (2000b, 2003) show that exclusion is a robust outcome if M can make discriminatory offers to the various buyers. They also show that exclusion is easier to support if M makes sequential offers to buyers. As stated by Whinston (2006, p. 146): "More generally, the ability of the incumbent to approach buyers sequentially both reduces the cost of successful exclusion, and makes it more likely that the incumbent will find exclusion profitable. In fact, as the number of symmetric buyers grows large, so that each buyer becomes a very small part of aggregate demand, the incumbent is certain to be able to exclude for free."[205]

Consider further how these ideas generalize to settings that often arise in practice in which exclusive dealing targets M's existing rivals, not just potential entrants, and products are differentiated.[206] Suppose now that E is an existing rival that, despite the presence of scale economies and its smaller relative size, is able to survive due to M's high price and product differentiation, with E's product especially well suited to some customers. In this scenario, M may successfully enter into exclusive arrangements with most dealers, thereby limiting E's ability to expand. Each of M's customers would

[203] Note that, since the required payment per buyer is trivial, it is essentially costless to M to sign up more buyers than necessary, ensuring that no buyer will believe that there is any real possibility that it would be pivotal.

[204] It is conceivable that a large group of buyers would attempt to agree not to deal with M, or at least not exclusively. In addition to the free-rider problem in organizing and enforcing such an agreement, it should be noted that it may well be illegal under antitrust law. See subsection 3.5.2. Typically, antitrust law does not allow buying cartels and other otherwise illegal arrangements to be justified on grounds that they create countervailing power.

[205] Another counter-strategy would be for E to create or induce the entry of additional dealers or to bypass dealers and directly serve customers at the next level in the distribution chain. Often this will be infeasible or impose substantial costs on E; think of products sold primarily through department stores, drug stores, or full-line wholesalers, where E supplies only one or a few of those products. This issue arose in the *Dentsply* case, discussed in subsection 5.4.3.

[206] See also Salop and Scheffman (1983) and Krattenmaker and Salop (1986), who consider a variety of strategies designed to elevate rivals' costs, which may or may not induce exit, and the survey of this work in Ordover and Saloner (1989, pp. 565–70).

prefer to be free to deal with E but nevertheless may find it very costly to resist M's exclusivity policy since this would require forgoing all purchases from M. In many settings, a wholesaler or retailer would find it difficult not to stock M's products, which by assumption dominate the market. As a result, E may indeed conclude that making the necessary investments to expand its product line, manufacturing capacity, sales and distribution network, or advertising would not yield a sufficient return. M's exclusivity may render unprofitable a strategy under which E gradually increases its share at the retail locations where M has traditionally been dominant. Many individual retailers choose to purchase from M on an exclusive basis, but collectively the retailers, and final consumers, are harmed in the long run by M's exclusive dealing.[207]

Additional variations and qualifications should be mentioned. Most notably, one or a few large buyers may find it profitable to support entry. If there is a single buyer, the free-rider problem does not arise, and with only a few the problem is attenuated, especially if one is large enough to support an entrant—although note that its behavior may still convey a positive externality on others, so its incentives may not be sufficient. Fumagalli and Motta (2006) show, however, that if the buyers compete against each other, one or a few might grow large enough to support entry if they can obtain more favorable terms from E than their rivals did by signing exclusive contracts with M. Even if entry does not occur, the threat of large buyers to sponsor entry might induce M to offer them better deals, reducing M's market power. On the other hand, as shown by Simpson and Wickelgren (in press), one buyer may have little incentive to resist anticompetitive exclusion of an upstream entrant so long as its rivals are equally disadvantaged; if all suffer similarly, higher input costs will largely be passed on to the next level, and it is those customers who will suffer from M's continued monopoly. Yet another possibility, along the lines of Aghion and Bolton (1987), discussed further below, is that M might find it more profitable to sign contracts with stipulated damages that lead to entry by only the relatively more efficient potential entrants. Segal and Whinston (2000b) show that this outcome can arise in a model with multiple buyers.[208] Finally, it should be noted that even when an incumbent monopolist can profitably exclude a rival using

[207] As we will discuss below and illustrate with cases in subsection 5.4.3, formal exclusivity contracts are not necessary for this result. Similar effects may arise from pricing strategies (such as quantity or loyalty discounts) or threats (implicit or explicit) of reduced services or a complete cut-off if dealers also sell E's wares.

[208] Another interesting and complex strand of the literature examines situations in which there is direct competition between buyers to sign contracts that may be exclusive. Exclusive contracts can affect oligopolistic competition even if no firms are excluded from the market as a result. See Besanko and Perry (1993, 1994). Bernheim and Whinston (1998) study the effects of banning exclusive dealing in a number of distinct models, stating (p. 64): "We demonstrate that a ban may have surprisingly subtle and unintended effects." For further discussion of the complexity and potential ambiguity of these models, see Whinston (2006, pp. 152–78). This literature overlaps with a broader literature on vertical integration and vertical contracting. See, for example, Hart and Tirole (1990), O'Brien and Shaffer (1992), McAfee and Schwartz (1994), and the survey by Rey and Tirole (2007).

exclusive dealing, the welfare analysis is further complicated because some profitable entry is inefficient, as shown by Mankiw and Whinston (1986).[209]

Although the case of multiple buyers is central to most antitrust cases involving exclusive dealing, it is useful to consider briefly the Chicago School argument about buyer resistance to exclusivity in the single-buyer case, one that has received substantial attention in the contract theory literature. Begin with a simple case in which the incumbent monopolist M has a constant marginal cost above that of the potential entrant E, who also bears a fixed cost to enter. In the case of interest, E enters if no exclusive dealing contract binds the sole buyer B to M, and we suppose that the price after entry is determined by Bertrand competition and thus equals M's marginal cost.[210] Suppose further that M wishes, before E is on the scene, to bind B to buy from M even if E should enter, and also suppose that the proposed contract does not specify the price.[211] In that case, B expects to pay the monopoly price, which is higher than what B would pay in the absence of the exclusive dealing contract. B's loss in surplus from exclusivity is simply the sum of M's gain in producer surplus and the deadweight loss, and this total is obviously greater than M's gain, by the amount of the deadweight loss. The most that M would pay for exclusivity is less than the least B would accept, so the Chicago School claim is valid under these assumptions. As pointed out by Whinston (2006, p. 139), this result does not rely on any specific bargaining model but rather reflects what he calls the bilateral bargaining principle: "if two parties (i) contract in isolation, (ii) have complete information about each others' payoffs, and (iii) lump-sum transfers are possible, then they will reach an agreement that maximizes their joint payoff." Readers will also recognize this claim as a version of the Coase theorem. Here, the joint payoff of M and B is reduced—by the amount of the deadweight loss—if they sign an exclusive contract. But this is not always the case, even with only a single buyer.[212]

An important set of extensions allows for other types of contract between M and E. Aghion and Bolton (1987) examine an exclusive contract that stipulates the price P but allows B to breach and purchase instead from E (at whatever price E might offer) upon payment of damages G to M.[213] Aghion and Bolton show how M and B can select G to extract rents from E. Since B must pay G to M if it wishes to buy from E, E needs

[209] The possible tradeoffs raise the question of the objectives of antitrust law and general principles governing exclusionary practices, discussed in subsection 5.2.2.

[210] This will be the case so long as E's profits are increasing in price at least up to that level. If not, price will be lower, but the conclusion to follow in the text will still hold.

[211] This incompleteness may arise for the usual reasons, such as future uncertainty and problems of verifiability. If the contract did specify price—as allowed by Aghion and Bolton (1987), discussed below—then we might suppose that the price would equal M's marginal cost with all rents extracted through an *ex ante* fixed charge.

[212] In addition to the variations considered in the text to follow, Farrell (2005) points out that the conclusion does not even generalize to alternative models of post-entry duopoly between M and E. With Cournot duopoly, exclusion can be profitable and inefficient.

[213] Many of the ideas elaborated in this literature first appeared in Diamond and Maskin (1979). See also Chung (1992).

to offer B a better deal than if there were no exclusionary contract. It is in M and B's joint interest to raise G just to the point that fully extracts E's profits, sharing this gain between themselves. Note that this outcome is efficient and does not exclude E, which is by assumption more efficient than M.

However, Aghion and Bolton (1987) show that this relatively benign rent-shifting result does not generalize to situations in which there is uncertainty about E's costs. In that case, perfect extraction from E is not possible since M and B do not know *ex ante* how much may be extracted. Setting G involves a familiar sort of tradeoff: raising G increases extraction from potential entrants with sufficiently low costs that they still enter but loses entry and thus forgoes extraction from those with higher costs. Thus, the privately optimal level of G will partially extract surplus from relatively efficient entrants and will exclude entrants that are not so efficient (but still more efficient than M). Through the contract, M and B act somewhat like a monopsonist purchasing from a distribution of potential entrants with different costs.[214] Even in this case, note that the buyer is not harmed by the use of stipulated damages on an *ex ante* basis. As emphasized by Farrell (2005), harm to the buyer from exclusivity is not possible in this simple setting because the buyer can always "just say no" by not agreeing to grant exclusivity to M. This result contrasts with that in the previously discussed models with multiple buyers.

The Chicago School argument about buyers' resistance to exclusive dealing is an application of a broader critique applied to a wide range of exclusionary practices that is referred to as the "one-monopoly-profit theorem" and under related rubrics. The essence of the wider attack draws on the idea that monopolists, no matter how powerful, cannot get something for nothing. That is, the argument holds there is some level of profit or rent that inheres in a given monopoly position, and monopolists cannot extract other concessions (potentially anticompetitive ones or otherwise) without giving up something in return.

This general point and many of its applications served as a useful corrective to superficial arguments that used to be prominent in both court opinions and commentary. A monopolist cannot generate monopoly returns on related products or in other markets—in addition to the monopoly profits it already is earning through charging the monopoly price—simply by threatening to withhold the product on which it enjoys the monopoly. Monopoly prices do not rise without limit, as we elaborated in section 2. Rather, there is an optimal price characterized by the property that a slightly higher price loses as much profit due to lost sales as is gained by the heightened margin on retained sales. Buyers who would be lost are, by definition, at the margin. Therefore,

[214] Spier and Whinston (1995) point out that this line of argument requires that M and B be able to commit themselves to the terms of their initial contract; perfect and costless renegotiation between them after E enters would undermine their ability to extract rents from E. However, Spier and Whinston show that this sort of contract would nevertheless benefit M and B through its influence on incentives to undertake investments before E arrives on the scene.

demanding other concessions, if they impose any positive cost, will lose these buyers as well unless price is reduced or other countervailing inducements are offered.

Despite the important element of truth in this proposition, it is now well known that there are substantial qualifications. Indeed, the so-called one-monopoly-profit theorem literally holds only in very special cases not often thought to be realistic, such as when a monopolist ties its product to another that is used in fixed proportions and is available competitively.[215] Within a static framework, the extent of monopoly profits may depend on practices that facilitate price discrimination or that limit substitution, such as by tying the sales of a monopolized product to sales of partial substitutes at an appropriate margin when otherwise the substitutes can be obtained at competitive prices. Many such practices have indeterminate effects on consumer and total welfare.

In a dynamic framework—especially when one also introduces externalities (which may be common with regard to effects on competition), asymmetric information, and other strategic dimensions—there are many more possibilities, including ones with anticompetitive effects. Moreover, most claims about exclusionary practices are expressly of a dynamic character. (Consider, for example, many proposed tests for exclusionary practices that we discussed in subsection 5.2.2, such as whether a short-run sacrifice is involved or whether a practice would be profitable but for its ultimate effect of excluding rivals.) The point that analysis can change qualitatively in such dynamic settings is well illustrated by the foregoing discussion of exclusive dealing with multiple buyers and also by our examination of credible predation in subsection 5.3.1. Indeed, substantial bodies of literature in industrial organization over the past few decades have been devoted to settings in which the one-monopoly-profit theorem does not hold. In any event, we have seen that it does not negate the anticompetitive potential of exclusive dealing.

5.4.2. *Efficiencies*

That exclusive arrangements can promote efficiency may be inferred from their use in situations where meaningful market power is clearly absent, such as in many employment contracts. Employees or members of a partnership may be forbidden from working elsewhere in order to avoid diversion of effort and to limit their ability to take personal advantage of opportunities developed by the enterprise. Marvel (1982) develops these notions in a context more pertinent to exclusive dealing arrangements challenged under the antitrust laws. For example, a manufacturer that makes investments to attract customers to a retailer may be concerned that the retailer would free ride by diverting these customers to competitors' products if not precluded from doing so by some form of exclusivity. As another example, Masten and Snyder (1993) revisit the famous *United*

[215] Many of these limitations have long been well known. Some are discussed and further references are offered in Kaplow (1985).

Shoe case, arguing that the contractual provisions inducing shoe manufacturers to exclusively use United Shoe's machines protected the investments made by United Shoe in training shoe manufacturers to organize their production processes more efficiently.[216]

Subsequent literature on contract theory has refined our understanding of the underlying mechanism. In Segal and Whinston's (2000a) model, a buyer and a seller, subsequent to entering into a contract, independently make noncontractible investments, after which they bargain over the terms of trade. If the initial contract is not exclusive, the buyer has the option of turning to an alternative supplier of the product. In a model in which the buyer only needs one unit and the seller can make investments that are specific to the relationship with the buyer—that is, they provide no value to either buyer or seller if the two do not end up dealing with each other—Segal and Whinston show that exclusivity has no effect on the seller's investment incentives. Exclusivity reduces the buyer's threat point and thus raises the seller's *ex post* payoff, but in a way that is unaffected by the seller's investment.

This result suggests that pro-competitive justifications for exclusivity based on free riding and investment incentives require investments that are not entirely relationship specific. Segal and Whinston (2000a) show that exclusivity promotes seller investments that are also valuable to the buyer when dealing with third parties, but discourages seller investments that raise value to the buyer from remaining with the seller relative to switching to third parties. Opposite results apply for buyer investments. The previously noted examples of seemingly efficient exclusive dealing fit this pattern in that they involve seller investments that the buyer can exploit in dealing with alternative suppliers. Although Segal and Whinston's analysis solidifies our understanding, they also point out that the full welfare analysis is more complex because increased investment need not mean increased welfare.

5.4.3. Legal test

It has long been believed that exclusive dealing contracts and related arrangements have the potential both to be anticompetitive and to promote efficiency. Accordingly, U.S. law has applied a balancing test along the lines of that under the rule of reason, which we discussed in subsection 3.5.2.[217] As a further legal note, contractual exclusivity may

[216] United States v. United Shoe Machinery Corp., 110 F. Supp. 295 (D. Mass. 1953), affirmed per curiam, 347 U.S. 521 (1954).

[217] Standard Oil Co. of California v. United States, 337 U.S. 293 (1949), discussed below, recognized that exclusive arrangements were potentially efficient even though it judged Standard Oil's arrangements harshly. Subsequent cases, including some of the others discussed below, have required a greater demonstration of anticompetitive effects and also have more clearly acknowledged that efficiencies count, even if they were unconvinced by those presented by the defendants. See Hovenkamp (2005, vol. 11, ch. 18D). In the European Union, Article 82 on abuse of a dominant position encompasses exclusivity and related agreements (such as loyalty or fidelity rebates and various quantity discounts) imposed by a dominant supplier. Also covered are "English clauses" (which in the United States are usually called price-matching clauses) that allow purchasers to buy from rivals offering lower prices, but only if they first inform the contract supplier and that supplier is unwilling to make the sale at an equivalent price.

be challenged under a variety of U.S. antitrust provisions. In addition to monopolization under Sherman Act Section 2, it may be reached under Section 1 since an agreement is involved, under Clayton Act Section 3's prohibition of contract provisions that anti-competitively restrict dealings with competitors, and under FTC Act Section 5's broad proscription against unfair methods of competition. The same is true of a variety of other forms of vertical restraints, including tying. As we noted in subsection 4.4.1 in connection with horizontal mergers, however, there has been a growing convergence in treatment regardless of the particular statutory provision invoked.

To state that the law applies a balancing test or, more particularly, applies a rule of reason, does not convey a very clear sense of how it actually operates, so it is useful to consider some cases. The *Standard Stations* decision in 1949 reflects the much stricter attitude of the Supreme Court at that time period.[218] Standard Oil had exclusive supply contracts with 16% of the retail outlets in the geographic market, most of which were terminable at six-month intervals upon giving thirty days notice. Although this arrangement does not seem to constitute an insuperable barrier to an entrant or a rival seeking to expand (despite the fact that other suppliers also used similar arrangements), the Court affirmed a determination that it was anticompetitive. A stronger case was presented in *Lorain Journal*, where an incumbent newspaper with a local news and advertising monopoly (in 1948) was found guilty of attempted monopolization for refusing to carry ads of those who also advertised on the newly entered radio station.[219]

More recently, variations on the exclusivity theme have appeared in many phases of the litigation involving Microsoft. In the mid-1990s, the government challenged and Microsoft ultimately agreed to cease the use of per-processor licensing fees for its operating system. Computer manufacturers who had wished to load Microsoft's operating system on some of their computers were charged for loading it on all of the computers they shipped, as a condition for dealing with Microsoft. Although not literally barred from dealing with competitors, computer manufacturers were discouraged from doing so since they had to pay for Microsoft's operating system even on computers shipped with an alternative operating system (or with none). Subsequent litigation successfully challenged other features of Microsoft's contracting and operating system design that exhibited some exclusivity.[220] In another recent case, *Dentsply*, the leading supplier of artificial teeth with a 75–80% market share was found to have violated §2 for imposing exclusivity on its dealers.[221]

[218] Standard Oil Co. of California v. United States, 337 U.S. 293 (1949).

[219] Lorain Journal Co. v. United States, 342 U.S. 143 (1951). This case is often discussed along with United States v. Griffith, 334 U.S. 100 (1948), in which the Supreme Court found monopolization where a chain of movie theatres with monopolies in many towns insisted on certain exclusive rights in all towns. Questions on both cases that outline the analysis in subsection 5.4.1 for the situation involving multiple buyers (though, for *Griffith*, it was multiple suppliers) appear in Areeda, Kaplow, and Edlin (2004, ch. 3).

[220] United States v. Microsoft Corp., 253 F.3d 34 (D.C. Cir. 2001).

[221] United States v. Dentsply International, Inc., 399 F.3d 181 (3d Cir. 2005). An interesting feature of this case is that Dentsply did not formally have exclusive contracts with its dealers, but it did have supply arrange-

A number of features of these cases are notable. First, all involved multiple buyers, although two of the dealers in *Dentsply* did have substantial market shares. Also, except for *Standard Stations*, the defendants seemed to possess monopoly power, and successful entry and ultimate expansion would seem to have required significant scale.[222] These features are consistent with the analysis of anticompetitive effects presented in subsection 5.4.1. Regarding efficiencies, none seemed apparent in *Lorain Journal*, and efficiency justifications offered by Microsoft and Dentsply were found to be unconvincing.

Like other allegedly exclusionary practices, exclusive dealing presents potentially difficult problems of balancing, raising both factual questions and issues about the precise content of the legal test, the subject of subsection 5.2.2.[223] A further challenge is raised by the possibility (as in horizontal merger cases) that both anticompetitive effects and efficiencies may be present simultaneously, given that the logics underlying the two considerations are essentially independent. Accordingly, it is also possible that a highly anticompetitive exclusive arrangement would involve no short-term profit sacrifice by the monopolist. As explained in subsection 5.4.1, with large numbers of buyers, exclusivity that has no efficiency consequences might be secured at a trivial cost; hence, even the slightest efficiency benefit would produce immediate (even if modest) gains, along with more substantial future profits due to anticompetitive effects. Note that in such cases any issue of recoupment would likewise be moot.

One way that the law addresses the problem of distinguishing legitimate and harmful exclusive dealing is through the monopoly power screen, which readily filters out a vast proportion of exclusive arrangements that may be efficient. Think of routine employment contracts, exclusivity provisions in partnerships, and most of the countless products sold (typically but not always without exclusivity provisions, in fact) in department stores and drug stores and distributed by wholesalers of all sorts. Relatedly, in subsection 5.1.2 we discussed how, in addition to the traditional monopoly power requirement, one can also assess whether a challenged practice has any prospect of significantly damaging competition by stipulating that it has the alleged anticompetitive

ments that were terminable by it at will, combined with a formal policy of terminating dealers who carried competing products (subject to some grandfathered exceptions). The government convinced the court, based in large part on a series of actual events, including threatened terminations followed by dealers' discontinuance of competitors' products, that dealers did not believe that they could be successful without Dentsply products. This example illustrates that exclusive dealing policies can have anticompetitive effects even without the use of formal exclusive-dealing contracts, much less long-term exclusive-dealing contracts.

[222] In *Dentsply*, there had long been a number of small suppliers, but the court was convinced that, without access to most dealers (some had access to certain dealers and many attempted to sell directly to the next level in the distribution chain, dental laboratories), the rivals could not realistically expand.

[223] Regarding the facts, Whinston (2006, pp. 189–97) indicates that there is "remarkably limited" empirical evidence on the motives and effects of exclusive contracting. Of course, it still may be possible to make determinations in specific cases. The court in *Dentsply* summarily dismissed the efficiency claim as "pretextual," seeing it to be clearly contradicted by the evidence. In *Microsoft*, evaluation of Microsoft's efficiency claims was more difficult regarding at least some of its challenged practices, particularly those in which the *de facto* exclusivity was not a feature of contracts but of the product (operating system) itself.

effect and then determining how large that effect could be. In addressing efficiencies, courts undertake an inquiry that is analogous to the requirement in horizontal merger cases that efficiencies be merger specific (see subsection 4.4.3). Thus, for purported benefits of exclusivity and other restrictive contractual features, courts typically ask whether there exists a less restrictive alternative. For example, if the manufacturer provides some service or training, it might impose separate charges rather than employ exclusivity to prevent free riding.[224]

6. Conclusion

Having surveyed several key economic underpinnings of antitrust policy and applied the lessons to the core features of existing regimes, we have seen that economics has had a tremendous influence on the law, but also that there is still much unfinished business for economists and lawyers alike. For each of our four main topics, we have noted that modern antitrust law in the United States—and to a substantial degree in the European Union—is aimed in large part at economic objectives and heavily employs economic tools in achieving them. At the same time, in every field there appear to be notable divergences, and ones that cannot fully be explained by administrative convenience or limitations on institutional competence. Some discrepancies may be the product of conscious choice; others no doubt reflect the inevitable lag in the dissemination of economic principles.

At least as important for economists, our primary audience, are the many ways that existing theoretical work and empirical methods, valuable as they are, do not yet adequately address many of the questions that those who formulate and apply competition policy need to answer. For example, we noted that antitrust law on collusion seems quite interested in forms of communication, whereas this matter plays a relatively minor role in economic models and empirical work. In many instances, the problem may be that the legal regime does not ask the right questions. Even in such cases, however, answers would be helpful, if for no other reason than to see in what sense and to what extent antitrust decisions have been led astray. In other cases, one cannot expect the law to answer economic questions sensibly, in real time, when leading economic research has not yet done so. One point of particular interest in many areas of competition policy is that the law is very concerned with minimizing error, especially false positives and associated chilling effects. But it is difficult to calibrate legal tests without better-informed priors. Forming such priors requires, in turn, an empirical and theoretical understanding of many legitimate practices as well as of the anticompetitive ones that naturally are the focus of legal disputes. This knowledge is important because, in many actual cases that must be adjudicated, these practices are not easily distinguished from each other.

[224] This approach is no panacea, however. A frequent problem with such alternatives—paralleling a problem with conduct remedies in monopolization cases—is that the terms, including the price, may need to be monitored. However, if exclusivity were forbidden, the monopolist that has effective services to offer would not have an incentive to charge a prohibitive price.

Acknowledgements

We are grateful to Jonathan Baker for extensive and very valuable comments, Stephanie Gabor, Jeffrey Harris, Christopher Lanese, Bradley Love, Stephen Mohr, Andrew Oldham, Peter Peremiczki, Michael Sabin, and Kevin Terrazas for research assistance, and the John M. Olin Center for Law, Economics, and Business at Harvard University for financial support.

References

Abreu, D. (1986). "Extremal equilibria of oligopolistic supergames". Journal of Economic Theory 39, 191–225.

Abreu, D., Pearce, D., Stacchetti, E. (1986). "Optimal cartel equilibria with imperfect monitoring". Journal of Economic Theory 39, 251–269.

Abreu, D., Pearce, D., Stacchetti, E. (1990). "Toward a theory of discounted repeated games with imperfect monitoring". Econometrica 58, 1041–1063.

Aghion, P., Bolton, P. (1987). "Contracts as a barrier to entry". American Economic Review 77, 388–401.

American Bar Association, Section of Antitrust Law (2005). Market Power Handbook. American Bar Association, Chicago.

Andrade, G., Mitchell, M., Stafford, E. (2001). "New evidence and perspectives on mergers". Journal of Economic Perspectives 15 (2), 103–120.

Areeda, P., Hovenkamp, H. (2002). Antitrust Law: An Analysis of Antitrust Principles and Their Application, 2nd edn, vols. 3, 3A, 6. Aspen Law & Business, New York.

Areeda, P., Turner, D. (1975). "Predatory pricing and related practices under Section 2 of the Sherman Act". Harvard Law Review 88, 697–733.

Areeda, P., Hovenkamp, H., Solow, J. (2006). Antitrust Law: An Analysis of Antitrust Principles and Their Application, 2nd edn, vols. 4 and 4A. Aspen Law & Business, New York.

Areeda, P., Kaplow, L., Edlin, A. (2004). Antitrust Analysis, 6th edn. Aspen Publishers, New York.

Armstrong, M., Porter, R. (Eds.) (2007). Handbook of Industrial Organization, vol. 3. Elsevier Science Publishers, New York, in press.

Arrow, K. (1962). "The economic implications of learning by doing". Review of Economic Studies 29, 155–173.

Athey, S., Bagwell, K. (2001). "Optimal collusion with private information". RAND Journal of Economics 32, 428–465.

Athey, S., Bagwell, K. (2006). "Collusion with persistent cost shocks", draft.

Athey, S., Bagwell, K., Sanchirico, C. (2004). "Collusion and price rigidity". Review of Economic Studies 71, 317–349.

Bagwell, K., Staiger, R. (1997). "Collusion over the business cycle". RAND Journal of Economics 28, 82–106.

Bajari, P., Summers, G. (2002). "Detecting collusion in procurement auctions". Antitrust Law Journal 70, 143–170.

Baker, J. (1994). "Predatory pricing after *Brooke Group*: An economic perspective". Antitrust Law Journal 62, 585–603.

Baker, J. (1999). "Econometric analysis in *FTC v. Staples*". Journal of Public Policy & Marketing 18, 11–21.

Baker, J. (2002). "Mavericks, mergers and exclusion: Proving coordinated competitive effects under the antitrust laws". New York University Law Review 77, 135–203.

Baker, J. (2003a). "The case for antitrust enforcement". Journal of Economic Perspectives 17 (2), 27–50.

Baker, J. (2003b). "Responding to developments in economics and the courts: Entry in the merger guidelines". Antitrust Law Journal 71, 189–206.

Baker, J., Bresnahan, T. (1985). "The gains from merger or collusion in product-differentiated industries". Journal of Industrial Economics 33, 427–444.

Baker, J., Bresnahan, T. (1988). "Estimating the residual demand curve facing a single firm". International Journal of Industrial Organization 6, 283–300.

Baker, J., Rubinfeld, D. (1999). "Empirical methods in antitrust litigation: Review and critique". American Law and Economics Review 1, 386–435.

Baker, J., Shapiro, C. (in press). "Reinvigorating horizontal merger enforcement". In: Pitofsky, R. (Ed.), Conservative Economic Influence on U.S. Antitrust Policy. Oxford University Press.

Bargeron, L., Schlingemann, F., Stulz, R., Zutter, C. (2007). "Why do private acquirers pay so little compared to public acquirers?" Dice Center for Research in Financial Economics WP 2007–8.

Barton, D., Sherman, R. (1984). "The price and profit effects of horizontal merger: A case study". Journal of Industrial Economics 33, 165–177.

Baumol, W. (1979). "Quasi-permanence of price reductions: A policy for prevention of predatory pricing". Yale Law Journal 89, 1–26.

Bellamy, C., Child, G. (2001). European Community Law of Competition, 5th edn. Sweet & Maxwell, London.

Benoit, J.-P. (1984). "Financially constrained entry in a game with incomplete information". RAND Journal of Economics 15, 490–499.

Benoit, J.-P., Krishna, V. (1987). "Dynamic duopoly: Prices and quantities". Review of Economic Studies 54, 23–35.

Bernheim, B.D., Ray, D. (1989). "Collective dynamic consistency in repeated games". Games and Economic Behavior 1, 295–326.

Bernheim, B.D., Whinston, M. (1990). "Multimarket contact and collusive behavior". RAND Journal of Economics 21, 1–26.

Bernheim, B.D., Whinston, M. (1998). "Exclusive dealing". Journal of Political Economy 106, 64–103.

Berry, S. (1994). "Estimating discrete-choice models of product differentiation". RAND Journal of Economics 25, 242–262.

Berry, S., Reiss, P. (2007). "Empirical models of entry and market structure". In: Armstrong, M., Porter, R. (Eds.), Handbook of Industrial Organization, vol. 3. Elsevier Science Publishers, New York, in press.

Berry, S., Levinsohn, J., Pakes, A. (1995). "Automobile prices in market equilibrium". Econometrica 63, 841–890.

Besanko, D., Perry, M. (1993). "Exclusive dealing in a spatial model of retail differentiation". International Journal of Industrial Organization 12, 297–329.

Besanko, D., Perry, M. (1994). "Equilibrium incentives for exclusive dealing in a differentiated products oligopoly". RAND Journal of Economics 24, 646–667.

Bolton, P., Scharfstein, D. (1990). "A theory of predation based on agency problems in financial contracting". American Economic Review 80, 93–106.

Bolton, P., Brodley, J., Riordan, M. (2000). "Predatory pricing: Strategic theory and legal policy". Georgetown Law Journal 88, 2239–2330.

Borenstein, S. (1990). "Airline mergers, airport dominance, and market power". American Economic Review Papers and Proceedings 80 (2), 400–404.

Borenstein, S. (2004). "Rapid price communication and coordination: The airline tariff publishing case (1994)". In: Kwoka, J., White, L. (Eds.), The Antitrust Revolution: Economics, Competition, and Policy, 4th edn. Oxford University Press, New York, pp. 233–251.

Borenstein, S., Shepard, A. (1996). "Dynamic pricing in retail gasoline markets". RAND Journal of Economics 27, 429–451.

Bork, R. (1978). The Antitrust Paradox: A Policy at War with Itself. Basic Books, New York.

Bresnahan, T. (1989). "Empirical studies of industries with market power". In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. 2. North-Holland, Amsterdam, pp. 1011–1057.

Bresnahan, T., Salop, S. (1986). "Quantifying the competitive effects of production joint ventures". International Journal of Industrial Organization 4, 155–175.

Brock, W., Scheinkman, J. (1985). "Price setting supergames with capacity constraints". Review of Economic Studies 52, 371–382.
Bulow, J., Pfleiderer, P. (1983). "A note on the effect of cost changes on prices". Journal of Political Economy 91, 182–185.
Bulow, J., Shapiro, C. (2004). "The BP Amoco–ARCO merger: Alaskan crude oil (2000)". In: Kwoka, J., White, L. (Eds.), The Antitrust Revolution: Economics, Competition, and Policy, 4th edn. Oxford University Press, New York, pp. 128–149.
Burns, M. (1986). "Predatory pricing and the acquisition cost of competitors". Journal of Political Economy 94, 266–296.
Capps, C., Dranove, D., Greenstein, S., Satterthwaite, M. (2002). "Antitrust policy and hospital mergers: Recommendations for a new approach". Antitrust Bulletin 27, 677–714.
Carlton, D. (2004). "Why barriers to entry are barriers to understanding". American Economic Review Papers and Proceedings 94 (2), 466–470.
Carlton, D., Perloff, J. (2005). Modern Industrial Organization, 4th edn. Pearson-Addison Wesley, Boston.
Chamberlin, E. (1933). The Theory of Monopolistic Competition. Harvard University Press, Cambridge, Mass.
Chung, T.-Y. (1992). "On the social optimality of liquidated damage clauses: An economic analysis". Journal of Law, Economics, & Organization 8, 280–305.
Coase, R. (1937). "The nature of the firm". Economica 4, 386–405.
Coate, M., Ulrick, S. (2005). "Transparency at the Federal Trade Commission: The Horizontal Merger Review Process 1996–2003", Bureau of Economics, Federal Trade Commission. (www.ftc.gov/os/2005/02/0502economicissues.pdf.)
Compte, O. (1998). "Communication in repeated games with private monitoring". Econometrica 66, 597–626.
Compte, O., Jenny, F., Rey, P. (2002). "Capacity constraints, mergers and collusion". European Economic Review 46, 1–29.
Connor, J. (2004). "Global cartels redux: The amino acid lysine antitrust litigation (1996)". In: Kwoka, J., White, L. (Eds.), The Antitrust Revolution: Economics, Competition, and Policy, 4th edn. Oxford University Press, New York, pp. 252–276.
Connor, J. (2007). "Price-fixing overcharges: Legal and economic evidence". In: Kirkwood, J. (Ed.), Research in Law and Economics, vol. 22. JAI/Elsevier, Amsterdam, pp. 59–153.
Cooper, R., DeJong, D., Forsythe, R., Ross, T. (1989). "Communication in the battle of the sexes game: Some experimental results". RAND Journal of Economics 20, 568–587.
Cooper, T. (1986). "Most-favored-customer pricing and tacit collusion". RAND Journal of Economics 17, 377–388.
Cournot, A.A. (1838). Researches into the Mathematical Principles of the Theory of Wealth, English edn. Kelley, New York.
Cramton, P., Schwartz, J. (2000). "Collusive bidding: Lessons from the FCC spectrum auctions". Journal of Regulatory Economics 17, 229–252.
Crandall, R., Winston, C. (2003). "Does antitrust policy improve consumer welfare? Assessing the evidence". Journal of Economic Perspectives 17 (4), 3–26.
Crawford, V. (1998). "A survey of experiments on communication via cheap talk". Journal of Economic Theory 78, 286–298.
Dabbah, M. (2004). EC and UK Competition Law. Cambridge University Press, Cambridge.
Dalkir, S., Warren-Boulton, F. (2004). "Prices, market definition, and the effects of merger: Staples–Office Depot (1997)". In: Kwoka, J., White, L. (Eds.), The Antitrust Revolution: Economics, Competition, and Policy, 4th edn. Oxford University Press, New York, pp. 52–72.
Davidson, C., Deneckere, R. (1984). "Horizontal mergers and collusive behavior". International Journal of Industrial Organization 2, 117–132.
Davidson, C., Deneckere, R. (1990). "Excess capacity and collusion". International Economic Review 31, 521–541.
Demsetz, H. (1973). "Industry structure, market rivalry, and public policy". Journal of Law and Economics 16, 1–10.

Deneckere, R., Davidson, C. (1985). "Incentives to form coalitions with Bertrand competition". RAND Journal of Economics 16, 473–486.
Department of Justice, Antitrust Division (2006). Workload Statistics, FY 1996–2005. (http://www.usdoj.gov/atr/public/workstats.htm.)
Diamond, P., Maskin, E. (1979). "An equilibrium analysis of search and breach of contract, I: Steady states". Bell Journal of Economics 10, 282–316.
Director, A., Levi, E. (1956). "Law and the future: Trade regulation". Northwestern University Law Review 51, 281–296.
Dixit, A. (1979). "A model of duopoly suggesting a theory of entry barriers". Bell Journal of Economics 10, 20–32.
Dixit, A. (1980). "The role of investment in entry-deterrence". Economic Journal 90, 95–106.
Edlin, A. (1997). "Do guaranteed-low-price policies guarantee high prices, and can antitrust rise to the challenge?" Harvard Law Review 111, 528–575.
Edlin, A. (2002). "Stopping above-cost predatory pricing". Yale Law Journal 111, 941–991.
Edlin, A., Emch, E. (1999). "The welfare losses from price matching policies". Journal of Industrial Economics 47, 145–167.
Edlin, A., Farrell, J. (2004). "The American Airlines case: A chance to clarify predation policy (2001)". In: Kwoka, J., White, L. (Eds.), The Antitrust Revolution: Economics, Competition, and Policy, 4th edn. Oxford University Press, New York, pp. 502–527.
Elhauge, E. (2003a). "Defining better monopolization standards". Stanford Law Review 56, 253–344.
Elhauge, E. (2003b). "Why above-cost price cuts to drive out entrants are not predatory and the implications for defining market power". Yale Law Journal 112, 681–828.
Ellison, G. (1994). "Theories of cartel stability and the joint executive committee". RAND Journal of Economics 25, 37–57.
Elzinga, K., Hogarty, T. (1973). "The problem of geographical market delineation in antimerger suits". Antitrust Bulletin 18, 45–81.
Elzinga, K., Mills, D. (2004). "The brand name prescription drugs antitrust litigation (1999)". In: Kwoka, J., White, L. (Eds.), The Antitrust Revolution: Economics, Competition, and Policy, 4th edn. Oxford University Press, New York, pp. 301–320.
Epstein, R., Rubinfeld, D. (2001). "Merger simulation: A simplified approach with new applications". Antitrust Law Journal 69, 883–919.
Epstein, R., Rubinfeld, D. (2004). "Merger simulation with brand-level margin data: Extending PCAIDS with nests". Advances in Economic Analysis and Policy 4, 1–26.
European Union (2001). "Guidelines on the applicability of Article 81 of the EC Treaty to horizontal cooperative agreements". European Commission, Official Journal of the European Communities, 6.1.2001, C 3/2.
European Union (2004a). "Guidelines on the assessment of horizontal mergers". European Commission, Regulation 139/2004, 1 May.
European Union (2004b). "Guidelines on the assessment of horizontal mergers under the Council regulation on the control of concentrations between undertakings". Official Journal of the European Union, February, 2004/C 31/03.
Evans, W., Kessides, I. (1994). "Living by the 'Golden Rule': Multimarket contact in the U.S. airline industry". Quarterly Journal of Economics 109, 341–366.
Farrell, J. (2000). "Renegotiation in repeated oligopoly interaction". In: Myles, G., Hammond, P. (Eds.), Incentives, Organisation, and Public Economics: Papers in Honour of Sir James Mirrlees. Oxford University Press, New York, pp. 303–322.
Farrell, J. (2005). "Deconstructing Chicago on exclusive dealing". Antitrust Bulletin 50, 465–480.
Farrell, J., Katz, M. (2006). "The economics of welfare standards in antitrust". Competition Policy International 2 (2), 3–28.
Farrell, J., Maskin, E. (1989). "Renegotiation in repeated games". Games and Economic Behavior 1, 327–360.
Farrell, J., Rabin, M. (1996). "Cheap talk". Journal of Economic Perspectives 10 (3), 103–118.

Farrell, J., Shapiro, C. (1990a). "Asset ownership and market structure in oligopoly". RAND Journal of Economics 21, 275–292.
Farrell, J., Shapiro, C. (1990b). "Horizontal mergers: An equilibrium analysis". American Economic Review 80, 107–126.
Farrell, J., Shapiro, C. (2001). "Scale economies and synergies in horizontal merger analysis". Antitrust Law Journal 68, 685–710.
Federal Trade Commission (1999). "A study of the Commission's divestiture process". Bureau of Competition. (http://www.ftc.gov/os/1999/08/divestiture.pdf.)
Federal Trade Commission (2002). "Understanding mergers: Strategy & planning, implementation and outcomes", Bureau of Economics, Merger Roundtable Papers. (http://www.ftc.gov/be/rt/mergerroundtable.htm.)
Federal Trade Commission and U.S. Department of Justice (1992). "Horizontal merger guidelines" (amended in 1997). (http://www.usdoj.gov/atr/public/guidelines/hmg.htm.)
Federal Trade Commission and U.S. Department of Justice (2003). "Merger challenges data: Fiscal years 1999–2003". (http://www.ftc.gov/os/2003/12/mdp.pdf.)
Federal Trade Commission and U.S. Department of Justice (2004). "FTC/DOJ joint workshop on merger enforcement", February 17–19. (http://www.ftc.gov/bc/mergerenforce/index.shtm.)
Federal Trade Commission and U.S. Department of Justice (2006). "Commentary on the horizontal merger guidelines". (www.ftc.gov/os/2006/03/CommentaryontheHorizontalMergerGuidelinesMarch2006.pdf.)
Fisher, F., McGowan, J. (1983). "On the misuse of accounting rates of return to infer monopoly profits". American Economic Review 73, 82–97.
Frech, H., Langenfeld, J., McCluer, R. (2004). "Elzinga-Hogarty tests and alternative approaches for market share calculations in hospital mergers". Antitrust Law Journal 71, 921–947.
Friedman, J. (1971). "A non-cooperative equilibrium for supergames". Review of Economic Studies 38, 1–12.
Froeb, L., Werden, G. (1998). "A robust test for consumer welfare enhancing mergers among sellers of a homogeneous product". Economics Letters 58, 367–369.
Fudenberg, D., Maskin, E. (1986). "The folk theorem in repeated games with discounting or with incomplete information". Econometrica 54, 533–554.
Fudenberg, D., Tirole, J. (1986). "A 'signal-jamming' theory of predation". RAND Journal of Economics 17, 366–376.
Fudenberg, D., Levine, D., Maskin, E. (1994). "The folk theorem with imperfect public information". Econometrica 62, 997–1039.
Fumagalli, C., Motta, M. (2006). "Exclusive dealing and entry: When buyers compete". American Economic Review 96, 785–795.
Genesove, D., Mullin, W. (2001). "Rules, communication, and collusion: Narrative evidence from the Sugar Institute case". American Economic Review 91, 379–398.
Genesove, D., Mullin, W. (2006). "Predation and its rate of return: The sugar industry, 1887–1914". RAND Journal of Economics 37, 47–69.
Gilbert, R. (1989). "Mobility barriers and the value of incumbency". In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. 1. North-Holland, Amsterdam, pp. 475–535.
Gilo, D., Moshe, Y., Spiegel, Y. (2006). "Partial cross ownership and tacit collusion". RAND Journal of Economics 37, 81–99.
Green, E., Porter, R. (1984). "Noncooperative collusion under imperfect price information". Econometrica 52, 87–100.
Grossman, S., Hart, O. (1986). "The costs and benefits of ownership: A theory of vertical and lateral integration". Journal of Political Economy 94, 691–719.
Guzman, A. (1998). "Is international antitrust possible?". New York University Law Review 73, 1501–1548.
Haltiwanger, J., Harrington, J. (1991). "The impact of cyclical demand movements on collusive behavior". RAND Journal of Economics 22, 89–106.
Harford, J., Li, K. (2007). "Decoupling CEO wealth and firm performance: The case of acquiring CEOs". Journal of Finance 62, 917–949.

Harrington, J. (1989). "Collusion among asymmetric firms: The case of different discount factors". International Journal of Industrial Organization 7, 289–307.
Harrington, J. (2004a). "Cartel pricing dynamics in the presence of an antitrust authority". RAND Journal of Economics 35, 651–673.
Harrington, J. (2004b). "Post-cartel pricing during litigation". Journal of Industrial Economics 52, 517–533.
Harrington, J. (2005). "Optimal cartel pricing in the presence of an antitrust authority". International Economic Review 46, 145–169.
Harrington, J. (2006a). "Corporate leniency programs and the role of the antitrust authority in detecting collusion". Competition Policy Research Center Discussion Paper CPDP–18–E.
Harrington, J. (2006b). "How do cartels operate?" Foundations and Trends in Microeconomics 2, 1–105.
Harrington, J. (2007). "Detecting cartels". In: Buccirossi, P. (Ed.), Advances in the Economics of Competition Law. MIT Press, Cambridge, in press.
Harrington, J., Skrzypacz, A. (2007). "Collusion under monitoring of sales", RAND Journal of Economics, in press.
Hart, O. (1995). Firms, Contracts, and Financial Structure. Clarendon Press, Oxford.
Hart, O., Moore, J. (1990). "Property rights and the nature of the firm". Journal of Political Economy 98, 1119–1158.
Hart, O., Tirole, J. (1990). "Vertical integration and market foreclosure". Brookings Papers on Economic Activity: Microeconomics 1990, 205–286.
Hastings, J. (2004). "Vertical relations and competition in the retail gasoline market: Empirical evidence from contract changes in Southern California". American Economic Review 94, 317–328.
Hausman, J., Leonard, G., Vellturo, C. (1996). "Market definition under price discrimination". Antitrust Law Journal 64, 367–386.
Hausman, J., Leonard, G., Zona, J. (1994). "Competitive analysis with differentiated products". Annales d'Economie et de Statistique 34, 159–180.
Hawk, B. (1988). "The American (anti-trust) revolution: Lessons for the EEC?" European Competition Law Review 9, 53–87.
Hayes, J., Shapiro, C., Town, R. (2007). "Market definition in crude oil: Estimating the effects of the BP/ARCO merger". Antitrust Bulletin, in press.
Healy, P., Palepu, K., Ruback, R. (1992). "Does corporate performance improve after mergers?" Journal of Financial Economics 31, 135–175.
Heyer, K. (2006). "Welfare standards and merger analysis: Why not the best?" Competition Policy International 2 (2), 29–54.
Holmstrom, B., Tirole, J. (1989). "The theory of the firm". In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. 1. North-Holland, Amsterdam, pp. 61–133.
Hovenkamp, H. (2005). Antitrust Law: An Analysis of Antitrust Principles and Their Application, 2nd edn, vol. 11. Aspen Publishers, New York.
Hylton, K. (2003). Antitrust Law: Economic Theory and Common Law Evolution. Cambridge University Press, Cambridge.
Innes, R., Sexton, R. (1994). "Strategic buyers and exclusionary contracts". American Economic Review 84, 566–584.
Isaac, R., Smith, V. (1985). "In search of predatory pricing". Journal of Political Economy 93, 320–345.
Ivaldi, M., Jullien, B., Rey, P., Seabright, P., Tirole, J. (2003a). "The economics of tacit collusion". IDEI Working Paper No. 186, Final Report for DG Competition, European Commission, IDEI, Toulouse.
Ivaldi, M., Jullien, B., Rey, P., Seabright, P., Tirole, J. (2003b). "The economics of unilateral effects". IDEI Working Paper No. 222, Interim Report for DG Competition, European Commission, IDEI, Toulouse.
Jacquemin, A., Slade, M. (1989). "Cartels, collusion, and horizontal merger". In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. 1. North-Holland, Amsterdam, pp. 415–473.
Joskow, P., Klevorick, A. (1979). "A framework for analyzing predatory pricing policy". Yale Law Journal 89, 213–270.
Jung, Y., Kagel, J., Levin, D. (1994). "On the existence of predatory pricing: An experimental study of reputation and entry deterrence in the chain-store game". RAND Journal of Economics 25, 72–93.

Kandori, M., Matsushima, H. (1998). "Private observation, communication and collusion". Econometrica 66, 627–652.
Kaplan, S. (Ed.) (2000). Mergers and Productivity. University of Chicago Press, Chicago.
Kaplow, L. (1982). "The accuracy of traditional market power analysis and a direct adjustment alternative". Harvard Law Review 95, 1817–1848.
Kaplow, L. (1985). "Extension of monopoly power through leverage". Columbia Law Review 85, 515–556.
Kaplow, L. (2004). "On the (ir)relevance of distribution and labor supply distortion to public goods provision and regulation". Journal of Economic Perspectives 18 (4), 159–175.
Kaplow, L., Shavell, S. (1994). "Why the legal system is less efficient than the income tax in redistributing income". Journal of Legal Studies 23, 667–681.
Katz, M. (1989). "Vertical contractual relations". In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. 1. North-Holland, Amsterdam, pp. 655–721.
Katz, M., Shapiro, C. (2003). "Critical loss: Let's tell the whole story". Antitrust 17 (2), 49–56.
Kim, E., Singal, V. (1993). "Mergers and market power: Evidence from the airline industry". American Economic Review 83, 549–569.
Klemperer, P. (2002). "How (not) to run auctions: The European 3G telecom auctions". European Economic Review 46, 829–845.
Klemperer, P. (2004). Auctions: Theory and Practice. Princeton University Press, Princeton.
Kolasky, W., Dick, A. (2003). "The Merger Guidelines and the integration of efficiencies into antitrust review of horizontal mergers". Antitrust Law Journal 71, 207–251.
Koller, R. (1971). "The myth of predatory pricing: An empirical study". Antitrust Law and Economics Review 4 (4), 105–123.
Kovacic, W., Marshall, R., Marx, L., Schulenberg, S. (2006). "Quantitative analysis of coordinated effects", working paper.
Krattenmaker, T., Salop, S. (1986). "Anticompetitive exclusion: Raising rivals' costs to achieve power over price". Yale Law Journal 96, 209–293.
Kreps, D., Scheinkman, J. (1983). "Quantity precommitment and Bertrand competition yield Cournot outcomes". Bell Journal of Economics 14, 326–337.
Kreps, D., Wilson, R. (1982). "Reputation and imperfect information". Journal of Economic Theory 27, 253–279.
Kühn, K.-U. (2001). "Fighting collusion by regulating communication between firms". Economic Policy: A European Forum 16 (32), 167–197.
Kwoka, J., White, L. (2004). "Manifest destiny? The Union Pacific and Southern Pacific railroad merger (1996)". In: Kwoka, J., White, L. (Eds.), The Antitrust Revolution: Economics, Competition, and Policy, 4th edn. Oxford University Press, New York, pp. 27–51.
Lambson, V. (1987). "Optimal penal codes in price-setting supergames with capacity constraints". Review of Economic Studies 54, 385–397.
Lambson, V. (1994). "Some results on optimal penal codes in asymmetric Bertrand supergames". Journal of Economic Theory 62, 444–468.
Lambson, V. (1995). "Optimal penal codes in nearly symmetric Bertrand supergames with capacity constraints". Journal of Mathematical Economics 24, 1–22.
Landes, W., Posner, R. (1981). "Market power in antitrust cases". Harvard Law Review 94, 937–996.
Leary, T. (2002). "The essential stability of merger policy in the United States". Antitrust Law Journal 70, 105–142.
Levenstein, M., Suslow, V. (2006). "What determines cartel success?" Journal of Economic Literature 44, 43–95.
Levin, D. (1990). "Horizontal mergers: The 50 percent benchmark". American Economic Review 80, 1238–1245.
Lichtenberg, F., Siegel, D. (1987). "Productivity and changes in ownership of manufacturing plants". Brookings Papers on Economic Activity 1987, 643–683.
Litan, R., Shapiro, C. (2002). "Antitrust policy in the Clinton administration". In: Frankel, J., Orszag, P. (Eds.), American Economic Policy in the 1990s. MIT Press, Cambridge, Mass., pp. 435–485.

Mankiw, N.G., Whinston, M. (1986). "Free entry and social inefficiency". RAND Journal of Economics 17, 48–58.

Marshall, R., Meurer, M. (2004). "Bidder collusion and antitrust law: Refining the analysis of price fixing to account for the special features of auction markets". Antitrust Law Journal 72, 83–118.

Marvel, H. (1982). "Exclusive dealing". Journal of Law and Economics 25, 1–25.

Mason, C., Phillips, O., Nowell, C. (1992). "Duopoly behavior in asymmetric markets: An experimental evaluation". Review of Economics and Statistics 74, 662–670.

Masten, S., Snyder, E. (1993). "United States versus United Shoe Machinery Corporation: On the merits". Journal of Law and Economics 36, 33–70.

McAfee, R.P., Schwartz, M. (1994). "Opportunism in multilateral vertical contracting: Nondiscrimination, exclusivity, and uniformity". American Economic Review 84, 210–230.

McAfee, R.P., Williams, M. (1992). "Horizontal mergers and antitrust policy". Journal of Industrial Economics 40, 181–187.

McAfee, R.P., Mialon, H., Williams, M. (2004). "When are sunk costs barriers to entry? Entry barriers in economic and antitrust analysis". American Economic Review Papers and Proceedings 94 (2), 461–465.

McCutcheon, B. (1997). "Do meetings in smoke-filled rooms facilitate collusion?" Journal of Political Economy 105, 330–350.

McGee, J. (1958). "Predatory price cutting: The Standard Oil (N.J.) case". Journal of Law and Economics 1, 137–169.

McGuckin, R., Nguyen, S. (1995). "On productivity and plant ownership change: New evidence from the Longitudinal Research Database". RAND Journal of Economics 26, 257–276.

Melamed, A.D. (2006). "Exclusive dealing agreements and other exclusionary conduct: Are there unifying principles?". Antitrust Law Journal 73, 375–412.

Milgrom, P., Roberts, J. (1982). "Predation, reputation, and entry deterrence". Journal of Economic Theory 27, 280–312.

Morton, F. (1997). "Entry and predation: British shipping cartels, 1879–1929". Journal of Economics & Management Strategy 6, 679–724.

Motta, M. (2004). Competition Policy: Theory and Practice. Cambridge University Press, Cambridge.

Motta, M., Polo, M. (2003). "Leniency programs and cartel prosecution". International Journal of Industrial Organization 21, 347–379.

Muris, T. (1999). "The government and merger efficiencies: Still hostile after all these years". George Mason Law Review 7, 729–752.

Nalebuff, B. (2003). "Bundling, tying, and portfolio effects, Part 1: Conceptual issues", Department of Trade and Industry Economics Paper No. 1. (http://www.dti.gov.uk/files/file14774.pdf.)

Nevo, A. (2000a). "Mergers with differentiated products: The case of the ready-to-eat cereal industry". RAND Journal of Economics 31, 395–421.

Nevo, A. (2000b). "A practitioner's guide to estimation of random-coefficients logit models of demand". Journal of Economics & Management Strategy 9, 513–548.

Nevo, A. (2001). "Measuring market power in the ready-to-eat cereal industry". Econometrica 69, 307–342.

O'Brien, D., Salop, S. (2000). "Competitive effects of partial ownership: Financial interest and corporate control". Antitrust Law Journal 67, 559–614.

O'Brien, D., Shaffer, G. (1992). "Vertical control with bilateral contracts". RAND Journal of Economics 23, 299–308.

O'Brien, D., Wickelgren, A. (2003). "A critical analysis of critical loss analysis". Antitrust Law Journal 71, 161–184.

Office of Fair Trading (2005). "Ex post evaluation of mergers", March 2005. (http://www.oft.gov.uk.)

Ordover, J., Saloner, G. (1989). "Predation, monopolization, and antitrust". In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. 1. North-Holland, Amsterdam, pp. 537–596.

Ordover, J., Willig, R. (1981). "An economic definition of predation: Pricing and product innovation". Yale Law Journal 91, 8–53.

Parker, P., Röller, L.-H. (1997). "Collusive conduct in duopolies: Multimarket contact and cross-ownership in the mobile telephone industry". RAND Journal of Economics 28, 304–322.
Pautler, P. (2003). "Evidence on mergers and acquisitions". Antitrust Bulletin 48, 119–221.
Perry, M., Porter, R. (1985). "Oligopoly and the incentive for horizontal merger". American Economic Review 75, 219–227.
Pesendorfer, M. (2003). "Horizontal mergers in the paper industry". RAND Journal of Economics 34, 495–515.
Peters, C. (2003). "Evaluating the performance of merger simulation: Evidence from the U.S. airline industry". Journal of Law and Economics 49, 627–649.
Pitofsky, R. (1999). "Efficiencies in defense of mergers: Two years after". George Mason Law Review 7, 485–493.
Popofsky, M. (2006). "Defining exclusionary conduct: Section 2, the rule of reason, and the unifying principle underlying antitrust rules". Antitrust Law Journal 73, 435–482.
Porter, R. (1983). "A study of cartel stability: The Joint Executive Committee, 1880–1886". Bell Journal of Economics 14, 301–314.
Porter, R. (2005). "Detecting collusion". Review of Industrial Organization 26, 147–167.
Porter, R., Zona, J. (2004). "Bidding, bid rigging, and school milk prices: *Ohio v. Trauth* (1994)". In: Kwoka, J., White, L. (Eds.), The Antitrust Revolution: Economics, Competition, and Policy, 4th edn. Oxford University Press, New York, pp. 211–232.
Posner, R. (1969). "Oligopoly and the antitrust laws: A suggested approach". Stanford Law Review 21, 1562–1606.
Posner, R. (2001). Antitrust Law, 2nd edn. University of Chicago Press, Chicago.
Prager, R., Hannan, T. (1998). "Do substantial horizontal mergers generate significant price effects? Evidence from the banking industry". Journal of Industrial Economics 46, 433–452.
Rasmusen, E., Ramseyer, M., Wiley, J. (1991). "Naked exclusion". American Economic Review 81, 1137–1145.
Ravenscraft, D., Scherer, F.M. (1987). Mergers, Sell-Offs, and Economic Efficiency. Brookings Institution, Washington, DC.
Ravenscraft, D., Scherer, F.M. (1989). "The profitability of mergers". International Journal of Industrial Organization 7, 101–116.
Rey, P., Tirole, J. (2007). "A primer on foreclosure". In: Armstrong, M., Porter, R. (Eds.), Handbook of Industrial Organization, vol. 3. Elsevier Science Publishers, New York, in press.
Reynolds, R., Snapp, B. (1986). "The competitive effects of partial equity interests and joint ventures". International Journal of Industrial Organization 4, 141–153.
Robinson, J. (1933). The Economics of Imperfect Competition. Macmillan, London.
Robinson, M. (1985). "Collusion and the choice of auction". RAND Journal of Economics 16, 141–145.
Ross, T. (1992). "Cartel stability and product differentiation". International Journal of Industrial Organization 10, 1–13.
Rotemberg, J., Saloner, G. (1986). "A supergame-theoretic model of price wars during booms". American Economic Review 76, 390–407.
Rothschild, R. (1999). "Cartel stability when costs are heterogeneous". International Journal of Industrial Organization 17, 717–734.
Salant, S., Switzer, S., Reynolds, R. (1983). "Losses from horizontal merger: The effects of an exogenous change in industry structure on Cournot-Nash equilibrium". Quarterly Journal of Economics 98, 185–199.
Salinger, M. (1990). "The concentration-margins relationship reconsidered". Brookings Papers on Economic Activity: Microeconomics 1990, 287–321.
Salop, S. (1986). "Practices that (credibly) facilitate oligopoly co-ordination". In: Stiglitz, J., Mathewson, G.F. (Eds.), New Developments in the Analysis of Market Structure. MIT Press, Cambridge, Mass., pp. 265–290.
Salop, S. (2005). "Question: What is the real and proper antitrust welfare standard? Answer: The *true* consumer welfare standard". Statement before the Antitrust Modernization Commission, November 4, 2005.

Salop, S. (2006). "Exclusionary conduct, effect on consumers, and the flawed profit-sacrifice standard". Antitrust Law Journal 73, 311–374.
Salop, S., Scheffman, D. (1983). "Raising rivals' costs". American Economic Review Papers and Proceedings 73 (2), 267–271.
Scharfstein, D. (1984). "A policy to prevent rational test-market predation". RAND Journal of Economics 15, 229–243.
Schelling, T. (1960). The Strategy of Conflict. Harvard University Press, Cambridge, Mass.
Scherer, F.M. (1976). "Predatory pricing and the Sherman Act: A comment". Harvard Law Review 89, 869–890.
Schmalensee, R. (1989). "Inter-industry studies of structure and performance". In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. 2. North-Holland, Amsterdam, pp. 951–1009.
Schmalensee, R. (2004). "Sunk costs and antitrust barriers to entry". American Economic Review Papers and Proceedings 94 (2), 471–475.
Schmalensee, R., Willig, R. (Eds.) (1989a). Handbook of Industrial Organization, vol. 1. North-Holland, Amsterdam.
Schmalensee, R., Willig, R. (Eds.) (1989b). Handbook of Industrial Organization, vol. 2. North-Holland, Amsterdam.
Segal, I., Whinston, M. (2000a). "Exclusive contracts and protection of investments". RAND Journal of Economics 31, 603–633.
Segal, I., Whinston, M. (2000b). "Naked exclusion: Comment". American Economic Review 90, 296–309.
Segal, I., Whinston, M. (2003). "Robust predictions for bilateral contracting with externalities". Econometrica 71, 757–791.
Selten, R. (1978). "The chain-store paradox". Theory and Decision 9, 127–159.
Shapiro, C. (1986). "Exchange of cost information in oligopoly". Review of Economic Studies 53, 433–446.
Shapiro, C. (1989). "Theories of oligopoly behavior". In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. 1. North-Holland, Amsterdam, pp. 329–414.
Shapiro, C. (1996). "Mergers with differentiated products". Antitrust 10 (2), 23–30.
Simpson, J., Wickelgren, A. (in press). "Naked exclusion, efficient breach, and downstream competition". American Economic Review.
Smith, A. [1776] (1970). An Inquiry into the Nature and Causes of the Wealth of Nations, Books I–III. Penguin Books, London, first published 1776.
Snyder, C. (1996). "A dynamic theory of countervailing power". RAND Journal of Economics 27, 747–769.
Spence, A.M. (1977). "Entry, capacity, investment and oligopolistic pricing". Bell Journal of Economics 8, 534–544.
Spence, A.M. (1978). "Tacit coordination and imperfect information". Canadian Journal of Economics 3, 490–505.
Spence, A.M. (1981). "The learning curve and competition". Bell Journal of Economics 12, 49–70.
Spier, K., Whinston, M. (1995). "On the efficiency of privately stipulated damages for breach of contract: Entry barriers, reliance, and renegotiation". RAND Journal of Economics 26, 180–202.
Stigler, G. (1964). "A theory of oligopoly". Journal of Political Economy 72, 44–61.
Stole, L. (2007). "Price discrimination and imperfect competition". In: Armstrong, M., Porter, R. (Eds.), Handbook of Industrial Organization, vol. 3. Elsevier Science Publishers, New York, in press.
Sutton, J. (2007). "Market structure: Theory and evidence". In: Armstrong, M., Porter, R. (Eds.), Handbook of Industrial Organization, vol. 3. Elsevier Science Publishers, New York, in press.
Telser, L. (1966). "Cutthroat competition and the long purse". Journal of Law and Economics 9, 259–277.
Tirole, J. (1988). The Theory of Industrial Organization. MIT Press, Cambridge, Mass.
Tirole, J. (2005). "The analysis of tying cases: A primer". Competition Policy International 1 (1), 1–25.
Varian, H. (1989). "Price discrimination". In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. 1. North-Holland, Amsterdam, pp. 597–654.
Vasconcelos, H. (2005). "Tacit collusion, cost asymmetries, and mergers". RAND Journal of Economics 36, 39–62.

Vickers, J. (2005). "Abuse of market power". Economic Journal 115, F244–F261.
Vita, M., Sacher, S. (2001). "The competitive effects of not-for-profit hospital mergers: A case study". Journal of Industrial Economics 49, 63–84.
Vives, X. (2001). Oligopoly Pricing: Old Ideas, New Tools. MIT Press, Cambridge, Mass.
Waehrer, K., Perry, M. (2003). "The effects of mergers in open-auction markets". RAND Journal of Economics 34, 287–304.
Walle de Ghelcke, B. van de, Gerven, G. van (2004). Competition Law of the European Community. Matthew Bender, New York.
Weiman, D., Levin, R. (1994). "Preying for monopoly? The case of Southern Bell Telephone Company, 1894–1912". Journal of Political Economy 102, 103–126.
Werden, G. (1996). "A robust test for consumer welfare enhancing mergers among sellers of differentiated products". Journal of Industrial Economics 44, 409–413.
Werden, G. (2006). "Identifying exclusionary conduct under Section 2: The 'no economic sense' test". Antitrust Law Journal 73, 413–433.
Werden, G., Froeb, L. (1994). "The effects of mergers in differentiated products industries: Logit demand and merger policy". Journal of Law, Economics, & Organization 10, 407–426.
Werden, G., Froeb, L. (1998). "The entry-inducing effect of horizontal mergers: An exploratory analysis". Journal of Industrial Economics 46, 525–543.
Werden, G., Froeb, L. (2007). "Unilateral competitive effects of horizontal mergers". In: Buccirossi, P. (Ed.), Advances in the Economics of Competition Law. MIT Press, Cambridge, Mass., in press.
Werden, G., Froeb, L., Scheffman, D. (2004). "A *Daubert* discipline for merger simulation". Antitrust 18 (3), 89–95.
Werden, G., Joskow, A., Johnson, R. (1991). "The effects of mergers on price and output: Two case studies from the airline industry". Managerial and Decision Economics 12, 341–352.
Whinston, M. (1990). "Tying, foreclosure, and exclusion". American Economic Review 80, 837–859.
Whinston, M. (2006). Lectures on Antitrust Economics. MIT Press, Cambridge, Mass.
White House Task Force Report on Antitrust Policy (1968). Antitrust Law and Economics Review 2 (2), 11–52.
Williamson, O. (1968). "Economies as an antitrust defense: The welfare trade-offs". American Economic Review 58, 18–36.
Williamson, O. (1975). Markets and Hierarchies: Analysis and Antitrust Implications. Free Press, New York.
Williamson, O. (1977). "Predatory pricing: A strategic and welfare analysis". Yale Law Journal 87, 284–340.
Williamson, O. (1985). The Economic Institutions of Capitalism. Free Press, New York.
Yamey, B. (1972). "Predatory price cutting: Notes and comments". Journal of Law and Economics 15, 129–142.
Zerbe, R., Cooper, D. (1982). "An empirical and theoretical comparison of alternative predation rules". Texas Law Review 61, 655–715.

Cases

American Column & Lumber Co. v. United States, 257 U.S. 377 (1921).
American Tobacco Co. v. United States, 328 U.S. 781 (1946).
Board of Trade of City of Chicago v. United States, 246 U.S. 231 (1918).
Broadcast Music, Inc. v. Columbia Broadcasting System, Inc., 441 U.S. 1 (1979).
Brooke Group Ltd. v. Brown & Williamson Tobacco Corp., 509 U.S. 209 (1993).
Brown Shoe Co. v. United States, 370 U.S. 294 (1962).
E.I. du Pont de Nemours & Co. v. Federal Trade Commission, 729 F.2d 128 (2nd Cir. 1984).
Federal Trade Commission v. Indiana Federation of Dentists, 476 U.S. 447 (1986).
Federal Trade Commission v. Staples, Inc., 970 F. Supp. 1066 (D.D.C. 1997).
Federal Trade Commission v. Swedish Match, 131 F. Supp. 2d 151 (D.D.C. 2000).

Federal Trade Commission v. Tenet Healthcare Corp., 17 F. Supp. 2d 937 (E.D. Mo. 1998), *rev'd*, 186 F. 3d 1045 (8th Cir. 1999).
In re High Fructose Corn Syrup Antitrust Litigation, 295 F.3d 651 (7th Cir. 2002).
Jefferson Parish Hospital District No. 2 v. Hyde, 466 U.S. 2 (1984).
LePage's, Inc. v. 3M, 324 F.3d 141 (3d Cir. 2003).
Lorain Journal Co. v. United States, 342 U.S. 143 (1951).
Matsushita Electric Industrial Co. Ltd. v. Zenith Radio Corp., 475 U.S. 574 (1986).
Monsanto Co. v. Spray-Rite Service Corp., 465 U.S. 752 (1984).
National Collegiate Athletic Association v. Board of Regents of the University of Oklahoma, 468 U.S. 85 (1984).
National Society of Professional Engineers v. United States, 435 U.S. 679 (1978).
Spectrum Sports v. McQuillan, 506 U.S. 447 (1993).
Standard Fashion Co. v. Magrane-Houston Co., 258 U.S. 346 (1922).
Standard Oil Co. of California (Standard Stations) v. United States, 337 U.S. 293 (1949).
Standard Oil Co. of New Jersey v. United States, 221 U.S. 1 (1911).
Theatre Enterprises v. Paramount Film Distributing Corp., 346 U.S. 537 (1954).
Toys "R" Us, Inc. v. Federal Trade Commission, 221 F.3d 928 (7th Cir. 2000).
United States v. Aluminum Co. of America, 148 F.2d 416 (2d Cir. 1945).
United States v. AMR Corp., 335 F.3d 1109 (10th Cir. 2003).
United States v. Container Corporation of America, 393 U.S. 333 (1969).
United States v. Dentsply International, Inc., 399 F.3d 181 (3d Cir. 2005).
United States v. E.I. du Pont de Nemours & Co., 351 U.S. 377 (1956).
United States v. General Dynamics Corp., 415 U.S. 486 (1974).
United States v. Griffith, 334 U.S. 100 (1948).
United States v. Grinnell Corp., 384 U.S. 563 (1966).
United States v. Microsoft Corp., 56 F.3d 1448 (D.C. Cir. 1995).
United States v. Microsoft Corp., 253 F.3d 34 (D.C. Cir. 2001).
United States v. Philadelphia National Bank, 374 U.S. 321 (1963).
United States v. United Shoe Machinery Corp., 110 F. Supp. 295 (D. Mass. 1953), affirmed per curiam, 347 U.S. 521 (1954).
United States v. Von's Grocery Co., 384 U.S. 270 (1966).

***

ICI, Ltd v. Commission, ECR 619, §64, Case 48/69 (1972).
Michelin v. European Commission, ECR 3461 §30, Case 322/81 (1983).
Tetra Pak International SA v. Commission, ECR I-5951, Case C-333/94P (1996).
ECS/Akzo II, Decision of the Commission, 1985 OJ L 374/1 (December 14, 1985).
Hilti, Decision of the Commission, 1988 OJ L 65/19 (December 22, 1987).

Chapter 16

REGULATION OF NATURAL MONOPOLY

PAUL L. JOSKOW*

Department of Economics, Massachusetts Institute of Technology

Contents

* A significant amount of the material in this chapter has been drawn from my lectures on the regulation of natural monopolies in the graduate course that I have taught at MIT for many years. I have had the privilege of teaching this course multiple times with each of my colleagues Nancy Rose, Dick Schmalensee and Jean Tirole. In many cases I can no longer distinguish what came initially from their lectures and what came from mine. To the extent that I have failed to give adequate credit to their contributions, I must apologize and thank them for what they have taught me over the years.

Handbook of Law and Economics, Volume 2
Edited by A. Mitchell Polinsky and Steven Shavell

DOI: 10.1016/S1574-0730(07)02016-6

Abstract

This chapter provides a comprehensive overview of the theoretical and empirical literature on the regulation of natural monopolies. It covers alternative definitions of natural monopoly, public interest regulatory goals, alternative regulatory institutions, price regulation with full information, price regulation with imperfect and asymmetric information, and topics on the measurement of the effects of price and entry regulation in practice. The chapter also discusses the literature on network access and pricing to support the introduction of competition into previously regulated monopoly industries.

Keywords

Natural monopoly, economies of scale, sunk costs, price regulation, public utilities, incentive regulation, performance based regulation, network access pricing

JEL classification: K20, K23, L43, L51, L90

1. Introduction

Textbook discussions of price and entry regulation typically are motivated by the asserted existence of an industry with "natural monopoly" characteristics [e.g. Pindyck and Rubinfeld (2001, p. 50)]. These characteristics make it economical for a single firm to supply services in the relevant market rather than two or more competing. Markets with natural monopoly characteristics are thought to lead to a variety of economic performance problems: excessive prices, production inefficiencies, costly duplication of facilities, poor service quality, and to have potentially undesirable distributional impacts.

Under U.S. antitrust law the possession of monopoly power itself is not illegal. Accordingly, where monopoly "naturally" emerges due to the attributes of the technology for producing certain services, innovation or unique skills, antitrust policy cannot be relied upon to constrain monopoly pricing. Nor are the antitrust laws well suited to responding to inefficiencies resulting from entry of multiple firms in the presence of economies of scale and scope. Accordingly, antitrust policy alone cannot be relied upon to respond to the performance problems that may emerge in markets with natural monopoly characteristics. Administrative regulation of prices, entry, and other aspects of firm behavior have instead been utilized extensively in the U.S. and other countries as policy instruments to deal with real or imagined natural monopoly problems.

American economists began analyzing natural monopolies and the economic performance issues that they may raise over 100 years ago [Lowry (1973), Sharkey (1982), Phillips (1993)] and refinements in the basic concepts of the cost and demand attributes that lead to natural monopoly have continued to evolve over time [Kahn (1970), Schmalensee (1979), Baumol, Panzar, and Willig (1982), Phillips (1993), Laffont and Tirole (1993, 2000), Armstrong, Cowan, and Vickers (1994)]. On the policy side, price and entry regulation supported by natural monopoly arguments began to be introduced in the U.S. in the late 19th century. The scope of price and entry regulation and its institutional infrastructure grew considerably during the first 75 years of the 20th century, covering additional industries, involving new and larger regulatory agencies, and expanding from the state to the federal levels. However, during the 1970s both the natural monopoly rationale for and the consequences of price and entry regulation came under attack from academic research and policy makers [Winston (1993)]. Since then, the scope of price and entry regulation has been scaled back in many regulated industries. Some industries have been completely deregulated. Other regulated industries have been or are being restructured to promote competition in potentially competitive segments and new performance-based regulatory mechanisms are being applied to core network segments of these industries that continue to have natural monopoly characteristics [Winston (1993), Winston and Peltzman (2000), Armstrong and Sappington (2006), Joskow (2006)]. Important segments of the electric power, natural gas distribution, water, and telecommunications industries are generally thought to continue to have natural monopoly characteristics and continue to be subject to price and entry regulation of some form.

Economic analysis of natural monopoly has focused on several questions which, while related, are somewhat different. One question is a normative question: What is the most efficient number of sellers (firms) to supply a particular good or service given firm cost characteristics and market demand characteristics? This question leads to technological or cost-based definitions of natural monopoly. A second and related question is a positive question: What are the firm production or cost characteristics and market demand characteristics that lead some industries "naturally" to evolve to a point where there is a single supplier (a monopoly) or a very small number of suppliers (an oligopoly)? This question leads to behavioral and market equilibrium definitions of natural monopoly which are in turn related to the technological attributes that characterize the cost-based definitions of natural monopoly. A third question is also a normative question: If an industry has "a tendency to monopoly" what are the potential economic performance problems that may result and how do we measure their social costs? This question leads to an evaluation of the losses in economic efficiency and other social costs resulting from an "unregulated" industry with one or a small number of sellers. This question in turn leads to a fourth set of questions: When is government regulation justified in an industry with natural monopoly characteristics and how can regulatory mechanisms best be designed to mitigate the performance problems of concern?

Answering this set of questions necessarily requires both theoretical and empirical examinations of the strengths and weaknesses of alternative regulatory mechanisms. Regulation is itself imperfect and can lead to costly and unanticipated firm responses to the incentives created by regulatory rules and procedures. The costs of regulation may exceed the costs of unregulated naturally monopoly or significantly reduce the net social benefits of regulation. These considerations lead to a very important policy-relevant question. Are imperfect unregulated markets better or worse than imperfectly regulated markets in practice?

Finally, firms with de facto legal monopolies that are subject to price and entry regulation inevitably are eventually challenged by policymakers, customers or potential competitors to allow competing suppliers to enter one or more segments of the lines of business in which they have de facto legal monopolies. Entry may be induced by changes in technology on the costs and demand sides or as a response to price, output and cost distortions created by regulation itself. These considerations lead to a final set of questions. How do changes in economic conditions or the performance of the institution of regulated monopoly lead to public and private interests in replacing regulated monopoly with competition? How can policymakers best go about evaluating the desirability of introducing competition into these industries and, if competition appears to be desirable, fashioning transition mechanisms to allow it to evolve efficiently?

Scholarly law and economics research focused on answering these positive and normative questions has involved extensive theoretical, empirical, and institutional analysis. Progress has been made as well through complementary research in law, political sciences, history, organizational behavior and corporate finance. This chapter adopts a similarly comprehensive perspective of the research on the natural monopoly problem relevant to a law and economics handbook by including theoretical, empirical,

policy and institutional research and identifying linkages with these other disciplines. Indeed, research on economic regulation has flourished because of cooperative research efforts involving scholars in several different fields. Nevertheless, the Chapter's primary perspective is through the lense of economic analysis and emphasizes the economic efficiency rationales for and economic efficiency consequences of government regulation of prices and entry of firms producing services with natural monopoly characteristics. In addition, several industries have been subject to price and entry regulation which clearly do not have natural monopoly characteristics (e.g. trucking, natural gas and petroleum production, airlines, agricultural commodities). These multi-firm regulated industries have been studied extensively and in many cases have now been deregulated [Joskow and Noll (1981), Joskow and Rose (1989)]. This chapter will not cover regulation of multi-firm industries where natural monopoly is an implausible rationale for regulation.

The chapter proceeds in the following way. The first substantive section discusses alternative definitions of natural monopoly and the attributes of technologies, demand and market behavior that are thought to lead to natural monopolies from either a normative or a positive (behavioral) perspective. The section that follows it examines the rationales for introducing price and entry regulation in sectors that are thought to have natural monopoly characteristics. This section enumerates the economic performance problems that may result from natural monopoly, focusing on economic efficiency considerations while identifying equity, distributional and political economy factors that have also played an important role in the evolution of regulatory policy. This discussion leads to a set of normative goals that are often defined for regulators that reflect these performance problems. Section 4 provides a brief discussion of the historical evolution of and legal foundations for price and entry regulation, emphasizing developments in the U.S. Section 5 discusses alternative institutional frameworks for regulating legal monopolies, including direct legislative regulation, franchise contracts, and regulation by independent regulatory commissions.

The chapter then turns to a discussion of optimal regulatory mechanisms given different assumptions about the information available to the regulator and the regulated firm and various economic and legal constraints. Section 6 discusses optimal price regulation of a monopoly with subadditive costs in a world where the regulator is perfectly informed about the regulated firm's costs and has the same information about the attributes of demand faced by the regulated firm as does the firm. This section includes a discussion of Ramsey-Boiteux pricing, two-part tariffs, more general models of nonlinear pricing, and peak load pricing. The section that follows it begins a discussion of regulatory mechanisms in a world where the regulator has limited or imperfect and asymmetric information about the attributes of the regulated firm's cost opportunities, the attributes of consumer demand for its services and the managerial effort exerted by its managers. It discusses how traditional cost-of-service regulation evolved in an effort to reduce the regulator's information disadvantage and the early analytical models that sought to understand the efficiency implications of cost of service or rate of return regulation. This discussion sets the stage for a review of the more recent theoretical literature on incentive or performance based regulation where the regulator has imperfect

and asymmetric information about firm's cost opportunities, demand, and managerial effort attributes and the basic practical lessons that can be learned from it. Section 9 turns to recent empirical research that seeks to measure the effects of price and entry regulation of legal monopolies using a variety of performance indicia. The section focuses on post-1990 research on the effects of incentive regulation in practice. Earlier empirical research is discussed in Joskow and Rose (1989).

Individual vertical segments or lines of business of many industries that had been regulated as vertically integrated monopolies for many years have been opened up to competition in recent years (e.g. intercity telecommunications, electricity generation, natural gas production) as remaining "network infrastructure" segments remain regulated and provide a platform for competition in the potentially competitive segments. The introduction and success of competition in one or more of these vertical segments often involves providing access to network facilities that continue to be controlled by the incumbent and subject to price regulation. Accordingly, introducing competition in these segments requires regulators to define the terms and conditions of access to these "essential" network facilities and ensure that they are implemented. Section 10 discusses theoretical research on competitive entry and network access pricing. A brief set of conclusions completes the chapter.

2. Definitions of natural monopoly

2.1. Technological definitions of natural monopoly

I have not been able to determine definitively when the term "natural monopoly" was first used. Sharkey (1982, pp. 12–20) provides an excellent overview of the intellectual history of economic analysis of natural monopolies and I draw on it and the references he sites here and elsewhere in this chapter. He concludes [Sharkey (1982, p. 14)] that John Stuart Mill was the first to speak of natural monopolies in 1848. In his *Principles of Economics*, Alfred Marshall (1890) discusses the role of "increasing returns" in fostering monopoly and oligopoly, though he appears to be skeptical that pure monopolies can endure for very long or profitably charge prices that are significantly above competitive levels without attracting competitive entry [Marshall (1890, pp. 238–239, 329, 380)]. Posner (1969, p. 548) writes that natural monopoly "does not refer to the actual number of sellers in a market but to the relationship between demand and the technology of supply." Carlton and Perloff (2004, p. 104) write that "When total production costs would rise if two or more firms produced instead of one, the single firm in a market is called a "natural monopoly."

These are simple expositions of the technological definition of natural monopoly: a firm producing a single homogeneous product is a natural monopoly when it is less costly to produce any level of output of this product within a single firm than with two or more firms. In addition, this "cost dominance" relationship must hold over the full range of market demand for this product $Q = D(p)$.

Consider a market for a homogeneous product where each of k firms produces output q^i and total output is given by $Q = \sum_k q^i$. Each firm has an identical cost function $C(q^i)$. According to the technological or cost-based definition of natural monopoly, a natural monopoly will exist when:

$$C(Q) < C(q^1) + C(q^2) + \cdots + C(q^k)$$

since it is less costly to supply output Q with a single firm rather than splitting production up between two or more competing firms. Firm cost functions that have this attribute are said to be *subadditive* at output level Q (Sharkey, 1982, p. 2). When firm cost functions have this attribute for all values of Q (or all values consistent with supplying all of the demand for the product $Q = D(p)$) then the cost function is said to be *globally subadditive*. As a result, according to the technological definition of natural monopoly, a necessary condition for a natural monopoly to exist for output Q of some good is that the cost of producing that good is subadditive at Q.

Assume that firm i's cost function is defined as:[1]

$$C^i = F + cq^i$$

then the firm's average cost of production

$$\mathrm{AC}^i = F/q^i + c$$

declines continuously as its output expands. When a firm's average cost of production declines as its output expands its production technology is characterized by *economies of scale*. A cost function for a single-product firm characterized by declining average total cost over the relevant range of industry output from 0 to $q^i = Q$ is subadditive over this output range. Accordingly, in the single product context, economies of scale over the relevant range of q is a sufficient condition to meet the technological definition of natural monopoly. Figure 1 depicts the cost function for a firm with economies of scale that extend well beyond the total market demand (Q) depicted by the inverse demand function $P = D(Q)$. We note as well that when there are economies of scale up to firm out level q it will also be the case that average cost will be greater than marginal cost over this range of output ($F/q^i + c > c$ in the simple example above).[2]

In the single product case, economies of scale up to $q^i = Q$ is a *sufficient* but not a *necessary* condition for subadditivity over this range or, by the technological definition, for natural monopoly. However, it may still be less costly for output to be produced in a single firm rather than multiple firms even if the output of a single firm has expanded

[1] It should be understood that cost functions utilized here are technically $C = C(q, \mathbf{w})$ where $\mathbf{w}$ is a vector of input prices that we are holding constant at this point. They also reflect cost-minimization by the firm in the sense that the marginal rate of transformation of one input into another is equal to the associated input price ratio.

[2] Some definitions of natural monopoly assert that the relevant characteristic is declining marginal cost. This is wrong.

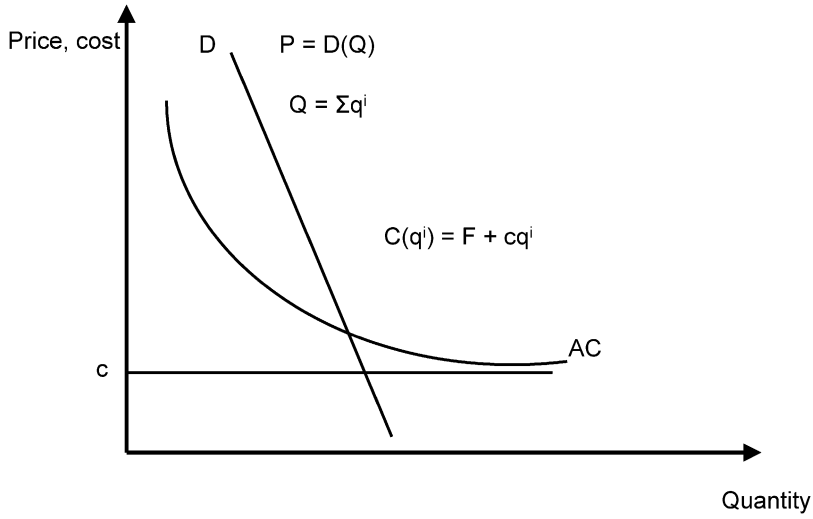


Figure 1. Economies of scale.

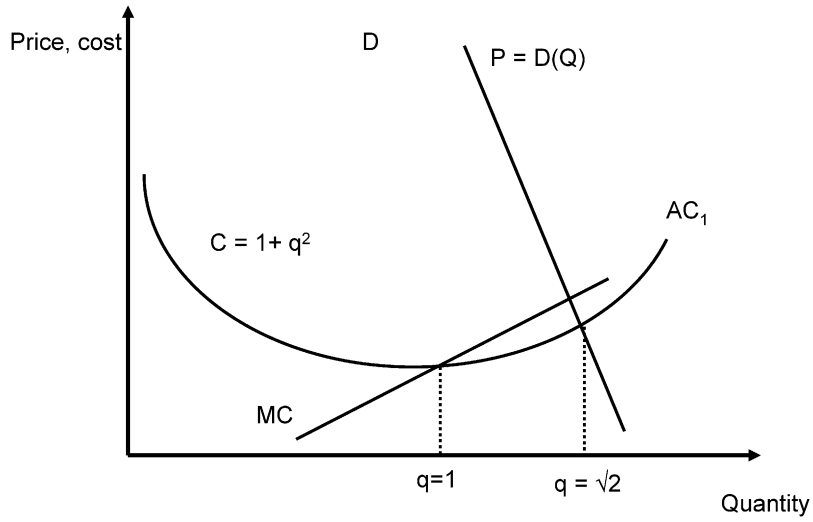


Figure 2. Subadditivity and diseconomies of scale.

beyond the point where there are economies of scale. Consider the total cost function for a firm $C = 1 + q^2$ and the associated average cost function $\text{AC} = q + 1/q$ depicted as AC_1 in Figure 2. There is a range of output where there are economies of scale ($q < 1$). The cost function then flattens out ($q = 1$) and then enters a range of decreasing returns

to scale ($q > 1$). However, this cost function is still subadditive for some values of $q > 1$, despite the fact that for $q > 1$ there is decreasing returns to scale. This is the case because the market demand $P = D(Q)$ is not large enough to support efficient production by two firms for some levels of industry output $Q > 1$.

Assume that firm 1 produces $q^1 = 1$ to produce at minimum efficient scale. Consider a second firm 2 with the same costs that could also produce at minimum efficient scale $q^2 = 1$. If both firms produced at minimum efficient scale total output would be 2 and total cost would be 4. If a single firm produced output $q = 2$, total cost would be 5, so it is more efficient to produce total industry output $Q = 2$ with two firms rather than one. However, it is apparent that for total output levels between $Q = 1$ and $Q = \sqrt{2}$ it is less costly to allow the first firm to operate in a range of decreasing returns to scale than it is to supply with two firms, both producing at greater than minimum efficient scale. Similarly for $q > \sqrt{2}$, it is less costly to supply with two firms rather than one and the cost function is not subadditive in this range.

Accordingly, the set of firm cost functions that are subadditive encompasses a wider range of cost functions than those that exhibit economies of scale over the entire (relevant) range of potential industry output. Specifically, in the single product case, the firm's cost functions must exhibit economies of scale over some range of output but it will still be subadditive in many cases beyond the point where economies of scale are exhausted and until industry output is large enough to make it economical to add a second firm.

There are some implicit assumptions regarding the firm's cost function $C(q)$ that should be noted here. First, it is a long run "economic cost function" in the sense that it reflects the assumption that the firm produces any particular output efficiently, given the underlying production function and input prices, and that inputs are fully adjusted to prevailing input prices and the quantity produced. That is, there is no "X-inefficiency" reflected in the firm's costs. Capital related costs in turn reflect the firm's opportunity cost of capital (r), economic depreciation (d), and the value of the capital invested in productive assets (K) measured at the current competitive market value of the associated assets. That is, the firm's total costs of production include the current period rental cost of capital $V = (r + d)K$. Accordingly, capital costs are not treated explicitly as being sunk costs for the technological definition of natural monopoly. These implicit assumptions have important implications for a variety of issues associated with behavioral definitions of natural monopoly, the measurement of the social costs of unregulated natural monopolies, the social costs of regulation, and the design of effective regulatory mechanisms. I will turn to these issues presently.

The technological definition of natural monopoly can be generalized to take account of multiproduct firms. For this purpose, multiproduct firms are firms that have technologies that make it more economical to produce two or more products within the same firm than in two or more firms. Production technologies with this attribute are characterized by *economies of scope*. Consider two products q_1 and q_2 that can be produced by a firm with a cost function $C(q_1, q_2)$. Define $\mathbf{q}^i$ as a vector of the two products $\mathbf{q}^i = (q_1^i, q_2^i)$. There are N vectors of the two products with the attribute that $\sum q_1^i = q_1$

and $\sum q_2^i = q_2$. Then the cost function $C(q_1, q_2)$ is subadditive if:

$$C\left(\sum q_1^i, \sum q_2^i\right) = C\left(\sum \mathbf{q}^i\right) < \sum C(\mathbf{q}^i)$$

for all N vectors of the products. This definition can be generalized to any number of products.

What attributes of a production technology/cost function will lead to multiproduct subadditivity? The technology must be characterized by some form of *economies of scope* and some form of *multiproduct economies of scale*.

By *economies of scope* we mean that it is more economical to produce the two products in one firm rather than multiple firms:

$$C(q_1, q_2) < C(q_1, 0) + C(0, q_2)$$

There are several concepts of *multiproduct economies of scale* depending upon how one slices the multiproduct cost function:

a. *Declining average incremental cost* for a specific product
b. *Declining ray average cost* for varying quantities of a set of multiple products that are bundled in fixed proportion

Define the *incremental cost* of producing product q_1 holding q_2 constant as

$$\text{IC}(q_1|q_2) = c(q_1, q_2) - c(0, q_2)$$

and define the average incremental cost of producing q_1 as

$$\text{AIC}(q_1|q_2) = \left[c(q_1, q_2) - c(0, q_2)\right]/q_1$$

If the AIC declines as the output of q_1 increases (holding q_2 constant) then we have *declining average incremental* cost of q_1. This is a measure of *single product economies of scale* in a multiproduct context. We can perform the same exercise for changes in q_2 holding q_1 constant to determine whether there are declining average incremental costs for q_2 and, in this way, determine whether the cost function is characterized by *declining average incremental cost* for each product.

We can think of fixing the *proportion* of the multiple products that are produced at some level (e.g. $q_1/q_2 = k$) in the two-product case) and then ask what happens to costs as we increase the quantity of both outputs produced holding their relative output proportions constant. Does the average cost of the bundle decline as the size of the bundle (holding the output proportions constant) increases?

Let λ be a number greater than one. If the total costs of producing this "bundle" of output increase less than proportionately with λ then there are multiproduct economies of scale along a ray defined by the product proportions k. This is called *declining ray average costs* for $q_1/q_2 = k$.

$$c(q_1, q_2|q_1/q_2 = k) > c(\lambda q_1, \lambda q_2|q_1/q_2 = k)/\lambda$$

By choosing different proportions of the products produced (alternative values for k) by the firm we can trace out the cost functions along different rays in q_1, q_2 space and

determine whether there are economies of scale or declining ray average costs along each ray. Then there are *multiproduct economies of scale* in the sense of declining ray average cost for any combination of q_1 and q_2 when:

$$C(\lambda q_1, \lambda q_2) < \lambda C(q_1, q_2)$$

For example, consider the cost function (Sharkey, 1982, p. 5)

$$C = q_1 + q_2 + (q_1 q_2)^{1/3}$$

This cost function exhibits multiproduct economies of scale since

$$\lambda C(q_1, q_2) = \lambda q_1 + \lambda q_2 + \lambda (q_1 q_2)^{1/3}$$
$$C(\lambda q_1, \lambda q_2) = \lambda q_1 + \lambda q_2 + \lambda^{2/3} (q_1 q_2)^{1/3}$$

and thus

$$C(\lambda q_1, \lambda q_2) < \lambda C(q_1, q_2)$$

However, this cost function exhibits *diseconomies of scope* rather than economies of scope since:

$$C(q_1, 0) = q_1$$
$$C(0, q_2) = q_2$$
$$C(q_1, 0) + C(0, q_2) = q_1 + q_2 < q_1 + q_2 + (q_1 q_2)^{1/3} = c(q_1, q_2)$$

As a result, this multiproduct cost function is not subadditive despite the fact that it exhibits declining ray average cost. It would be less costly to produce the two products in separate firms.

Subadditivity of the cost function, or natural monopoly, in the multiproduct context requires both a form of multiproduct cost complementarity (e.g. economies of scope[3]) and a form of multiproduct economies of scale over at least some range of the output of the products. For example, the multiproduct cost function [discussed by Sharkey (1982, p. 7)

$$C(q_1, q_2) = (q_1)^{1/4} + (q_2)^{1/4} - (q_1 q_2)^{1/4}$$

exhibits economies of scope. It also exhibits economies of scale in terms of both declining average incremental cost and declining ray average cost at every level of output of the two products. It is obvious that costs are lower when the products are produced together rather than separately by virtue of the term $(-(q_1 q_2)^{1/4})$ in the cost function. There are also declining ray average cost and declining average incremental cost for each product. This is the case because the cost of producing a particular combination of the two outputs increases less than proportionately with increases in the scale of the

[3] Or one of a number of other measures of cost complementarity.

bundle of two products produced by virtue of the power 1/4 in the cost function. Similarly for the average incremental cost of q_1 and q_2 individually. It can be shown that this cost function is subadditive at every output level or *globally subadditive*.

The necessary and sufficient conditions for global subadditivity of a multiproduct cost function are complex and it is not particularly useful to go into those details here. Interested readers should refer to Sharkey (1982) and to Baumol, Panzar, and Willig (1982). As already discussed, economies of scope is a necessary condition for a multiproduct cost function to be subadditive. One set of sufficient conditions for subadditivity of a multiproduct cost function is that it exhibit both economies of scope and declining average incremental cost for all products. An alternative set of sufficient conditions is that the cost function exhibit both declining average incremental cost for all products plus an alternative measure of multiproduct cost complementarity called *trans-ray convexity*. Trans-ray convexity requires that multiproduct economies outweigh any single product diseconomies of scale. For example, it may be that there are single product economies of scale for product 1, diseconomies of scale for product 2, but large multiproduct economies. Then it could be less costly to produce q_1 and q_2 together despite the diseconomies of scale in producing q_2 to take advantage of the multiproduct economies available from joint production. A third alternative sufficient condition is that the cost function exhibit *cost complementarity*, defined as the property that increased production of any output reduces (does not increase) marginal costs of all other outputs. As in the case of single product cost functions, the necessary conditions regarding scale economies are less strict and allow for output to expand into a range of diseconomies of scale or diseconomies of scope since it may be less costly to produce at a point where there are diseconomies than it is to incur the costs of suboptimal production from a second firm.

2.2. *Behavioral and market equilibrium considerations*

The previous section discussed the attributes of a firm's cost function that would make it most efficient from a cost of production perspective (assuming costs are minimized given technology and input prices as discussed earlier) to concentrate production in a single firm rather than in multiple firms. However, the intellectual evolution of the natural monopoly concept and public policy responses to it focused much more on the consequences for unregulated market outcomes of production technologies having such "natural monopoly" attributes. Moreover, historical discussions of the natural monopoly problem focus on more than economies of scale and related multiproduct cost complementarity concepts as potential sources of market distortions. Sharkey (1982, pp. 12–20) discusses this aspect of the intellectual history of economic analysis of natural monopoly as well. For example, in addition to economies of scale he notes that Thomas Farrer (1902) [referenced by Sharkey (1982, p. 15)] associated natural monopoly with supply and demand characteristics that included (a) the product or service supplied must be essential, (b) the products must be non-storable, (c) the supplier must have a favorable production location. In addition, Richard Ely (1937) [referenced

by Sharkey (1982, p. 15)] added the criteria that (a) the proportion of fixed to variable costs must be high and (b) the products produced from competing firms must be close substitutes. Bonbright (1961, pp. 11–17) suggested that economies of scale was a sufficient but not a necessary condition for natural monopoly and Posner (1969, p. 548) observed that "network effects" could lead to subadditive costs even if the cost per customer increased as the number of customers connected to the network increased; as more subscribers are connected to a telephone network, the average cost per subscriber may rise, but it may still be less costly for a single firm to supply the network service. Kaysen and Turner (1959, pp. 191, 195–196) note that economies of scale is a relative concept that depends on the proper definition of the relevant product and geographic markets and also argue that "ruinous competition" leading to monopoly may occur when the ratio of fixed to variable costs is high and identify what we would now call "sunk costs" as playing an important role leading to monopoly outcomes. Kahn (1970, pp. 119, 173) refers to both economies of scale and the presence of sunk or fixed costs that are a large fraction of total costs as attributes leading to destructive competition that will in turn lead a single firm or a very small number of firms in the market in the long run. He also recognizes the potential social costs of "duplicated facilities" when there are economies of scale or related cost-side economic attributes that lead single firm production to be less costly than multiple firm production.

These expanded definitions of the attributes of natural monopoly appear to me to confuse a set of different but related questions. In particular, they go beyond the normative concept of natural monopoly as reflecting technological and associated cost attributes that imply that a single firm can produce at lower cost than multiple firms, to examine the factors that "naturally" lead a market to evolve to a point where there is a single supplier (or not). That is, they include in their definition of natural monopoly the answer to the positive or behavioral question of what cost and demand attributes lead industries to evolve so that only a single firm survives in the long run? To some extent, some of these definitions also begin to raise normative questions about the consequences of the dynamics of the competitive process for costs, prices, and other aspects of social welfare in industries with natural monopoly characteristics. For example, Kaysen and Turner (1959, p. 191) associated natural monopoly that "leaves the field to one firm ... competition here is self-destructive." They go on to assert that "The major prerequisites for competition to be destructive are fixed or sunk costs that bulk large as a percentage of total costs" [Kaysen and Turner (1959, p. 173)]. Kahn (1970) observes that sunk costs must be combined with significant economies of scale for monopoly to "naturally" emerge in the market. So, the historical evolution of the natural monopoly doctrine reflects both a normative interest in identifying situations in which a single firm is necessary to achieve all economies of scale and multiproduct cost complementarities as well as a positive interest in identifying the attributes of costs and demand that lead to market conditions that are "unsuitable for competition" to prevail and the associated normative performance implications for prices, costs and other attributes of social welfare.

Absent regulatory constraints on pricing and entry, the presence of subadditive costs per se do not necessarily lead to the conclusion that a single firm—a monopoly—will "naturally" emerge in equilibrium. And if a monopoly "naturally" does emerge in equilibrium, a variety of alternative pricing patterns may result depending on cost, demand, and behavioral attributes that affect opportunities for price discrimination, competitive entry and the effects of potential entry on incumbent behavior. After all, many models of imperfect competition with two or more firms are consistent with the assumption that the competing firms have cost functions that are characterized by economies of scale over at least some range of output. Nor, as we shall see presently, if a single firm emerges in equilibrium is it *necessarily* the case that it will charge prices that yield revenues that exceed a breakeven level. On the other, hand, if a single firm (or a small number of firms) emerges in equilibrium it may have market power and charge prices that yield revenues that exceed the breakeven level for at least some period of time, leading to lower output and higher unit costs than is either first-best or second-best efficient (i.e. given a break-even constraint).

In order to draw positive conclusions about the consequences of subadditive costs for the attributes of short run and long run firm and market behavior and performance we must make additional assumptions about other attributes of a firm's costs, the nature of competitive interactions between firms *in* the market and interactions between firms in the market with potential entrants *into* the market when the firm's long run production costs are subadditive. Moreover, if more than one firm survives in equilibrium—e.g. a duopoly—the equilibrium prices, quantities and costs may be less desirable from an economic performance perspective than what is theoretically *feasible* given the presence of subadditive costs and other constraints (e.g. breakeven constraints). This latter kind of result is the foundation for arguments for introducing price and entry regulation in industries with natural monopoly characteristics despite the fact that multiple firms may survive in equilibrium and compete, but compete imperfectly.

2.3. Sunk costs

The most important cost attribute that is not reflected explicitly in the traditional technological definitions of natural monopoly that turn on the presence of subadditive firm production costs is the existence and importance of sunk costs. Sunk cost considerations also provide the linkage between subadditivity, behavioral definitions of natural monopoly, and the economic performance problems that are thought to arise from unregulated natural monopolies. Sunk costs are associated with investments made in long-lived physical or human assets whose value in alternative uses (i.e. to produce different products) or at different locations (when transportation costs are high) is lower than in its intended use. At the extreme, an investment might be worthless in an alternative use. Sunk costs are a "short run" cost concept in the sense that the associated assets eventually are valueless in their intended use and are retired. However, because the assets are long-lived, the short run may be quite long from an economic perspective. Sunk costs are not directly captured in long run neoclassical cost functions since

these cost functions reflect the assumption that capital assets can be rented on a period by period basis and input proportions are fully adjusted to prevailing input prices and output levels. Accordingly, sunk costs have not been considered directly in technological definitions of natural monopoly that turn only on cost subadditivity. Yet, sunk costs are quite important both theoretically and empirically for obtaining a comprehensive understanding of the natural monopoly problem as it has emerged in practice. Sunk cost considerations are important both to explain why some industries "naturally" evolve to a point where one or a very small number of firms survive and to measure the social welfare consequences of the market structures and associated, price, cost and quality attributes of these markets in the absence of price and entry regulation (Sutton, 1991). As discussed below, sunk cost considerations are also important for establishing regulated prices for incumbents when their industries are opened up to competition [Hausman (1997), Pindyck (2004)].

Most of the industries that have been regulated based on natural monopoly arguments—railroads, electric power, telephone, gas pipelines, water networks, cable television networks, etc.—have the attribute that a large fraction of their total costs are sunk capital costs. Moreover, it has been argued that a meaningful economic definition of economies of scale requires that there be at least some sunk costs and, for these purposes, thinking about there being fixed costs without there also being sunk costs is not particularly useful (Weitzman, 1983). Indeed, Weitzman argues that sunk costs introduce a time dimension into the cost commitment and recovery process that is essential to obtaining a useful concept of economies of scale. I will return to this issue presently.

2.4. Contestible markets: subadditivity without sunk costs

In order to get a better feeling for the importance of sunk cost and the behavioral attributes of firms in the market and potential entrants into the market, it is useful to focus first on the model of *contestable markets* developed by Baumol, Panzar, and Willig (1982) which assumes that costs are subadditive but generally ignores sunk costs. The examples that follow will focus on a single product case, but the extension to multiple products is straightforward, at least conceptually. Consider the single product situation in which there are n identical firms (where n is large) with identical cost functions $C(q^i) = F + cq^i$. This cost function is assumed to exhibit economies of scale over the entire range of q and thus is subadditive. One of the n firms (the incumbent) is in the market and the remaining $(n-1)$ firms are potential entrants. The declining average cost curve for the firm in the market is depicted in Figure 3 along with the inverse *market* demand for the product $p = D(q)$ (where the market demand is $Q = \sum q_i = D(p)$). F is assumed initially to be a fixed cost but not a sunk cost. It is not a sunk cost in the sense that firms can enter or exit the market freely without facing the risk of losing any of these fixed costs up to the point in time that the firm actually produces output q^i and incurs operating costs cq^i. If prices are not high enough to cover both a firm's operating cost cq^i and its associated fixed cost F, the firm will either not enter the market or will exit the market before committing to produce and avoiding incurring the associ-

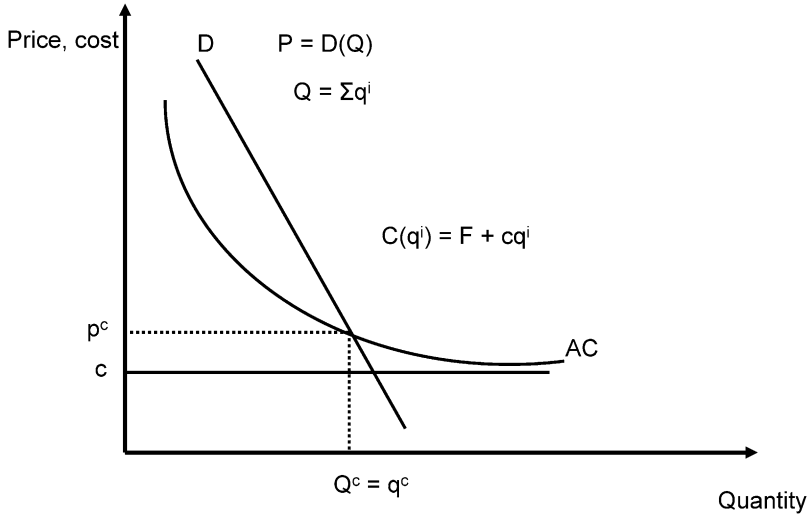


Figure 3. Economies of scale and break-even price.

ated costs. Thus, assuming that fixed costs are not sunk costs is equivalent to assuming that there is hyper-free entry and exit into and out of this market—there are no fixed commitment costs prior to actual production and the fixed costs of production can be avoided by a firm that has "entered" the market by simply not producing any output and effectively exiting the market without incurring any entry or exit costs.[4]

We are looking for an equilibrium where it is (a) *profitable* for one or more firms to enter (or remain in) the market and produce output ($pq^i \geq C(q^i)$), (b) *feasible* in the sense that supply and demand are in balance ($\sum q^i = Q = D(p)$), and (c) *sustainable* in the sense that no entrant can make a profit given the price charged by the incumbent(s)—there does not exist a price $p^a < p$ and an output $Q^a \leq D(p^a)$ such that $p^a q^a \geq C(q^a)$.

Figure 3 depicts an equilibrium that satisfies these conditions. At price p^c and output Q^c ($Q^c = q^c$) the incumbent firm exactly covers its costs and earns zero economic profit since $p^c = F/q^c + c = \mathrm{AC}_c$. It is not profitable for a second firm to enter with a price lower than p^c since it could not break even at any output level at a price less that p^c. The incumbent cannot charge a price higher than p^c (that is, p^c is not sustainable) because if the incumbent committed to a higher price one of the potential entrants could profitably offer a lower price, enter the market and take all of the incumbent's sales away. Moreover, with the incumbent committing to $p > p^c$, competition among potential entrants would drive the price down to p^c and it would be profitable for only

[4] Weitzman (1983) argues that there are no economies of scale in any meaningful sense in this case. Also see Tirole (1988, p. 307). This issue is discussed presently.

one of them to supply in equilibrium due to economies of scale. So, under these conditions the industry equilibrium is characterized by a single firm (a "natural monopoly"). However, the price p^c is the lowest uniform per unit (linear) price consistent with a firm breakeven (zero profit) constraint; the equilibrium price is not the classical textbook monopoly price but the lowest uniform per unit price that allows the single firm producing output to just cover its total costs of production. This price and output configuration is both feasible and sustainable. Thus, the threat of entry effectively forces the single incumbent supplier to charge the lowest uniform (linear) per unit price consistent with a breakeven constraint. As we shall see below, this equilibrium price which is equal to average total cost at the quantity that clears the market is the second-best efficient uniform (Ramsey-Boiteux) price when the firm is subject to a break-even constraint. Obviously, as I shall discuss in more detail below, it is not first best since the equilibrium price is greater than marginal cost (c).

These are remarkable results. They suggest that even with significant increasing returns we "naturally" get to a competitive equilibrium characterized by both a single firm exploiting the cost savings associated with global subadditivity and the lowest price that just allows a single firm exploiting all economies of scale to break even. This is as close to efficient uniform per unit (linear) pricing as we can expect in a market with private firms that are subject to a break-even constraint and have cost functions characterized by economies of scale. The classical textbook problem of monopoly pricing by an incumbent monopoly does not emerge here in equilibrium. In this case potential competition is extremely effective at constraining the ability of the incumbent to exercise market power when it sets prices, with no regulatory intervention at all. If this situation accurately reflected the attributes of the industries that are generally thought of as having "natural monopoly" characteristics then they would not appear to be particularly interesting targets for regulatory intervention (see the next section) since a fully informed regulator relying on uniform per unit prices could do no better than this.

Note, that even in this peculiar setting, an equilibrium with these attributes may not be *sustainable*. Consider the average cost function depicted in Figure 4 that has increasing returns up to point q^o and then enters a range of decreasing returns (perhaps due to managerial inefficiencies as the firm gets very large). The market demand curve crosses the average cost curve at the output level q^a and the average cost at this output level is equal to AC_a. In this case, the price that allows the single firm supplying the entire market to break even and that balances supply and demand is $p_a = AC_a$. However, this price is not sustainable against free entry. An entrant could, for example, profitably enter the market by offering to supply q^o at a price p_o equal to $AC_o + \varepsilon$. In this case, the entrant would have to ration demand to limit its output to q_o. The incumbent could continue to supply to meet the demand that has not been served by the new entrant, but would incur very high average costs to do so and would have to charge higher prices to break even. If we assume that the entrant supplied the consumers with the highest willingness to pay, there would not be any consumers willing to pay a price high enough for the incumbent to cover its average costs. Thus, the zero profit "natural monopoly" equilibrium is unstable.

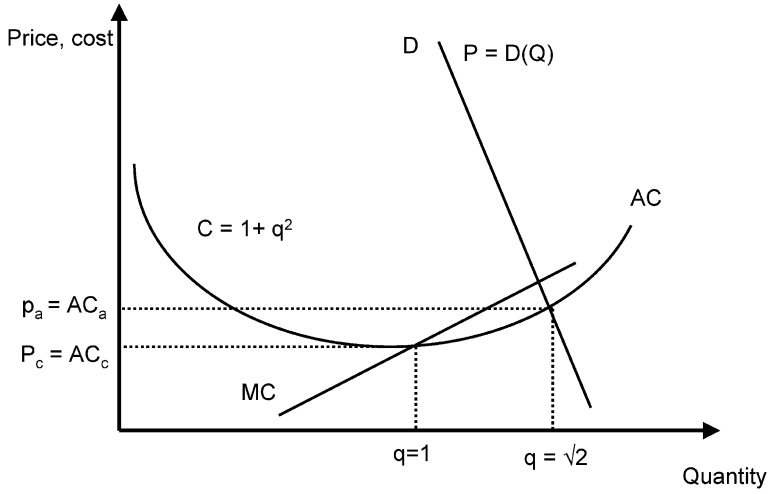


Figure 4. Subadditivity and diseconomies of scale.

In a multiproduct context, perfectly contestable markets (no sunk costs, free entry) have a symmetrical set of attributes. Following, Baumol, Panzar, and Willig (1982), if a sustainable allocation exists, it has the following attributes: (a) there is a single firm to take advantage of cost subadditivity, (b) the firm earns zero profits, (c), the revenues that a firm earns from any subset of products is greater than or equal to the incremental cost of producing that subset of products—there is no "cross-subsidization" in the sense that the prices charged for any product or set of products covers the *incremental* costs incurred to produce them, (d) the price of each product exceeds its (single product) marginal cost given the output of the other products, (e) under certain conditions the firm will voluntarily charge the second-best linear (Ramsey-Boiteux) prices [Baumol, Bailey, and Willig (1977)]. As in the case of a single product firm, the existence of a subadditive multiproduct cost function does not guarantee that a sustainable single-firm zero profit (break-even) configuration exists.

It seems to me that the primary point that emerges from the lengthy literature on contestable markets is that one cannot conclude that there are necessarily "monopoly problems" from the observation that there is one or a very small number of firms producing in a market. Prices may still be competitive in the second best sense ($P = \text{AC}$) in the presence of increasing returns because entry is so easy that it constrains the incumbent's prices. A monopoly naturally emerges, but it may have no or small social costs compared to feasible alternative allocations.

2.5. Sunk costs and barriers to entry

As I have already noted, the assumption that there are fixed costs but no sunk costs does not make a lot of economic sense [Weitzman (1983), Tirole (1988, p. 307)]. Sunk costs

introduce a time dimension into the analysis since sunk costs convey a stream of potential benefits over some period of time and once the associated cost commitments are made they cannot be shifted to alternative uses without reducing their value from that in the intended use. Sunk costs are what make the distinction between incumbents and potential entrants meaningful. Absent sunk costs there is no real difference between firms in the market and firms that are potentially in the market since entry and exit are costless. Sunk costs also create potential opportunities for strategic behavior by the incumbent designed both to sustain prices about the break-even level while simultaneously discouraging entry. If the fixed costs are fully avoidable up to the point that production actually takes place, a firm incurs no opportunity cost merely by entering the market. Whether a firm is "in" the market or "out" of the market is in some sense irrelevant in this case since there is no time dimension to the fixed costs. Firms are only "in" when they start to produce and can avoid incurring any fixed costs if they don't. From an entry and exit perspective, all costs are effectively variable over even the shortest time period relevant for determining prices and output.

An alternative approach that retains the notion that fixed costs are also at least partially sunk involves specifying a price competition game in which fixed cost (capacity) commitments can be adjusted more quickly than can the prices set by the firm and the associated quantities it commits to see [Tirole (1988, pp. 310–311)]. The fixed costs are sunk, but they are sunk for a shorter period of time than it takes to adjust prices. In this case, the contestable market result emerges as a generalization of Bertrand competition to the case where there are economies of scale [Tirole (1988, p. 310)]. However, for most industries, especially those that have typically been associated with the concept of natural monopoly, prices adjust much more quickly than can production capacity and its associated sunk costs. Accordingly, this approach to a contestable market equilibrium does not appear to be of much practical interest either.

A case for price and entry regulation based on a natural monopoly rationale therefore requires both significant increasing returns and long-lived sunk costs that represent a significant fraction of total costs. Indeed, this conclusion reflects a century of economic thinking about monopoly and oligopoly issues, with the development of contestable market theories being an intellectual diversion that, at best, clarifies the important role of sunk costs in theories of monopoly and oligopoly behavior.

Models of "wars of attrition" represent an interesting approach to natural monopoly that allows for increasing returns, sunk costs, exit, textbook monopoly pricing, and no incentives for re-entry in the face of textbook monopoly pricing at the end of the war [e.g. Tirole (1988, p. 311)]. In these models (to simplify considerably) there are two identical firms in the market at time 0. They compete Bertrand (for a random length of time) until one of them drops out of the market because the expected profits from continuing to stay in the market is zero. The remaining firm charges the monopoly price until there is entry by a second firm. However, re-entry by a competing firm is not profitable because the potential entrant sees that post-entry it will have to live through a war of attrition ($p = c$) and, even if it turned out to be the survivor, the expected profits from entry are zero. In this kind of model there is a period of intense competition when

prices are driven to marginal costs.[5] There is also inefficient duplication of facilities during this time period. Then there is a monopoly that "naturally" emerges at some point which charges a textbook pure monopoly price since it is not profitable for an entrant to undercut this price when faced with the threat of a price war. (There remains the question of why both firms entered in the first place.) This kind of war of attrition has been observed repeatedly in the early history of a number of industries that are often considered to have natural monopoly attributes: competing electric power distribution companies, railroad and urban transit lines in the late 19th and early 20th centuries and competing cable TV companies more recently.

War of attrition models also have interesting implications for the kind of "rent seeking" behavior identified by Posner (1975). Monopolies are valuable to their owners because they produce monopoly profits. These potential profits create incentives for firms to expend resources to attain or maintain a monopoly position. These resource expenditures could include things like investments in excess capacity to deter entry, duplication of facilities in the face of increasing returns as multiple firms enter the market to compete to be the monopoly survivor, and expenditures to curry political favor to obtain a legal monopoly through patent or franchise. In the extreme, all of the monopoly rents could be dissipated as a result of these types of expenditures being made as firms compete to secure a monopoly position. The worst of all worlds from a welfare perspective is that all of the monopoly profits are competed away through wasteful expenditures and consumers end up paying the monopoly price.[6]

The combination of increasing returns (and the multiproduct equivalents) combined with a significant component of long-lived sunk costs brings us naturally to more conventional monopoly and oligopoly models involving barriers to entry, entry deterrence and predation [Tirole (1988, Chapters 8 and 9)]. The natural monopoly problem and general models of barriers to entry, entry deterrence and oligopoly behavior are linked together, with natural monopoly being an extreme case. Sunk "capacity" costs create an asymmetry between firms that are "in" the market and potential entrants. This asymmetry can act as a *barrier to entry* by giving the first mover advantage to the firm that is the first to enter the market (the incumbent). Once costs have been sunk by an entrant they no longer are included in the opportunity costs that are relevant to the incumbent firm's pricing decisions. Sunk costs have commitment value because they cannot be reversed. This creates opportunities for an incumbent or first mover to behave strategically to deter entry or reduce the scale of entry.

[5] Since this is a repeated game it is possible that there are dynamic equilibria where the firms tacitly collude and keep prices high or non-cooperative price games with fixed capacity which lead to Cournot outcomes with higher prices.

[6] The war of attrition model that I outlined above is not this bad. There is wasteful "duplication" of facilities prior to the exit of one of the firms but prices are low so consumers benefit during the price war period. After exit consumers must pay the monopoly price, but the costs of duplication are gone. This outcome is worse than the second-best associated with the perfectly contestable market outcome with increasing returns by (effectively) no sunk costs. [Tirole (1988, pp. 311–314)].

In the simplest models of sequential entry with sunk costs and increasing returns [Tirole (1988, pp. 314–323)] firms compete in the long run by making capacity commitments, including how much capacity to accumulate upon entering a market and, for a potential entrant considering to enter to compete with an incumbent, whether or not it will commit capital to support even a modest quantity of capacity needed to enter the market at all. In making this decision the potential entrant must take account of the nature of the competition that will determine prices and entry post entry, at post-entry capacity and output levels. If the incumbent can profitably and credibly make commitments that indicate to the potential entrant that it will be unprofitable to enter due to the nature of the post-entry competition it will face, then competitive entry may be deterred. In these sequential entry games, the presence of sunk costs alone does not generally deter entry, but rather the strategic behavior of the first mover can reduce the amount of capacity the entrant commits to the markets and as a result, sustain post-entry prices above competitive levels and post-entry output below competitive levels. The combination of sunk costs and increasing returns can make small scale entry unprofitable so that the incumbent may deter entry completely.

Joe Bain (1956) characterized alternative equilibria that may arise in the context of significant economies of scale (to which today we would add multiproduct cost complementarities and sunk costs as well) that were subsequently verified in the context of more precise game-theoretic models [Tirole (1988, Chapter 8)]. These cases are:

Blockaded entry: Situations where there is a single firm in the market that can set the pure monopoly price without attracting entry. The incumbent competes as if there is no threat of entry. A situation like this may emerge where economies of scale are very important compared to the size of the market and where sunk costs are a large fraction of total costs. In this case, potential entrants would have to believe that if they entered, the post-entry competitive equilibrium would yield prices and a division of output that would not generate enough revenues to cover the entrant's total costs. This is the classic "pure monopoly" case depicted in microeconomics textbooks.

Entry deterrence: There is still no entry to compete with the incumbent, but the incumbent had to take costly actions to convince potential entrants that entry would be unprofitable. This might involve wasteful investments in excess capacity to signal a commitment to lower post-entry prices or long-term contracts with buyers to limit ("foreclose") the market available for a new entrant profitable to serve (Aghion and Bolton).

Accommodated entry: It is more profitable for the incumbent to engage in strategic behavior that *accommodates* profitable entry but limits the profitability of entry at other than small scale. Here the incumbent sacrifices some short-term pre-entry profits to reduce the scale of entry to keep prices higher than they would be if entry occurred at large scale.

2.6. Empirical evidence on cost subadditivity

Despite the extensive theoretical literature on natural monopoly, there is surprisingly little empirical work that measures the extent to which the costs of producing services that are typically thought of as natural monopolies are in fact subadditive. The most extensive research on the shape of firm level cost functions has been done for electricity [e.g. Christiansen and Greene (1976), Cowing (1974), Joskow and Rose (1985, 1989); Jamasb and Pollitt (2003)]. There has also been empirical work on cost attributes of water companies [Teeples and Glyer (1987)], telecommications firms [Evans (1983), Gasmi, Laffont, and Sharkey (2002)], cable television companies [Crawford (2000)], urban transit enterprises [Gagnepain, P. and M. Ivaldi (2002)], and multi-product utilities [Fraquelli, G., M. Picenza, and D. Vannoni (2004)]. Empirical analysis tends to find economies of scale (broadly defined) out to some level of firm output. However, much of this work fails properly to distinguish between classical economies of scale and what is best thought of as economies of density. Thus, for example, economies of scale in the distribution of natural gas may be exhausted by a firm serving let's say 3 million customers on an exclusive basis in a specific geographic area. However, whatever the size of the geographic area covered by the firm it would still be very costly to run two competing gas distribution systems down the same streets, because there are economies of scale or "density" associated with the installation and size of the pipes running down each street.

3. Why regulate natural monopolies?

It is important to recognize that in reality there is not likely to be a bright line between industries that are "natural monopolies" and those that are (imperfectly) "competitive." Whether an industry is judged to have classical natural monopoly characteristics inevitably depends on judgments about the set of substitute products that are included in the definition of the relevant product market (e.g. are Cheerios and Rice Crispies close enough products to be considered to be in the same product market? Are cable TV and Direct Broadcast Satellite in the same relevant product market?) and the geographic expanse over which the market is regulated (e.g. a supermarket may technically have natural monopoly characteristics if the geographic market is defined very narrowly, but may have no market power since consumers can easily switch between outlets at different geographic locations and the market cannot discriminate between consumers with good substitutes and those without). Moreover, many "competitive" industries are imperfectly competitive rather than perfectly competitive. They may have production technologies that give individual firms economies of scale but there is little cost sacrifice if there are several firms in the market. Or firms may have technologies that exhibit economies of scale over the production of a narrowly defined product or brand but there are many "natural monopolies" producing competing products or brands that are close substitutes for it and constrain the ability of suppliers to exercise market power.

In these cases, competition may be imperfect but the (theoretical) social welfare costs compared to the best *feasible* alternative industry configurations given economies of scale, differentiated product attributes, and break-even constraints may be quite small. This suggests that the technical definitions of natural monopoly employed (normative or positive) must be carefully separated from the questions of whether and how to regulate a particular industry.

The standard normative economic case for imposing price and entry regulations in industries where suppliers have natural monopoly characteristics is that (a) industries with natural monopoly characteristics will exhibit poor economic performance in a number of dimensions and (b) it is feasible in theory and practice for governments to implement price, entry and related supporting regulations in ways that improve performance (net) compared to the economic performance that would otherwise be associated with the unregulated market allocations. That is, the case for government regulation is that there are costly market failures whose social costs (consequences) can in principle be reduced (net) by implementing appropriate government regulatory mechanisms.

This "market failures" case for government regulation naturally leads to four sets of questions. First, what is the nature and magnitude of the performance problems that would emerge absent price and entry regulation in industries with natural monopoly characteristics? Second, what regulatory instruments are practically available to stimulate performance improvements and what are their strengths and weaknesses? Third, what are the performance attributes of the industry configuration that would be expected to emerge in a regulated environment? Fourth, are imperfect regulatory outcomes, on balance, likely to be superior to imperfect market outcomes taking all relevant performance criteria into account, including the direct and indirect costs of government regulation itself?

3.1. Economic efficiency considerations

The *economic efficiency* case for government regulation when an industry has natural monopoly characteristics has focused on a number of presumed attributes and the associated inefficiencies of market outcomes that are thought would arise in the absence of government regulation. Figure 5A displays two potential equilibria for an industry supplied by one single-product firm with subadditive costs. These equilibria provide normative benchmarks against which the performance attributes of "unregulated natural monopoly" can be compared. The firm's costs (AC_e and MC_e) assume that the firm produces a given level of output efficiently given input prices and technology. The price p_o reflects a second-best linear price that allows the firm just to cover its production costs and clears supply and demand. The price p_e is the first-best efficient price ($p = MC$) that leaves the regulated firm with a deficit and therefore requires government subsidies. Note that p_e is efficient in a broader general equilibrium sense only if we ignore the costs the government incurs to raise the revenues required to raise the funds to pay subsidies these through taxation. I will focus here on the case where the

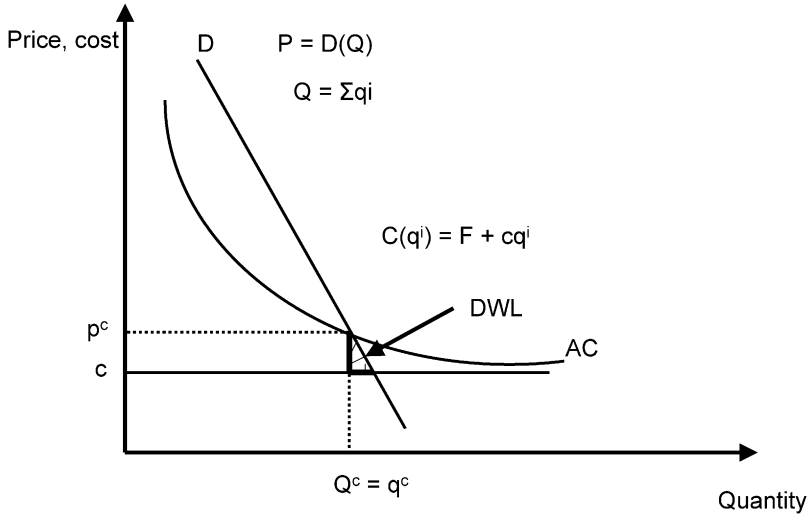


Figure 5A. Break-even price and dead-weight loss.

firm must break-even from the revenues it earns by selling services subject to price regulation to consumers.

Figure 5B depicts an alternative "unregulated natural monopoly" equilibrium where there are sunk costs and barriers to entry. The firm's production costs are now depicted as c_m (to keep the figure from becoming too confused, I have left out the average cost curve AC_M from Figure 5A which we should think of as being higher than AC_e and c_e), reflecting inefficient production by the monopoly, and the price charged by the firm is now $p_m > p_o > p_e$. In Figure 5B the rectangle marked with an "X" depicts the cost or "X-inefficiency" at output level Q_M associated with the monopoly configuration. The firm also spends real resources equal to R per year to maintain its monopoly position, say through lobbying activity or carrying excess capacity to deter entry. The case for regulation starts with a comparison of the attributes of the unregulated natural monopoly equilibrium depicted in Figure 5B with the efficient (first or second best with linear prices) equilibria depicted in Figure 5A.

Inefficient Price Signals: Prices greater than marginal cost: As have seen above, if a single or multiproduct monopoly naturally emerges (and is sustainable) in markets that are "contestable," then the resulting monopoly will not have much market power. At worst, the monopoly will set prices above marginal cost to satisfy a break-even constraint ($p = AC$ in the single product case and under certain conditions Ramsey prices in the multiproduct case (Baumol, Bailey, and Willig, 1977—more on this below). This in turn leads to the standard dead- weight loss triangle associated with the gap between prices and marginal cost (depicted by the triangle marked DWL in Figure 5A). However, these are the second-best linear prices and, assuming that public policy requires

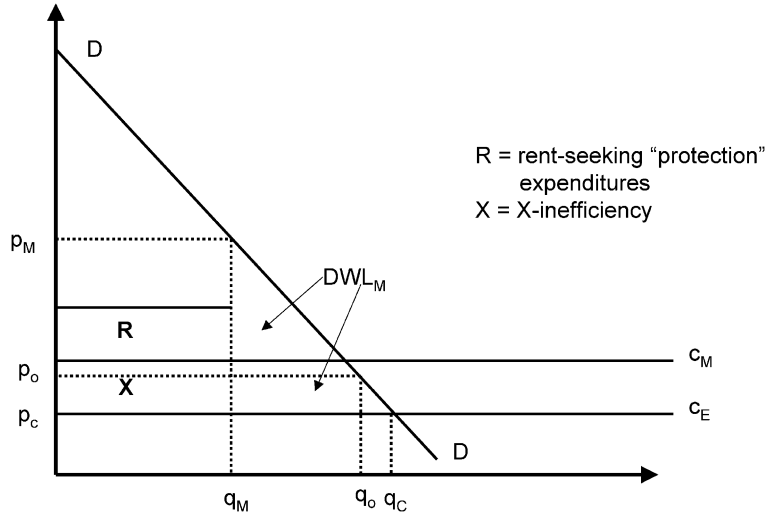


Figure 5B. Potential monopoly inefficiencies.

regulated firms to break-even and to charge linear prices, a regulator could not do any better. This is the second-best price p_o depicted in Figure 5A.

It has been argued that even with contestable markets we could do even better by regulating the monopoly and forcing it to sell at prices equal to marginal costs, using government subsidies to make up the difference between revenues and total costs. This argument normally assumes that the government can raise funds to finance the deficit without incurring any distortionary costs from the tax system put in place to generate the associated government revenues. Since governments do not generally rely on non-distortionary lump sum taxes to raise revenues, the theoretical case for regulating a natural/legal monopoly so as to constrain prices to equal marginal cost must depend on a comparison between the costs of distortions created by prices charged for the regulated services that exceed marginal cost and the costs of distortionary taxes that are otherwise required to pay for the firm's deficit. If the demands for the products and services sold by the regulated firm are fairly inelastic, as is often the case, the distortions resulting from raising prices above marginal cost to balance revenues and costs may not be larger than the distortions caused by increasing taxes to raise the revenues required to close the gap between revenues and costs when prices for the regulated product are force set equal to the relevant marginal costs [Laffont (1999)].

Putting the government subsidy arguments aside for the moment, if one believes that a monopoly has naturally emerged in a setting consistent with the assumptions associated with contestable markets then monopoly price distortions do not create a very good argument for price and/or entry regulation. That is, if the prices in Figure 5B where the same as those in 5A then from a pricing perspective there would be no loss from unregulated natural monopolies.

The more interesting "market failures" case for regulation to mitigate distortions associated with monopoly prices arise in situations in which there are significant barriers to entry and unregulated prices can be sustained at levels far above both marginal cost and average cost. This is the case depicted in Figure 5B where $p_M > p_o$. Since the market power possessed by an incumbent monopoly depends on both the presence of entry barriers and the elasticity of demand for the products sold by the firm, the social costs of monopoly will be higher the more important are entry barriers and the more inelastic is the demand for the relevant products. The polar case is one of blockaded entry (Bain, 1956) where the incumbent dominant firm faces a market demand with elasticity ε_d and sets the monopoly price:

$$P_M = MC/(1 + 1/\varepsilon_d)$$

and the Lerner Index of monopoly power is given by

$$(P_M - MC)/P_M = 1/\varepsilon_d$$

In this case, P_M is the highest price that a monopoly profitably can charge. The incumbent may charge a lower price to accommodate entry or through contracts to deter entry. After entry occurs, prices will likely fall as a result of there being more competition in the market, but they may not fall to the level where total revenues and total costs are equal ($P = AC$). That is, oligopoly price distortions may remain for some period of time. In all of these cases, the firm will charge prices greater than p_o, produce positive ("excess") economic profits that it will have an incentive to invest resources in order to protect, and yield a dead-weight loss from excessive prices alone (area DWL_M in Figure 5B) relative to the dead-weight loss at the break-even uniform unit (linear) price level ($P = AC$ in the single product case). However, if the elasticity of demand is very large in absolute value, any distortion resulting from monopoly pricing will be small.

Inefficient costs of production (including inefficient entry and exit): By definition, a natural monopoly involves production conditions such that it is less costly to produce output in a single firm than in two or more firms. In a contestable markets environment the monopoly in the market has high powered incentives to minimize production costs since it can be replaced instantly by a firm that will to supply at a price equal to average "minimum" (efficient production) total cost. Accordingly, firms or markets that are candidates for regulation must depart from the assumptions associated with contestable markets. That is, we should focus on cases where there are significant scale and scope economies and sunk costs represent a significant fraction of total costs.

In such cases, one potential source of increased production costs arises from the strategic behavior that an incumbent monopoly may engage in order to deter entry and protect its monopoly position. This may entail building excess capacity or spending resources in other ways ("rent seeking" behavior) to obtain or protect a monopoly position. Potentially all of the monopoly profits associated with the pure monopoly outcome may be "wasted" in this way. This type of social cost is depicted as the rectangle marked "R" in Figure 5B.

A second potential source of higher production costs results from inefficient entry of competitors. If the industry has natural monopoly attributes and multiple firms enter the market to supply output—even if competitors eventually exit after a war of attrition—excessive costs are naturally incurred due to duplication of facilities the failure to exploit all available economies of scale. Even in a contestable market the natural monopoly equilibrium may not be sustainable and inefficient entry may occur. The cost of duplicated facilities is not reflected in Figure 5B, but can be conceptualized as being related to the increase in average costs caused by each firm producing at a lower (suboptimal) output level.

A third potential source of production cost inefficiencies is the failure of the incumbent monopoly to minimize production costs—produce efficiently—at the output level it is producing, given technology and input prices. Cost minimization requires that the marginal rate of technical substitution between inputs equal the ratio of their respective prices. If we have a two input production function $q = F(K, L)$ where the rental rates for capital (K) and the wage rate for labor (L) are respectively r and w, then cost minimization at any output level requires that $F_K/F_L = r/w$, where F_K is the marginal product of capital and F_L is the marginal product of labor. Neoclassical profit maximizing monopoly firms minimize costs in this way. However, when there is separation of ownership and management and management gets satisfaction from managerial emoluments and gets disutility from effort, monopoly firms that are insulated from competition may exhibit "X-inefficiency" or managerial slack that leads to higher production costs. There is also some evidence that monopolies are more easily organized by unions which may extract some of the monopoly profits in the form of higher wages ($w_M > w$) [Salinger (1984), Rose (1987), Hendricks (1977)]. If wages are driven above competitive levels this will lead firms inefficiently to substitute capital for labor in production. These costs are depicted as $c_m > c_e$ and the associated social cost is depicted as rectangle marked "X" in Figure 5B.

Product quality and dynamic inefficiencies: Although, the issue has largely been unexplored in the context of natural monopoly *per se*, related literature in industrial organization that examines research and development, adoption of innovations in the production and product dimensions and the choice of product quality suggests that monopoly outcomes are likely to differ from competitive outcomes. Moreover, issues associated with the reliability of service (e.g. outages of the electric power network) and various aspects of the quality of service (e.g. queues for obtaining connections to the telephone network) are significant policy issues in many regulated industries. As a general matter, we know that monopoly will introduce a bias in the selection of quality, the speed of adoption of innovations, and investment in R&D. In simple static models of monopoly the bias turns on the fact that a profit maximizing monopoly looks at the willingness to pay for quality of the marginal consumer while social welfare is maximized by focusing on the surplus achieved by the average consumer (Spence, 1975).

However, the size and magnitude of any quality bias, compared to a social welfare-maximizing norm is ambiguous. The monopoly may supply too much or too little quality or have too little or too much incentive to invest in R&D and adopt innova-

tions depending on the circumstances, in particular whether the incumbent monopoly is threatened by potential entry, as well as the existence and nature of patent protection and spillovers from R&D [Tirole (1988, pp. 100–106, 361–414)]. This is not the place to review the extensive literature on the relationship between market structure and innovation, but I note only that it raises potentially important dynamic efficiency issues with market structures that evolve into monopolies. On the one hand, in situations where there are significant spillovers from R&D and innovation that would otherwise be captured by competing firms and lead to underinvestment in innovation in the product and process dimensions, regulatory policies that facilitate the internalization of these spillover effects, for example, by having a single firm serving the entire sector or providing for the recovery of R&D costs in product prices, might increase social welfare. On the other hand, depending on the circumstances, creating a monopoly and regulating the prices it can charge for new products could increase rather than decrease inefficiencies associated with product quality, R&D, and the adoption of product and process innovations.

Firm viability and breakeven constraints: As I have already noted, if the regulated monopoly is a private firm and there are no government subsidies available to support it, the government may be able to regulate the firm's prices and service quality, but it cannot compel it to supply output to balance supply and demand in the long run if it is unprofitable for it to do so. Accordingly, price and entry regulation also must confront one important set of constraints even in an ideal world where regulators have full information about a firm's cost opportunities, managerial effort levels, and attributes of demand faced by the regulated firm (we discuss regulation with asymmetric information in more detail below). Private firms will only supply goods and services if they expect to at least recover the costs of providing these goods and services. The relevant costs include the costs of materials and supplies, compensation necessary to attract suitable employees and to induce them to exert appropriate levels of effort, the direct cost of capital investments in the enterprise, a return of and on those investments, reflecting the opportunity cost of capital, economic depreciation, taxes, and other costs incurred to provide service. If the process through which regulated prices are set does not lead private firms to expect to earn enough revenues to cover these production and distribution costs the firm will not voluntarily supply the services. Since prices are regulated, supply and demand will not necessarily clear and prices that are set too low will lead to shortages in the short run and/or the long run and the use of non-price rationing to allocate scarce supplies. Accordingly, if we are to rely on regulated private monopolies to provide services, the regulatory process must have a price-setting process that provides the regulated firm with adequate financial incentives to induce them to provide services whose value to consumers exceeds the costs of supplying them.

At this point I will simply refer to this requirement as a breakeven-constraint defined as:

$$\sum_n p_i q_i \geq C(q_1, \ldots, q_{n-1}, q_n)$$

where q_i defines the total output of the different products supplied by the firm (or the same output supplied to different groups of consumers that are charged different prices or a combination of both) and $C(*)$ defines the associated costs. For now, let's think about $C(*)$ as being a static measure of the "efficient" level of costs given any particular output configuration. We will address differences between expected costs and realized costs and issues of cost inefficiency in more detail below.

There is an inherent conflict between the firm viability constraint and efficient pricing when costs are subadditive. Efficient pricing considerations would dictate that prices be set equal to marginal cost. But marginal cost pricing will not produce enough revenues to cover total costs, thus violating the firm viability or break-even constraint.[7] A great deal of the literature on price regulation has focused on responding to this conflict by implementing price structures that achieve the break-even constraint in ways that minimize the efficiency losses associated with departures from marginal cost pricing.

Moreover, because the interesting cases involve technologies where long-lived sunk costs are a significant fraction of total costs, the long-term credibility of regulatory rules plays an important role in convincing potential suppliers that the rules of the regulatory game will in fact fairly compensate them for the sunk costs that they must incur to provide service [Laffont and Tirole (1993, Chapter 10), Armstrong, Cowan, and Vickers (1994, pp. 85–91), Levy and Spiller (1994)]. This is the case because once costs are sunk, suppliers must be concerned that they will be "held-up" by the regulator. That is, once the costs are sunk, the regulator is potentially in a position to lower prices to a point where they cover only avoidable costs, causing the firm that has committed the sunk costs to fail to recover them. As I shall discuss presently, creating a regulatory process and judicial oversight system that constrains the ability of a regulatory agency to hold up a regulated firm in this way has proven to be a central component of regulatory systems that have been successful in attracting adequate investment and associated supplies to the regulated sectors. These "credibility" institutions include legal principles governing the formulas used to set prices and to review "allowable" costs, the structure of regulatory procedures and opportunities for judicial review, as well as de jure and de facto restrictions on competitive entry.

3.2. Other considerations

While this chapter will focus on the economic efficiency rationales for and consequences of the regulation of natural monopolies, we must recognize that the nature and performance of the institutions associated with regulated monopoly in practice reflect additional normative public policy goals and the outcomes of interest group politics.

Income distribution, "essential services," cross-subsidization and taxation by regulation: Although simple conceptualizations of economic efficiency are "indifferent" to

[7] In the single product case declining average cost is a necessary condition for marginal cost pricing to be unprofitable. In the multiproduct case, declining ray average cost is a necessary condition for marginal cost pricing to yield revenues that are less than total costs.

the distribution of surplus between consumers and producers, public policy generally is not. Thus, while the efficiency losses from classical monopoly pricing are measured by a welfare triangle reflecting the loss in the sum of consumers' and producers' surplus from higher prices and lower output, public policy has also been concerned with the transfer of income and wealth associated with the excess profits resulting from monopoly pricing as well. Even ignoring the fact that some of the monopoly profits may be eaten up by wasteful "rent seeking" expenditures, and the difficulties of calculating the ultimate effects on the distribution of income and wealth from monopoly pricing, it is clear that regulatory policy has historically been very concerned with mitigating monopoly profits by keeping prices at a level that roughly reflect the regulated firm's total production costs.

It also is quite clear that several of the industries that have evolved as regulated monopolies produce products access to which has come to be viewed as being "essential" for all of a nation's citizens. I use the term "access" here broadly to reflect both physical access (e.g. "universal service")[8] as well as "affordability" considerations. Electricity, telephone, and clean water services fall in this category. The argument is that absent price and entry regulation, suppliers of these services will not find it economical to expand into certain areas (e.g. rural areas) or if they do will charge prices that are too high given the incomes of the individuals living or firms producing (e.g. farms) in those areas. While there are no clear definitions of what kinds of services are essential, how much is essential, or what are the "reasonable" prices at which such services should be provided, these concepts have clearly played a role in the development of regulatory policies in many countries. This being said, it is hard to argue that food, for example, is any less essential than electricity. Yet there has been no interest in creating regulated legal monopolies for the production and distribution of food. Low-income consumers or residents of rural areas could simply be given subsidies by the government to help them to pay for the costs of services deemed essential by policymakers, as is the case for food stamps. Accordingly, the case for regulated monopoly and the case for subsidies for particular geographic areas or types of consumers appear to be separable policy issues that can in principle be addressed with different policy instruments.

These issues are joined when an industry does have natural monopoly characteristics, and the introduction of government regulation of prices and entry creates opportunities to use the regulated monopoly itself as a vehicle for implementing a product-specific, geographic, customer-type specific internal subsidy program rather than relying on the government's general budget to provide the subsidies directly. With regulated *legal monopoly* that bars competitive entry, regulated prices in some geographic areas or the prices charged to some classes of consumers or for some products can be set at levels above what would prevail if economic efficiency criteria alone were applied to set prices (more on this presently). The excess revenues generated by increasing these

[8] A universal service rationale may also be justified by the desire to internalize network externalities (Katz and Shapiro, 1986). Network externalities may also be a source of cost subadditivity.

prices above their efficient level can then be used to reduce prices to the target classes of customers, leaving the overall level of revenues produced from the menu of regulated prices equal to the total costs incurred by the firm. Richard Posner has referred to this phenomenon as *taxation by regulation* (Posner, 1971) and views government regulation of prices as one instrument of public finance [see also Hausman (1998)].

This phenomenon is also often referred to loosely as *cross-subsidization*. The notion is that one group of consumers subsidizes the provision of service to another group of customers by paying more than it costs to provide them with service while the other group pays less. However, when a firm has natural monopoly characteristics, an objective definition of "cross-subsidization" is not straightforward. When cost functions are subadditive and a natural monopoly is sustainable, break-even prices will generally be above the marginal cost of providing service to any individual or group of consumers. At least some consumers of some products produced by the natural monopoly must pay more than the incremental cost of serving them to satisfy a break-even constraint for the regulated firm. And, as we shall see below, efficient prices will generally vary from customer to customer when marginal cost-based prices do not yield sufficient revenues to cover total costs. Are consumer's paying prices that yield relatively high margins (difference between price and marginal cost) necessarily "subsidizing" consumers paying prices that yield lower margins?

More refined definitions of cross-subsidization have evolved that better reflect the attributes of subadditive cost functions [Sharkey (1982), Faulhaber (1975)]. A price configuration does not involve cross-subsidies (it is "subsidy free") if:

(a) All consumers pay at least the average incremental costs of providing them with service and

(b) No consumers or groups of consumers pay more than the "stand-alone costs" of providing them with service. Stand-alone costs refer to the costs of supplying only one or more groups of consumers that are a subset of the entire population of consumers that demand service at the prices at issue.

If these conditions prevail, consumers who are charged relatively high prices may be no worse off as a consequence of other consumers being charged lower prices and may be made better off than if the latter consumers purchased less (or nothing) from the firm if the prices they are being charged were to increase. This the case because if the contribution to meeting the firm's budget constraint made by the consumers being charged the lower prices is greater than or equal to zero then the remaining consumers will have to pay a smaller fraction of the firm's total costs and are better off than if they had to support the costs of the enterprise on a *stand-alone* basis.

Moreover, if subsidy free prices exist the natural monopoly will also be sustainable [Baumol, Bailey, and Willig (1977), Baumol, Panzar, and Willig (1982)]. On the other hand, if the government endeavors to engage in taxation by regulation in ways that involve setting prices that are not subsidy free, the resulting configuration may not be sustainable. In this case, restrictions on entry—legal monopoly—will be necessary to keep entrants from *cream skimming* the high margin customers away from the incumbent when the stand-alone costs make it profitable to do so [Laffont and Tirole (1990b)].

So, for example, when the U.S. federal government implemented policies in the 1920s to keep regulated local telephone service charges low in order to encourage universal service, subsidize customers in rural areas, etc., it simultaneously kept long-distance prices high to generate enough net revenues from long distance service to cover the costs of the local telephone network that were in excess of local service revenues [Palmer (1992), Crandall and Hausman (2000), Joskow and Noll (1999)]. This created potential opportunities for firms inefficiently to enter the market to supply some of the high-margin long-distance service (the prices were therefore greater than the stand alone costs), potentially undermining the government's ability to utilize taxation by regulation to implement the universal service and income distribution goals. When the costs of creating a competing long distance network were very high, this price structure was sustainable. However, as the costs of long distance telecommunications facilities fell, it became profitable, though not necessarily efficient, for competing entrants to supply, a subset of long distance services: the price structure was no longer sustainable.

Price Discrimination: In the single product case price discrimination involves a firm charging different prices for identical products to different consumers. The discrimination may involve distinguishing between different types of consumers (e.g. residential and commercial customers) and charging different per unit (linear) prices to each group for the same quantities purchased (third-degree price discrimination) or prices may vary depending on the quantities purchased by individual consumers (second-degree price discrimination). In a multiproduct context, price discrimination also encompasses situations where prices are set to yield different "margins" between price and marginal/average incremental cost for different products or groups of products (a form of third-degree price discrimination). The welfare/efficiency consequences of price discrimination by a monopoly in comparison to simple uniform monopoly pricing are ambiguous [Schmalensee (1981)]. Price discrimination could increase or reduce efficiency compared to uniform price-cost margins, depending on the shapes of the underlying demands for the services as well as attributes of the firm's cost function. In a regulated monopoly context, when firms are subject to a breakeven constraint, price discrimination of various kinds can reduce the efficiency losses associated with departures from marginal cost pricing. We will explore these issues presently.

Whatever the efficiency implications of price discrimination, it is important to recognize that real or imagined price discrimination by unregulated monopolies played an important political role in stimulating the introduction of price regulation of "natural monopolies" in the United States. The creation of the Interstate Commerce Commission in 1887 to supervise rail freight rates was heavily influenced by arguments made by shippers served by one railroad that they were being charged much higher prices per mile shipped for similar commodities by the same railroad than where shippers served by competing railroads or with transport alternatives that were close substitutes (e.g. barges) [Kolko (1965), Mullin (2000), Prager (1989a), Gilligan, Marshall, and Weingast (1990)]. Many regulatory statutes passed in the U.S. in the last century have (or had) text saying something like "rates shall be just, reasonable, *and not unduly discriminator*" [Bonbright (1961, p. 22), Clark (1911)]. The development of regulation in the

U.S. has been heavily influenced by the perceived inequities of charging different consumers different prices for what appear to be the same products. When combined with monopoly or very limited competition it has been both a source of political pressure to introduce price regulation and has led to legal and policy constraints on the nature of the price structures that regulatory agencies have at their disposal.

Political economy considerations: By this point it should be obvious that the decision to introduce price and entry regulation, as well as the behavior and performance of regulatory agencies, reflects a broader set of considerations than simply a public interest goal of mitigating the distortions created by unregulated markets with natural monopoly characteristics. Price and entry regulation can and does convey benefits on some groups and impose costs on other groups compared to alternatives, whether these alternatives are no price and entry regulation or alternative mechanisms for implementing price and entry regulation. The potential effects of price and entry regulation on the welfare of different interest groups—different groups of consumers, different groups of suppliers, environmental and other "public interest" groups—has played a significant role in where, when and how price and entry regulation are introduced, when and how regulatory mechanisms are changed, and when and how price and entry regulation may be removed. The nature and magnitude of alternative configurations of price and entry regulation on different interest groups, the costs and benefits these groups face to organize to influence regulatory laws and the behavior of regulatory agencies, and how these groups can use the institutions of government (legislature, executive, judicial) to create regulatory (or deregulatory) laws and influence regulatory behavior and outcomes is a very complex subject. The extensive relevant literature has been reviewed elsewhere (e.g. Noll, 1989) and much of it is covered as well in Chapter 22 (McNollgast) of this handbook. It is not my intention to review it again here. However, there are a number of general lessons learned from this literature that are worth noting as background for the rest of the material in this chapter.

For many years students were taught that regulation had been introduced to respond to natural monopoly problems—a "public interest" view of the introduction of price and entry regulation [Stigler (1971), Posner (1974)]. This view confused the *normative* market failures case for why it might be desirable to introduce price and entry regulation to achieve public interest goals with the *positive* question of why price and entry regulation was actually introduced in a particular industry at a particular time. One cannot and should not assume that because an industry is subject to price and entry regulation it is necessarily a "natural monopoly" in any meaningful sense. The introduction of price and entry regulation and the nature of the regulatory mechanisms used to implement it reflect political considerations that are the outcome of interest group politics [Peltzman (1989)]. There are many industries that have been subject to price and entry regulation (e.g. trucking, oil and natural gas production, various agricultural commodities) where there is no evidence of natural monopoly characteristics or the associated economic performance problems. Because regulation typically involves regulation of both prices and entry, it can be and has been used in some cases to keep prices high rather than low and to restrict competition where it would otherwise lead to lower prices, lower costs, and

other efficiency benefits. Each situation must be judged on the merits based on relevant empirical analysis of firm and industry cost and demand characteristics as well as the effects of regulation on firm behavior and performance.

Whatever the rationale for introducing price and entry regulation, we should not assume that regulatory agencies can and will use the most effective mechanisms for achieving public interest goals that may be available to them. Political considerations driven by interest group politics not only play a role in the introduction of price and entry regulation, but in how it is implemented by regulatory authorities [Weingast and Moran (1983), Noll (1989)]. While policymakers frequently refer to "independent" regulatory agencies in the abstract, the reality is that no regulatory agency is completely independent of political influences. This political influence is articulated by who is appointed to lead regulatory authorities, by legislative oversight and budget control, by the election of commissioners in states with elected commissions, and by the resources that different interest groups can bring to the regulatory process itself [McCubbins (1985), McCubbins, Noll, and Weingast (1987), Joskow, Rose, and Wolfram (1996), Hadlock, Lee, and Parrino (2002)].

Even under the best of circumstances, regulatory institutions can respond effectively to the goals established for them only imperfectly. Regulation leads to direct costs incurred by the agency and those groups who are involved with the regulatory process as well as indirect costs associated with distortions in regulated firm prices, costs, profits, etc., that may result from poorly designed or implemented regulatory mechanisms. The direct costs are relatively small. The indirect costs are potentially very large.

Firms may seek to enter an industry subject to price and entry regulation even if entry is inefficient. This result may flow from political constraints that influence the level and structure of regulated prices and make entry look profitable even though it is inefficient because the regulated price signals are inefficient. Distinguishing between efficient entry requests (e.g. due to technological change, new products, excessive costs of the regulated incumbent) and inefficient entry (e.g. responding to a price structure that reflects significant cross-subsidies) is a significant challenge that requires industry-specific assessments of the presence of natural monopoly characteristics and the distortions that may be caused by inefficient regulation.

3.3. Regulatory goals

Since the focus of this essay is on the economic efficiency rationales for price and entry regulation, the regulatory goals that will guide the design of effective regulatory mechanisms and institutions and against which the performance of regulatory institutions will be evaluated should reflect the same efficiency considerations. In what follows I will focus on the following regulatory goals:

Efficient pricing of goods and services: Regulated prices should provide consumers with efficient price signals to guide their consumption decisions. Ideally, prices will equal the relevant marginal or incremental costs. However, firm-viability and potentially

other constraints will necessarily lead to departures from first-best prices. Accordingly, second-best pricing given these constraints will be the goal on the pricing front.

Efficient production costs: The natural monopoly rationale for restricting entry to a single firm is to make it possible for the firm to exploit all economies of scale and economies of scope that are made feasible by the underlying technology, taking into account the organizational and related transactions costs associated with firms of different horizontal and vertical scales. Textbook presentations of natural monopoly regulation typically take the firm's cost function as given and focus on specification of optimal prices given the firm's costs and break-even constraint. However, by controlling a regulated firm's prices and profits and eliminating the threat of competitive entry, we may simultaneously sharply curtail the incentives that lead competing firms to seek to minimize costs from both static and dynamic perspectives. Moreover, regulation may significantly reduce the efficiency incentives that are potentially created by the market for corporate control by imposing lengthy regulatory review requirements and capturing the bulk of any cost savings resulting from mergers and acquisitions for consumers through lower regulated prices. Regulators need to be focused on creating substitute incentive mechanisms to induce regulated firms to minimize costs by adjusting inputs to reflect the relative input prices, to exert the optimal amounts of managerial effort to control costs, to constrain costly managerial emoluments and other sources of X-inefficiency, and to adopt new process innovations in a timely and efficient manner.

Efficient levels of output and investment (firm participation and firm-viability constraints): The regulated firm should supply the quantities of services demanded by consumers and make the investments in facilities necessary to do so in a timely and efficient manner. If private firms are to be induced to supply efficiently they must perceive that it is privately profitable to do so. Accordingly, regulatory mechanisms need to respect the constraint that private firms will only invest if they expect the investment to be profitable ex ante and will only continue to produce if they can cover their avoidable costs ex post.

Efficient levels of service quality and product variety: Products may be provided with varying levels of service quality and reliability. Different levels of service quality and reliability carry with them different costs. Consumer valuations of service quality and reliability may vary widely as well. Regulators should be concerned that the levels of service quality and reliability, and the variety of quality and reliability options available to consumers reflect consumer valuations and any costs associated with providing consumers with a variety of levels of quality and reliability from which they can choose. Physical attributes of the networks which characterize industries that have often been subject to price and entry regulation may limit the array of product qualities that can be offered economically to consumers. For example, on a typical electric distribution network, individual consumers cannot be offered different levels of network reliability because the physical control of the distribution network is at the "neighborhood" rather than the individual levels (Joskow and Tirole, 2005).

Monopoly profit and rent extraction considerations: While simple models of social welfare (e.g. the sum of consumers' plus producers' surplus) are agnostic about the dis-

tribution of surplus between consumers and producers, it is clear that regulatory policies are not. In addition to the efficiency distortions caused by monopoly pricing, extracting the excess profits associated with monopoly profits for the benefit of consumers is also an important goal of most regulatory laws. It is the flip side of the firm viability constraint. The regulated firm's profits must be "high enough" to induce it to supply efficiently, but "no higher" than is necessary to do so. This goal can be rationalized in a number of ways. I prefer to view it as an articulation of a social welfare function that weights consumers' surplus more than producers' surplus subject to a firm viability or breakeven constraint. Alternatively, one might rationalize it as reflecting a concern that some or all of the monopoly profits will be transformed into wasteful "rent seeking" expenditures by the regulated firm to enable it to retain its monopoly position.

Distributional Goals: To the extent that other income distribution goals (e.g. universal service goals) are assigned to the regulated firm, price and quantity mechanisms should be adopted to achieve these goals at minimum cost.

Ultimately, sound public policy must ask whether the potential improvements in performance along the various performance dimensions discussed above relative to unregulated market outcomes—depicted in a simple fashion in Figures 5A and 5B—are likely to be greater than the direct and indirect costs of government regulatory mechanisms. Accordingly, sensible decisions about whether and how to regulate should consider both the costs of imperfect markets and the costs of imperfect regulation.

4. Historical and legal foundations for price regulation

Government regulation of prices can be traced back at least to the period of the Roman Empire when the emperor established maximum prices for roughly 800 items. These actions found support in the doctrine of "the just price" developed by Church authorities [Phillips (1993 p. 90)]. During the Middle Ages, craft guilds developed which licensed and controlled the individuals who could work in specific occupations. Because these guilds had monopoly control over who could work in particular crafts they were regulated. "The obligation of the guilds was to provide service to anyone who wanted it at reasonable prices. The various crafts were known therefore as 'common carriers,' 'common innkeepers,' 'common tailors' and so forth. Since each craft had a monopoly of its trade, they were closely regulated" [Phillips (1993, p. 90)]. During the 16th century, the French government began to issue Royal charters to trading companies and plantations which gave them special privileges, including monopoly status, and in turn subjected them to government regulation [Phillips (1993, p. 90)]. These charters, analogous to modern franchises, have been rationalized as reflecting efforts by governments to induce private investment in activities that advanced various social goals [Glaeser (1927, p. 201)].

The antecedents of American legal concepts of "public interest" and "public utilities" that were the initial legal foundations for government price and entry regulation can be found in English Common law. "Under the common law, certain occupations or

callings were singled out and subjected to special rights and duties. These occupations became known as 'common callings,'. . . . A person engaged in a common employment had special obligations . . . , particularly the duty to provide, at reasonable prices, adequate services and facilities to all who wanted them" [Phillips (1993, p. 91)]. English common law regulations were carried over to the English colonies and during the Revolution several colonies regulated prices for many commodities and wages [Phillips (1993, pp. 91–92)]. However, after the American Revolution, government regulation of prices and entry faded away as the United States developed a free market philosophy that relied on competition and was hostile to government regulation of prices and entry [Phillips (1993, p. 92)]. Following the Civil War, and especially with the development of the railroads and the great merger wave of the late 1890s, policymakers and the courts began again to look favorably on price and entry regulation under certain circumstances. The Granger Movement of the 1870s focused on pressuring the states and then the federal government to regulate railroad freight rates. State regulation of railroads by special commissions began in the Midwestern states and then spread to the rest of the country [Phillips (1993, p. 93)]. The first federal economic regulatory agency, the Interstate Railroad Commission (ICC), was established in 1888 with limited authority to regulate the structure of interstate railroad rates. This authority was greatly expanded during the first two decades of the 20th century [Gilligan, Marshall, and Weingast (1989), Mullin (2000), Prager (1989a), Kolko (1965), Clark (1911)].

In the U.S., it was widely accepted as a legal matter that a state or municipality (with state authorization) could issue franchises or concessions to firms seeking to provide certain services using rights of way owned by the municipality and to negotiate the terms of the associated contracts with willing suppliers seeking to use such state and municipal rights of way [Hughes (1983), McDonald (1962)]. These firms proposed to use state or municipal property and the state could define what the associated terms and conditions of contracts to use that property would be. However, the notion that a municipal, state or the federal government could on its own initiative independently impose price regulations on otherwise unwilling private entities was a more hotly contested legal issue about which the Supreme Court's views have changed over time.[9]

Until the 1930s, the U.S. Supreme Court was generally fairly hostile to actions by state and federal authorities to restrict the ability of private enterprises to set prices freely without any restrictions imposed by government [Clemens (1950, pp. 12–37)] except under very special circumstances. Such actions were viewed as potentially violating Constitutional protections of private property rights, due process and contracts. On the one hand, the commerce clause (Article I, Section 8, Clause 3) gives the federal government the power to "regulate commerce with foreign nations and among the several states . . . ." On the other hand, the due process clause (Fifth Amendment) and the equal protection of the laws clause (Article Fourteen), and the obligation of contracts clause (Article I, Section 10) restricts the regulatory powers of the government

[9] The relevant Court decisions are discussed in Clemens (1950, pp. 49–54).

[Clemens (1950, pp. 45–48)]. The courts initially recognized some narrow exceptions to the general rule that the government could not regulate prices in light of the protections provided by the Fifth and Fourteenth Amendments; for example when there were emergencies that threatened public health and safety [Bonbright (1961, p. 6)]. And gradually over time the courts carved out additional exceptions "for certain types of business said to have been 'dedicated to a public use' or 'affected with the public interest,'. . ." [Bonbright (1961, p. 6)]. Railroads, municipal rail transit systems, local gas and electricity systems and other "public utilities" became covered by these exceptions.

One would not have to be very creative to come up with a long list of industries that are "affected with the public interest" and where investments had been "dedicated to a public use." And if such vague criteria were applied to define industries that could be subject to price and entry regulation, there would be almost no limit to the government's ability to regulate prices for reasons that go well beyond performance problems associated with natural monopoly characteristics. However, at least up until the 1930s, the courts had in mind a much less expansive notion of what constituted a "public utility" whose prices and other terms and conditions of service could be legitimately regulated by state or federal authorities (or municipal authorities by virtue of power delegated to them by their state government).[10] The two criteria where (a) the product had to be "important" or a "necessity" and (b) the production technology had natural monopoly characteristics [Bonbright (1961, p. 8)]. Clemens (1950, p. 25) argues that "[N]ecessity and monopoly are almost prerequisites of public utility status." One could read this as saying that the combination of relatively inelastic demand for a product that was highly valued by consumers and natural monopoly characteristics on the supply side leading to significant losses in social welfare are a necessary pre-condition for permitting government price and entry regulation. An alternative interpretation is that the "necessity" refers not so much to the product itself, but rather for the "necessity of price and entry regulation" to achieve acceptable price, output and service quality outcomes when industries had natural monopoly characteristics. In either case, until the 1930s, it is clear that the Supreme Court intended that the situations in which government price regulation would be constitutionally permissible were quite narrow.[11]

The conditions under which governments could regulate price, entry and other terms and conditions of service without violating constitutional protections were expanded during the 1930s.[12] Since the 1930s, federal and state governments have imposed

[10] The landmark case is *Munn v. Illinois 94 U.S. 113 (1877)* where the Illinois state legislature passed a law that required grain elevators and warehouses in Chicago to obtain licenses and to charge prices that did not exceed levels specified in the stature. The importance of the grain storage facilities to the grain shipping business in Chicago and that fact that the ownership of the facilities constituted a virtual monopoly were important factors in the Court's decision. See also *Budd v. New York, 143 U.S. 517 (1892)*.

[11] In a series of subsequent cases the Court made it clear that the conditions under which states could regulate prices were narrow. See *German Alliance Insurance Co. v. Lewis 233 U.S. 389 (1914)*, *Wolff Packing Co. v. Court of Industrial Relations, 262 U.S. 522 (1923), Williams v. Standard Oil Co. 278 U.S. 235 (1929)*.

[12] In *Nebbia vs. New York 291 U.S. 502 (1934)* the Supreme Court upheld a New York State law that created a milk control board that could set the maximum and minimum retail prices for milk sold in the State.

price regulation on a wide variety of industries that clearly do not meet the "necessity and natural monopoly" test discussed above—milk, petroleum and natural gas, taxis, apartment rents, insurance, etc.—without violating the Constitution. Nevertheless, the natural monopoly problem, the concept of the public utility developed in the late 19th and early 20th centuries, and the structure, rules and procedures governing state and federal regulatory commissions that are responsible for regulating industries that meet the traditional public utility criteria go hand in hand.

It should also be recognized that just because an industry can as a legal matter be subject to government price and entry regulation does not mean that the owners of the enterprises affected give up their Constitutional protections under the Fifth and Fourteenth amendments. The evolution of legal rules supporting the right of government to regulate prices and entry and impose various obligations on regulated monopolies were accompanied by a parallel set of legal rules that required government regulatory actions to adhere to these constitutional guarantees. This requirement in turn has implications for regulatory procedures and regulatory mechanisms. They must be consistent with the principle that private property cannot be taken by government action without just compensation. This interrelationship between the conditions under which government may regulate prices and the Constitutional protections that the associated rules and procedures must adhere to are very fundamental attributes of U.S. regulatory law and policy. In particular, they have important implications for the incentives regulated firms have to invest in facilities to expand supplies of services efficiently to satisfy the demand for these service whose prices are subject to government regulation [Sidak and Spulber (1997), Kolbe and Tye (1991)].

5. Alternative regulatory institutions

5.1. Overview

There are a variety of organizational arrangements through which prices, entry and other terms and conditions of service might be regulated by one or more government entities. Legislatures may enact statutes that establish licensing conditions, maximum and minimum prices and other terms and conditions of trade in certain goods and services. This was the approach that led to the Supreme Court's decision in *Munn v. Illinois* where prices were regulated by a statute passed by the Illinois legislature. Indeed, the first "public utilities" were created by legislative acts that granted franchises that specified maximum prices and/or profit rates and provide the first examples of rate of return regulation [Phillips (1993, p. 129)]. When changes in supply and demand conditions led to the need for price changes the legislature could, in principle, amend the statute to make these changes. This type of regulation by legislative act was both clumsy and politically inconvenient [McCubbins (1985), Fiorina (1982), McCubbins, Noll, and Weingast (1987), Hughes (1983)].

Governments can also use the terms of the contracts that they issue to firms which require authorization to use public streets and other rights of way to provide service by including in these "franchise contracts" terms and conditions specifying prices and how they can be adjusted over time [McDonald (1962), Hughes (1983)]. The sectors that are most often categorized as "public utilities" typically began life as local companies that received franchises from the individual municipal governments to whose streets and rights of way they required access to provide service. City councils and agencies negotiated and monitored the associated franchise contracts and were effectively the regulators of these franchisees. However, as contracts, the ability of the municipality to alter the terms and conditions of the franchise agreement without the consent of the franchisee was quite limited [National Civic Federation (1907)]. Most gas, electric, telephone, water and cable TV companies that provide local service and use municipal streets and rights of way still must have municipal franchises, but these franchises typically are little more than mechanisms to collect fees for the use of municipal property as state and federal laws have transferred most regulation of prices and entry to state and/or federal regulatory agencies. The strengths and weaknesses of municipal franchise contracts allocated through competitive bidding are discussed further below.

The "independent" regulatory commission eventually became the favored method for economic regulation in the U.S. at both the state and federal levels [Clemens (1950, Chapter 3), Kahn (1970, p. 10), Phillips (1993, Chapter 4)]. Independent regulatory commissions have been given the responsibility to set prices and other terms and conditions of service and to establish rules regarding the organization of public utilities and their finances. This approach creates a separate board or commission, typically with a staff of engineers, accountants, finance specialists and economists, and gives it the responsibility to regulate prices and other terms and conditions of services provided by the companies that have been given charters, franchises, licenses or other permissions to provide a specific service "in the public interest." The responsibilities typically extend to the corporate forms of the regulated firms, their finances, the lines of business they may enter and their relationships with affiliates. Regulatory agencies are also given various authorities to establish accounting standards and access to the books, records and other information relevant for fulfilling their regulatory responsibilities, to approve investment plans and financings, and to establish service quality standards. Regulated firms are required to file their schedules of prices or "tariffs" with the regulatory commission and all eligible consumers must be served at these prices. Changes in price schedules or tariffs must be approved by the regulatory agency. We will discuss commission regulation in more detail presently.

A final approach to "the natural monopoly problem" has been to rely on public ownership. Under a public ownership model, the government owns the entity providing the services, is responsible for its governance, including the choice of senior management, and sets prices and other terms and conditions. Public ownership may be affected through the creation of a bureau or department of the municipal or state government that provides the services by creating a separate corporate entity organized as a public benefit corporation with the government as its sole owner. In the latter case, the state-

owned company will typically then be "regulated" by a municipal or state department which will approve prices, budgets and external financing decisions. In the U.S. there has been only limited use of public ownership as a response to the natural monopoly problem. The primary exceptions are electricity where roughly 20% of the electricity distributed or generated in the U.S. is accounted for by municipal or state public utility districts (e.g. Los Angeles Department of Water and Power) or federal power marketing agencies (e.g. TVA) and the public distribution of water where state-owned enterprises play a much larger role. Natural gas transmission and distribution, telephone and related communications, and cable television networks are almost entirely private in the U.S. This has not been the case in many other countries in Europe, Latin America, and Asia where state-owned enterprises dominated these sectors until the last decade or so.

There is a long literature on public enterprise and privatization that covers both traditional natural monopoly industries and other sectors where public enterprise spread [e.g. Vickers and Yarrow (1991), Armstrong, Cowan, and Vickers (1994), Megginson and Netter (2001)]. The literature covers price regulation as well as many other topics related to the performance of state-owned utilities. I will not cover the literature on public enterprise or privatization in this essay.

5.2. Franchise contracts and competition for the market

When the supply of a good or service has natural monopoly characteristics "*competition within the market*" will lead to a variety of performance failures as discussed above. While "competition within the market" may lead to these types of inefficiencies, Harold Demsetz (1968) suggested that "*competition for the market*" could rely on competitive market processes, rather than regulation, to select the most efficient supplier and (perhaps) a second-best break-even price structure. The essence of the Demsetz proposal is to use competitive bidding to award monopoly franchise contracts between a government entity and the supplier, effectively to try to replicate the outcomes that would emerge in a perfectly contestable market. The franchise could go to the bidder that offers to supply the service at the lowest price (for a single product monopoly) or the most efficient (second-best) price structure. The franchising authority can add additional normative criteria to the bidding process. Whatever the criteria, the idea is that the power of competitive markets can still be harnessed at the ex ante franchise contract execution stage even though ex post there is only a single firm in the market. Ex post, regulation effectively takes place via the terms and conditions of the contract which are, in turn, determined by competitive bidding ex ante.

For a franchise bidding system to work well there must, at the very least, be an adequate number of ex ante competitors and they must act independently (no collusion). In this regard, one cannot presume that ex ante competition will be perfect competition due to differences among firms in access to productive resources, information and other attributes. Competition among two or more potential suppliers may still be imperfect. The efficiency and rent distribution attributes of the auction will also depend on the specific auction rules used to select the winner and the distribution of information about

costs and demand among the bidders [Klemperer (2002)]. And, of course, the selection criteria used to choose the winner may be influenced by the same kinds of political economy considerations noted above.

More recent theoretical developments in auction theory and incentive theory lead to a natural bridge between franchise bidding mechanisms and incentive regulation mechanisms, a subject that we will explore in more detail below. Laffont and Tirole (1993, Chapter 7) show that the primary benefit of the optimal auction compared to the outcome of optimal regulation with asymmetric information in this context is that competition lowers the prices (rents) at which the product is supplied. In addition, as is the case for optimal regulation with asymmetric information (more below) the franchise contract resulting from an optimal auction is not necessarily a fixed price contract but rather a contract that is partially contingent on realized (audited) costs. The latter result depends on the number of competitors. As the number of competitors grows, the result of the optimal auction converges to a fixed price contract granted to the lowest cost supplier, who exerts optimal effort and leaves no excess profits on the table [Laffont and Tirole (1993, p. 318)]. Armstrong and Sappington (2003a, 2003b) show (proposition 14) that the optimal franchise auction in a static setting with independent costs has the following features: (a) The franchise is awarded to the firm with the lowest costs; (b) A high-cost firm makes zero rent; (b) the rent enjoyed by a low-cost firm that wins the contest decreases with the number of bidders; (c) the total expected rent of the industry decreases with the number of bidders; (d) the prices that the winning firm charges do not depend on the number of bidders and are the optimal prices in the single-firm setting. That is, in theory, with a properly designed auction and a large number of competitors, the outcome converges to the one suggested by Demsetz.

The Demsetz proposal and the related theoretical research seems to be most relevant to natural monopoly services like community trash collection or ambulance services where assets are highly mobile from one community to another (i.e. minimal location-specific sunk costs), the attributes of the service can be easily defined and suppliers are willing to offer services based on a series of repeated short-term contracts mediated through repeated use of competitive bidding. That is, it is most relevant to market environments that are closer to being contestable. It ignores the implications of significant long-lived sunk costs, asymmetric information between the incumbent and non-incumbent bidders, strategic actions changing input prices, changing technology, product quality and variety issues, and incomplete contracts.

As Williamson (1976) has observed, these attributes of the classical real-world natural monopoly industries make once-and-for-all long-term contracts inefficient and not credible. One alternative is to rely on repeated fixed-price short-term contracts. But in the presence of sunk costs and asymmetric information, repeated fixed-price auctions for short-term franchise contracts lead to what are now well known ex ante investment and ex post adaptation problems associated with incomplete complex long-term contracts and opportunistic behavior by one or both parties to the franchise agreement [Williamson (1985)]. Where sunk costs are an important component of total costs, repeated auctions for short-term fixed-price contracts are unlikely to support efficient

investments in long-lived assets and efficient prices for the associated services. This in turn leads to the need for an institutional mechanism to adjudicate contractual disputes. This could be a court or a government agency created by the government to monitor contractual performance, to negotiate adjustments to the franchise contract over time, and to resolve disputes with the franchisee. Goldberg (1976) argues that in these circumstances the franchising agency effectively becomes a regulatory agency that deals with a single incumbent to enforce and adjust the terms of its contract. Joskow and Schmalensee (1986) suggest that government regulation is productively viewed from this contract enforcement and adjustment perspective. For the kinds of industries that are typically thought of a regulated natural monopolies, the complications identified by Williamson and Goldberg are likely to be important.

5.3. *Franchise contracts in practice*

In fact, franchise bidding for natural monopoly services is not a new idea but a rather old idea with which there is extensive historical experience. Many sectors with (arguably) natural monopoly characteristics in the U.S., Europe, Canada and other countries that started their lives during the late 19th and early part of the 20th century, started off life as suppliers under (typically) municipal franchise contracts that were issued through some type of competitive bidding process [Phillips (1993, pp. 130–131), Hughes (1983, Chapter 9)]. The franchise contracts were often exclusive to a geographic area, but in many cases there were multiple legal (and illegal) franchisees that competed with one another [Jarrell (1978), McDonald (1962)] in the same geographic area.

In many cases the initial long-term contracts between municipalities and suppliers broke down over time as economic conditions changed dramatically and the contracts did not contain enforceable conditions to adapt prices, services, and quality to changing conditions, including competitive conditions, and expectations changed [Hughes (1983), McDonald (1962)]. The historical evolution is consistent with the considerations raised by Williamson and Goldberg.[13] Municipal corruption also played a role, as did wars of attrition when there were competing franchises and adverse public reaction to multiple companies stringing telephone and electric wires on poles and across city streets and disruptions caused by multiple suppliers opening up streets to bury pipes and wires [McDonald (1962), National Civic Federation (1907)]. Utilities with municipal franchises began to expand to include many municipalities, unincorporated areas of the state and to cross state lines [Hughes (1983)]. These expansions reflect further exploitation of economies of scale, growing demand for the services as costs and prices fell due to economies of scale, economies of density, technological change, and extensive merger and acquisition activity. Municipalities faced increasing difficulties in regulat-

[13] Though municipal franchise contracts for cable TV service appear not to have had the significant performance problems identified by Williamson (1976). See [Prager (1989b), Zupan (1989a, 1989b)] while federal efforts to regulate cable TV prices have encountered significant challenges [Crawford (2000)].

ing large corporate entities that provided service in many municipalities from common facilities [National Civic Federation (1907), Hughes (1983)]. By around the turn of the 20th century, problems associated with the governance of municipal franchise contracts and their regulation led progressive economists like John R. Commons to favor replacing municipal franchise contracting and municipal regulation with state regulation by independent expert regulatory agencies that could be better insulated from interest group politics generally [McDonald (1962)] and have access to better information and relevant expertise to more effectively determine reasonable prices, costs, service quality benchmarks, etc. [Prager (1990)].

5.4. Independent "expert" regulatory commission

5.4.1. Historical evolution

Prior to the Civil War, several states established special commissions or boards to collect information and provide advice to state legislatures regarding railroads in their states. These commissions were advisory and did not have authority to set prices or other terms and conditions of service [Phillips (1993, p. 132), Clemens (1950, p. 38)]. The earliest state commissions with power over railroad rates were established by "the Granger laws" in several Midwestern states in the 1870s.[14] These commissions had various powers to set maximum rates, limit price discrimination and to review mergers of competing railroads. By 1887, twenty-five states had created commissions with various powers over railroad rates and mergers and to assist state legislatures in the oversight of the railroads [Phillips (1993, p. 132)]. In 1887, the federal government created the Interstate Commerce Commission (ICC) to oversee and potentially regulate certain aspects of interstate railroad freight rates, though the ICC initially had limited authority and shared responsibilities with the states. [Clemens (1950, p. 40)]. The ICC's regulatory authority over railroads was expanded considerable during the first two decades of the 20th century [Mullin (2000), Prager (1989a), Kolko (1965), Gilligan, Marshall, and Weingast (1989)] and was extended to telephone and telegraph (until these responsibilities were taken over by the Federal Communications Commission in 1934) and to interstate trucking in 1935 and domestic water carriers in 1940.

State commission regulation of other "public utility" sectors spread much more slowly as they continued to be subject to local regulation through the franchise contract and renewal process. Massachusetts established the Board of Gas Commissioners in 1885 which had power to set maximum prices and to order improvements in service [Clemens (1950, p. 41)]. Its power was extended to electric light companies two years later. However, the transfer of regulatory power from local governments to state commissions began in earnest in 1907 when New York and Wisconsin created state commissions with jurisdiction over gas distribution, electric power, water, telephone and

[14] Earlier state railroad commissions had fact finding and advisory roles.

telegraph service prices. By 1920 more than two-thirds of the states had created state public utility commissions [Stigler and Friedland (1962), Phillips (1993, p. 133), Jarrell (1978)], a very rapid rate of diffusion of a new form of government regulatory authority, and today all states have such commissions. The authority of the early commissions over the firms they regulated was much less extensive than it is today, and their legal authorities, organization and staffing evolved considerably over time [Clemens (1950, p. 42)].

Federal commission regulation expanded greatly during the 1930s with the Communications Act of 1934 and the associated creation of the Federal Communications Commission (FCC) with authority over the radio spectrum and interstate telephone and telegraph rates, the expansion of the powers of the Federal Power Commission (FPC, now the Federal Energy Regulatory Commission or FERC) by the Federal Power Act of 1935 to include interstate sales of wholesale electric power and transmission service, and interstate transportation and sales of natural gas to gas distribution companies and large industrial consumers, the passage of the Public Utility Holding Company Act of 1935 which gave the new Securities and Exchange Commission (SEC) regulatory responsibilities for interstate gas and electric public utility holding companies, the expansion of the ICC's authority to regulate rates for interstate freight transportation by trucks in 1935, and the creation of the Civil Aeronautics Board (CAB) to regulated interstate air fares in 1938.

It is hard to argue that the growth of federal regulation at this time reflected a renewed concern about performance problems associated with "natural monopolies." The expansion of federal authority reflected a number of factors: the general expansion of federal authority over the economy during the Great Depression and in particular the popularization of views that "destructive competition" and other types of market failure were a major source of the country's economic problems; efforts by a number of industries to use federal regulatory authority to insulate themselves from competition, especially in the transportation areas (railroads, trucks, airlines); as well as the growth of *interstate* gas pipelines, electric power networks, and telephone networks that could not be regulated effectively by individual states.

5.4.2. Evolution of regulatory practice

It became clear to students of regulation and policymakers that effective regulation by the government required expertise in areas such as engineering, accounting, finance, and economics. Government regulators also needed information about the regulated firms' costs, demand, investment, management, financing, productivity, reliability and safety attributes to regulate effectively. Powerful interest groups were affected by decisions about prices, service quality, service extensions, investment, etc. and had incentives to exert any available political and other influence on regulators. The regulated firms and larger industrial and commercial consumer groups were likely to be well organized to exert this kind of influence, but residential and small commercial consumers were likely to find it costly and difficult to organize to represent their interests effectively

through the same political processes. At the same time, the industries subject to regulation were capital intensive, incurred significant sunk costs associated with investments in long-lived and immobile assets and were potentially subject to regulatory hold-ups. The threat of such hold-ups would reduce or destroy incentives to make adequate investments to balance supply and demand efficiently.

The chosen organizational solution to this web of challenges for price and entry regulation in the U.S. during most of the 20th century was the independent regulatory commission [Phillips (1993)]. The commission would have a quasi-judicial structure that applied transparent administrative procedures to establish prices, review investment and financing plans, and to specify and monitor other terms and conditions of service. At the top of the commission would be three to seven public utility "commissioners" who were responsible for voting "yes" or "no" on all major regulatory actions. In most jurisdictions the commissioners are appointed by the executive (governor or the President) and approved by the legislature. They are often appointed for fixed terms and sometimes for terms that are coterminous with the term of the governor. At the federal level and in a number of states no more than a simple majority of the commissioners can be registered in the same political party. In about a dozen states the public utility commissioners are elected by popular vote [Joskow, Rose, and Wolfram (1996)].

Underneath the commissioners is a commission staff which consists of professionals with training in engineering, accounting, finance, and economics and often a set of administrative law judges who are responsible for conducting public hearings and making recommendations to the commissioners. The composition and size of commission staffs varies widely across the states. Commissions adopt uniform systems of accounts and require regulated firms to report extensive financial and operating data to the commission on a continuing basis consistent with these accounting and reporting protocols. Each commission adopts a set of administrative procedures that specifies how the commission will go about making decisions. These procedures are designed to give all interest groups the opportunity to participate in hearings and other administrative procedures, to make information and decisions transparent, and generally to provide due process to all affected interest groups. These procedures include rules governing private meetings between groups that may be affected by regulatory commission proceedings (so-called ex parte rules), rules about the number of commissioners who may meet together privately, and various "sunshine" and "open meeting" rules that require commissioners to make their deliberations public. Regulatory decisions must be based on a reasonable assessment of the relevant facts in light of the agency's statutory responsibilities. Prices must be "just, reasonable and not unduly discriminatory," insuring the consumers are charged no more than necessary to give the regulated firms a reasonable opportunity to recover efficiently incurred costs, including a fair rate of return of and on their investments [*Federal Power Commission v. Hope Natural Gas* 320 U.S. 591 (1944)].

In light of the evolution of constitutional principles governing economic regulation, providing adequate protection for the investments made by regulated firms in assets dedicated to public use plays an important role in the regulatory process and has important implications for attracting investments to regulated sectors. Not surprisingly, these

administrative procedures have evolved considerably over time, with the general trend being to provide more opportunities for interest group participation, more transparency, and fewer opportunities for closed-door influence peddling [Chapter 22, McNollgast (in this handbook)]. Regulatory decisions may be appealed to state or federal appeals courts.

Of course this idealized vision of the independent regulatory commission making reasoned decisions based on an expert assessment of all of the relevant information available often does not match the reality very well. No regulatory agency can be completely independent of political influences. Commissioners and senior staff members are political appointments and while they cannot be fired without just cause they are also unlikely to be appointed or reappointed if their general policy views are not acceptable to the executive or the public (where commissioners are elected). Regulatory agencies are also subject to legislative oversight and their behavior may be constrained through the legislative budgetary process [Weingast and Moran (1983)]. Regulators may have career ambitions that may lead them to curry favor with one interest group or another [Laffont and Tirole (1993, Chapter 16)]. Staffs may be underfunded and weak. Reporting requirements may not be adequate and/or the staff may have inadequate resources properly to analyze data and evaluate reports submitted by the parties to regulatory proceedings. Ex parte rules may be difficult to enforce. The administrative process may be too slow and cumbersome to allow actions to be taken in a timely way. Under extreme economic conditions, regulatory principles that evolved to protect investments in regulated enterprises from regulatory expropriation come under great stress [Joskow (1974), Kolbe and Tye (1991), Sidak and Spulber (1997)]. On the other hand, both the executive branch and the legislature may find it politically attractive to devolve complicated and controversial decisions to agencies that are both expert and arguably independent [McCubbins, Noll, and Weingast (1987)].

All things considered, the performance of the U.S. institution of the independent expert regulatory agency turns on several attributes: a reasonable level of independence of the commission and its staff from the legislative and executive branches supported by detailed due process and transparency requirements included in enforceable administrative procedures, the power to specify uniform accounting rules and to require regulated firms to make their books and operating records available to the commission, a professional staff with the expertise and resources necessary to analyze and evaluate this information, constitutional protections against unreasonable "takings" of investments made by regulated firms, and the opportunity to appeal regulatory decisions to an independent judiciary.

6. Price regulation by a fully informed regulator

Much of the traditional theoretical literature on price regulation of natural monopolies assumes that there is a legal monopoly providing one or more services and a regulatory agency whose job it is to set prices. The regulated firm has natural monopoly characteristics (generally economies of scale in its single product and multiproduct variations)

and the firm is assumed to minimize costs given technology, input prices and output levels (i.e., no X-inefficiency). That is, the firm's cost function is taken as given and issues of production inefficiency are ignored. In the presence of scale economies, marginal cost pricing will typically not yield sufficient revenues to cover total cost. Fully efficient pricing is typically not feasible for a private firm that must meet a break-even constraint in the presence of economies of scale (even with government transfers since government taxation required to raise revenues to transfer to the regulated firm creates its own inefficiencies). Accordingly, the traditional literature on price regulation of natural/legal monopolies focused on normative issues related to the development of second-best pricing rules for the regulated firm given a break-even constraint (or given a cost of government subsidies that ultimately rely on a tax system that also creates inefficiencies). A secondary focus of the literature has been on pricing of services like electricity which are non-storable, have widely varying temporal demand, have high capital intensities and capital must be invested to provide enough capacity to meet the peak demand—the so-called peak-load or variable-load pricing (PLP) problem.

The traditional literature on second-best pricing for natural monopolies assumes that the regulator is fully informed about the regulated firm's costs and knows as much about the attributes of the demand for the services that the firm supplies as does the regulated firm. The regulator's goal is to identify and implement normative pricing rules that maximize total surplus given a budget constraint faced by the regulated firm. Neither the regulated firm nor the regulator acts strategically. This literature represents a normative theory of what regulators *should* do if they are fully informed. It is not a positive theory of what regulators or regulated firms actually do in practice. (Although there is a sense of "normative as positive theory of regulation" in much of the pre-1970s literature on price regulation.)

6.1. Optimal linear prices: Ramsey-Boiteux pricing

In order for the firm with increasing returns to break-even it appears that the prices the firm charges for the services it provides will have to exceed marginal cost. One way to proceed in the single product context is simply to set a single price for each unit of the product equal to its average cost (p_{AC}). Then the expenditures made by each consumer i will be equal to $E_i = p_{\mathrm{AC}} q_i$. In this case p_{AC} is a uniform linear price schedule since the firm charges the same price for each unit consumed and each consumer's expenditures on the product varies proportionately with the output she consumes. In the multiproduct context, we could charge a uniform price per unit for each product supplied by the firm that departs from its marginal or average incremental cost by a common percentage mark-up consistent with meeting the regulated firm's budget constraint. Again the prices charged for each product are linear in the sense that the unit price for each product is a constant and yields a linear expenditure schedule for consumers of each product.

The first question to address is whether, within the class of linear prices, we can do better than charging a uniform price per unit supplied that embodies an equal mark-up over marginal cost to all consumers for all products sold by the regulated firm?

Alternatively, can we do better by engaging in third degree price discrimination, in the case of a single product firm, by charging different unit prices to different types of consumers (e.g. residential and industrial and assuming that resale is restricted) or in the case of multiproduct firms by charging a constant unit price for each product but where each unit price embodies a different markup over its incremental cost?

Following Laffont and Tirole (2000, p. 64), the regulated firm produces n products whose quantities supplied are represented by the vector $\mathbf{q} = (q_1, \ldots, q_n)$. Assume that the demand functions for the price vector $\mathbf{p} = (p_1, \ldots, p_n)$ are $q_k = D_k(p_1, \ldots, p_n)$. The firm's total revenue function is then $R(\mathbf{q}) = \sum_{(i=1,n)} p_k q_k$. Let the firm's total cost function be $C(\mathbf{q}) = C(q_1, \ldots, q_n)$ and denote the marginal cost for each product k as $C_k(q_1, \ldots, q_n)$.

Let $S(q)$ denote the gross surplus for output vector $\mathbf{q}$ with $\frac{\partial S}{\partial q_k} = p_k$. The Ramsey-Boiteux pricing problem [Ramsey, 1927, Boiteux, 1971 (1956)] is then to find the vector of constant unit (linear) prices for the n products that maximizes net social surplus subject to the regulated firm's break-even or balanced budget constraint:

$$\max_q \big\{S(\mathbf{q}) - C(\mathbf{q})\big\} \quad \text{subject to} \tag{1}$$

$$R(q) - C(q) \geq 0 \tag{2}$$

or equivalently, maximizing the firm's profit subject to achieving the Ramsey-Boiteux level of net social surplus:

$$\max_q \big\{R(\mathbf{q}) - C(\mathbf{q})\big\} \quad \text{subject to} \tag{3}$$

$$S(q) - C(q) \geq S(\mathbf{q}^*) - C(\mathbf{q}^*) \tag{4}$$

Where $\mathbf{q}^*$ represent the Ramsey-Boiteux levels of output.

Let $1/\lambda$ represent the shadow price of the constraint in the second formulation above. Then the first order condition for the maximization of (3) subject to (4) for each q_k is given by:

$$\lambda\left(p_k - c_k + \sum_{j=1}^{n} \frac{\partial p_j}{\partial q_k} q_j\right) + p_k - c_k = 0 \tag{5}$$

When the demands for the products produced by the regulated firm are *independent* this reduces to:

$$\frac{p_k - c_k}{p_k} = \frac{\lambda}{1+\lambda}\frac{1}{\eta_k} \tag{6}$$

for all products $k = 1, \ldots, n$ and where η_k is the own-price elasticity of demand for product k. This is often referred to as the *inverse elasticity rule* [Baumol and Bradford (1970)]. Prices are set so that the difference between a product's price and its marginal cost varies inversely with the elasticity of demand for the product. The margin is higher for products that have less elastic demands than for products that have more elastic demand (at the equilibrium prices).

When the products produced by the regulated firm are not independent—they are substitutes or complements—the own-price elasticities in (6) must be replaced with "super-elasticities" that reflect the cross-price effects as well as own-price effects. If the products are substitutes, the Ramsey-Boiteux prices are higher than would be implied by ignoring the substitution effect (the relevant superelasticity is less elastic than the own-price elasticity of good k) and vice versa.

Note that Ramsey-Boiteux prices involve third-degree price discrimination that results in a set of prices that lie between marginal cost pricing and the prices that would be set by a pure monopoly engaging in third-degree price discrimination. For example, rather than being different products, assume that q_1 and q_2 are the same product consumed by two groups of consumers who have different demand elasticities (e.g. residential and industrial consumers) and that resale can be blocked, eliminating the opportunity to arbitrage away differences in prices charged to the two groups of consumers. Then the price will be higher for the group with the less elastic demand despite the fact that the product and the associated marginal cost of producing it are the same. Note as well that the structure, though not the level, of the Ramsey-Boiteux prices is the same as the prices that would be charged by an unregulated monopoly with the opportunity to engage in third-degree price discrimination.

6.2. *Non-linear prices: simple two-part tariffs*

Ramsey-Boiteux prices are still only second-best prices because the per unit usage prices are not equal to marginal cost. The distortion is smaller than for uniform ($p = \text{AC}$ in the single product case) pricing since we are taking advantage of differences in the elasticities of demand for different types of consumers or different products to satisfy the budget constraint yielding a smaller dead-weight loss from departures from marginal cost pricing. That is, there is still a wedge between the price for a product and its marginal cost leading to an associated dead-weight loss. The question is whether we can do better by further relaxing the restriction on the kinds of prices that the regulated firm can charge? Specifically, can we do better if we were to allow the regulated firm to charge a "two-part" price that includes a non-distortionary uniform fixed "access charge" (F) and then a separate per unit usage price (p). A price schedule or tariff of this form would yield a consumer expenditure or outlay schedule of the form:

$$T_i = F + pq_i$$

Such a price schedule is "non-linear" because the average expenditure per unit consumed T_i/q_i is no longer constant, but falls as q_i increases. We can indeed do (much) better from an efficiency perspective with two-part prices than we can with second-best (Ramsey-Boiteux) linear prices (Brown and Sibley, 1986, pp. 167–183).

Assume that there are N identical consumers in the market each with demand $q_i = d(p)$ and gross surplus of S_i evaluated at $p = 0$. The regulated firm's total cost function is given by $C = f_o + cq$. That is, there is a fixed cost f_o and a marginal cost c. Consider a tariff structure that requires each consumer to pay an access charge $A = f_o/N$ and

then a unit charge $p = c$. Consumer i's expenditure schedule is then:

$$T_i = A + pq_i = f_o/N + cq_i$$

This two-part tariff structure is first-best (ignoring income effects). On the margin, each consumer pays a usage price equal to marginal cost and the difference between the revenues generated from the usage charges and the firm's total costs are covered with a fixed fee that acts as a lump sum tax. As long as $A < (S_i - pq_i)$ then consumers will pay the access fee and consume at the efficient level. If $A > (S_i - pq_i)$ then it is not economical to supply the service at all because the gross surplus is less than the total cost of supplying the service (recall S_i is the same for all consumers and $p_i = c$).

Two-part tariffs provide a neat solution to the problem of setting efficient prices in this context when consumers are identical (or almost identical) or A is very small compared to the net surplus retained by consumers (i.e. after paying $pq_i = cq_i$). However, in reality consumers may have very different demands for the regulated service and A may be large relative to $(S_i - cq_i)$ for at least some consumers. In this case, if a single access fee $A = f_o/N$ is charged consumers with relatively low valuations will choose not to pay the access fee and consumer zero units of the regulated service even though their net surplus exceeds cq_i and they would be willing to make at least some contribution to the firm's fixed costs. A uniform two-part tariff would not be efficient in this case. However, if the regulator were truly fully informed about each consumer's individual demand and could prevent consumers from reselling the service, then a "discriminatory" two-part tariff could be tailored to match each consumer's valuation. In this case the customized/access fee A_i charged to each consumer would simply have to satisfy the condition $A_i < (S_i - cq_i)$ and there will exist at least one vector of A_i values that will allow the firm to satisfy the break-even constraint as long as it is efficient to supply the service at all.

If any of the conditions are met for two-part tariffs to be an efficient solution, the welfare gains compared to Ramsey-Boiteux pricing are likely to be relatively large [Brown and Sibley (1986, Chapter 7)].

6.3. *Optimal non-linear prices*

In reality, consumers are likely to be quite diverse and the regulator will not know each individual consumer's demand for the services whose prices they regulate. Can we use a variant of two-part tariffs to realize efficiency gains compared to either Ramsey-Boiteux prices or uniform two-part tariffs? In general, we can do better with non-linear pricing than with simple Ramsey-Boiteux pricing as long as the regulator is informed about the distribution of consumer demands/valuations for the regulated service in the population.

Consider the case where there are two types of consumers, one type (of which there are n_1 consumers) with a "low demand" and another type (of which there are n_2 consumers) with a "high demand." The inverse demand functions for representative type 1 and type 2 consumers are depicted in Figure 6 as $p = d_1(q_1)$ and $p = d_2(q_2)$. The cost function is as before with marginal cost $= c$. If we charge a uniform unit price of $p = c$,

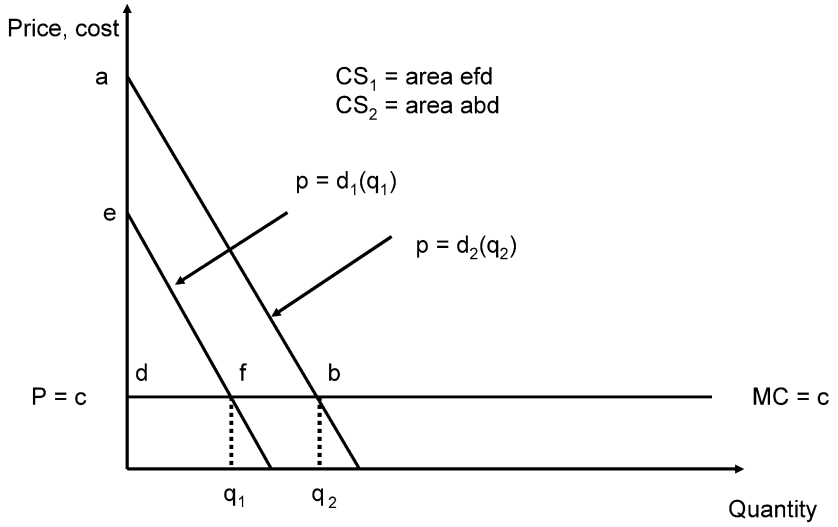


Figure 6. Heterogeneous consumers.

the net surplus for a low-type consumer is $\mathrm{CS}_1 = (S_1 - cq_1)$ and the net surplus for a high-type consumer is $\mathrm{CS}_2 = (S_2 - cq_2)$ where $\mathrm{CS}_1 < \mathrm{CS}_2$ and $n_1\mathrm{CS}_1 + n_2\mathrm{CS}_2 > f_o$. If the regulator were restricted to only a uniform two-part tariff, the highest access charge that could be assessed without forcing the low-value types off of the network would be $A = \mathrm{CS}_1$. If the total revenues generated when all consumers are charged an access fee equal to CS_1 is less than f_o then the break-even constraint would not be satisfied and the product would not be supplied even if it's total value is greater than its total cost. How can we extract more of the consumer's surplus out of the high demand types to cover the regulated firm's budget constraint with the minimum distortion to consumption decisions of both consumer types?

This is a simple example of the more general non-linear pricing problem. Intuitively, we can think of offering a menu of two-part tariffs of the form:

$$T_1 = A_1 + p_1 q_1$$
$$T_2 = A_2 + p_2 q_2$$

Where $A_1 < A_2$ and $p_1 > p_2 \geq c$ as in Figure 7 so that the low demand consumers find it most economical to choose T_1 and the high demand types choose T_2. In order to achieve this *incentive compatibility* property, the tariff T_1 with the low access fee must have a price p_1 that is sufficiently greater than p_2 to make T_1 unattractive to the high demand type. At the same time we would like to keep p_1 and p_2 as close to c as we can to minimize the distortion in consumption arising from prices being greater than marginal cost. The low-demand and high-demand types face a different price on the margin and the optimal prices are chosen to meet the break-even constraint

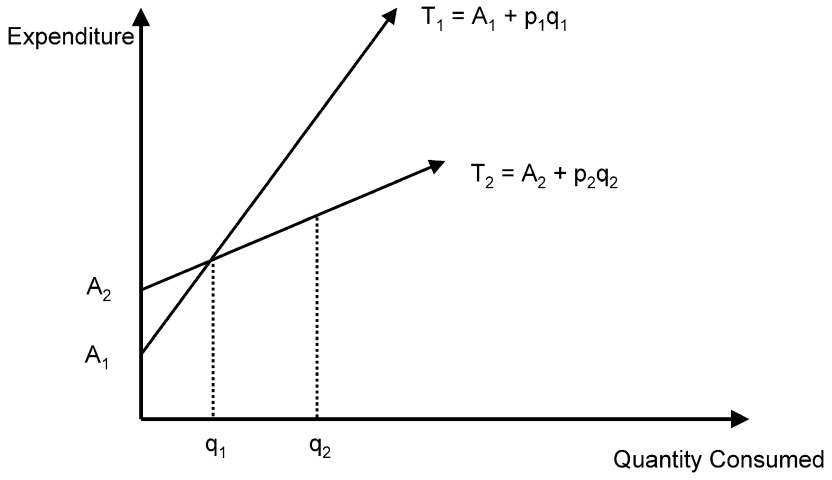


Figure 7. Two-part tariff.

with the minimum distortion. Note, that the menu above is equivalent to a single price schedule that has a single fixed fee A^* and then a usage fee that declines as consumption increases:

$$T(q) = A^* + p_1 q_1 + p_2\left(q_2 - q_1^*\right)$$

for q_1 between 0 and q_1^* and $q_2 > q_1^*$.

Let us turn to a more general case. Following Laffont and Tirole (2000, pp. 70–71) assume that the regulated firm's cost function is as before:

$$C = f_0 + cq$$

There is then a continuum of consumers with different demands for the regulated service and the consumer types are indexed by the parameter θ. A consumer of type θ will be confronted with a non-linear tariff $T(q)$ which has the property that the average expenditure per unit purchased on the service declines as q increases. Assume that the consumer of type θ consumes $q(\theta)$ when she faces $T(q)$ and has net utility $U(\theta) = \theta V[q(\theta)] - T[q(\theta)]$. (Note, this effectively assumes that the distribution of consumer demands shifts outward as θ increases and that the associated individual consumer demand curves do not cross. See Braeutigam (1989) and Brown and Sibley (1986) for more general treatments.)

Assume next that the parameter θ is distributed according to the c.d.f. $G(\theta)$ with density $g(\theta)$ with lower and upper bounds on θ of θ_L and θ_H respectively with the hazard rate $g(\theta)/[1 - G(\theta)]$ increasing with θ. Let $(1 + \lambda)$ denote the shadow cost of the firm's budget constraint. Then maximizing social welfare (gross consumers' surplus

net of the total costs of production) is equivalent to maximizing:

$$\int_{\theta_L}^{\theta_H} \{\theta V[q(\theta)] - T[q(\theta)]\}\, dG(\theta) - (1+\lambda)\int_{\theta_L}^{\theta_H} \{cq(\theta) + k_0 - T[q(\theta)]\}\, dG(\theta) \tag{7}$$

Let $U(\theta) \equiv \theta V[q(\theta)] - T[q(\theta)]$ and we obtain the constrained maximization problem for deriving the properties of the optimal non-linear prices

$$\max \int_{\theta_L}^{\theta_H} \big((1+\lambda)\{\theta V[q(\theta)] - cq(\theta) - k_0\} - \lambda U(\theta)\big)\, dG(\theta) \tag{8}$$

subject to:

$$\dot{U} = V[q(\theta)] \quad \text{and} \quad \dot{q} \geq 0 \tag{9}$$

$$U(\theta \geq 0 \quad \text{for all } \theta \tag{10}$$

where the first constraint is the incentive compatibility constraint and the second constraint is the constraint that all consumers with positive net surplus participate in the market.

Letting $\theta(q) = p(q) = T'(q)$ denote the marginal price that characterizes the optimal non-linear tariff, we obtain

$$\frac{p(q) - c}{p(q)} = \frac{\lambda}{1+\lambda}\frac{1 - G(\theta)}{\theta g(\theta)} \tag{11}$$

which implies that the optimal two-part tariff has the property that the marginal price falls toward marginal cost as θ increases or, alternatively we move from lower to higher demand types.

Willig (1978) shows that any second-best (Ramsey-Boiteux) uniform price schedule can be dominated from a welfare perspective by a non-linear price schedule. In some sense this should not be surprising. By capturing some infra-marginal surplus to help to cover the regulated firm's fixed costs, marginal prices can be moved closer to marginal cost, reducing the pricing distortion, while still satisfying the firm's budget balance constraint.

In fact, non-linear pricing has been used in the pricing of electricity, gas and telephone service since early in the 20th century [Clark (1911, 1913)]. Early proponents of non-linear pricing such as Samuel Insull viewed these pricing methods as a way to expand demand and lower average costs while meeting a break-even constraint [Hughes (1983, pp. 218–226)]. As is the case for uniform prices, the basic structure, though not the level, of the optimal non-linear prices is identical to the structure that would be chosen by a profit maximizing monopoly with the same information about demand patterns and the ability to restrict resale.

6.4. Peak-load pricing

Many public utility services cannot be stored and the demand for these services may vary widely from hour to hour, day to day and season to season. Because these services cannot be stored, the physical capacity of the network must be expanded sufficiently to meet peak demand. Services like electricity distribution and generation, gas distribution, and telephone networks are very capital intensive and the carrying costs (depreciation, interest on debt, return on equity investment) of the capital invested in this capacity is a relatively large fraction of total cost. For example, the demand for electricity varies widely between day and night, between weekdays and weekends and between days with extreme rather than moderate temperatures. Over the course of a year, the difference in demand between peak and trough may be on the order of a factor of three or more. The demand during the peak hour of a very hot day may be double the demand at night on that same day. Since electricity cannot be stored economically, the generating, transmission and distribution capacity of an electric power system must be sufficient to meet these peak demand days, taking into account equipment outages as well as variations in demand. Traditional telephone and natural gas distribution network have similar attributes.

The "peak load pricing" literature, which has been developed primarily in connection with the pricing of electricity, has focused on the specification of efficient prices and investment levels that take account of the variability of demand, the non-storability of the service, the attributes of alternative types of capital equipment available to supply electricity, equipment outages, and the types of metering equipment this is available and at what cost. There is a very extensive theoretical literature on efficient pricing and investment programs for electric power services that was developed mostly during the period 1950–1980 and primarily by French, British and American economists [Nelson (1964), Steiner (1957), Boiteux (1960), Turvey (1968a, 1968b), Kahn (1970), Crew and Kleinforfer (1976), Dreze (1964), Joskow (1976), Brown and Sibley (1986), Panzar (1976), Carlton (1977)]. This theoretical work was applied extensively to the pricing of electricity in France and in England during the 1950s and 1960s. There is little new of late on this topic and I refer interested readers to the references cited above.

The intuition behind the basic peak load pricing results is quite straightforward. If capacity must be built to meet peak demand then when demand is below the peak there will be surplus capacity available. The long run marginal cost of increasing supply to meet an increment in peak demand includes both the additional capital and operating costs of building and operating an increment of peak capacity. The long run marginal cost of increasing supply to meet an increment of off peak demand reflects only the additional operating costs or short run marginal cost of running more of the surplus capacity to meet the higher demand as long as off-peak demand does not increase to a level greater than the peak capacity on the system. Accordingly, the marginal social cost of increasing supply to meet an increase in peak demand will be much higher than the marginal cost of increasing supply to meet an increment of off-peak demand. Efficient price signals should convey these different marginal costs to consumers. Accordingly,

the peak price should be relatively high, reflecting both marginal operating and capital costs, and the off-peak prices low to reflect only the off-peak marginal costs of operating the surplus capacity more intensively.

The following simple model demonstrates this intuitive result and one of several interesting twists to it.

Let $q_D = q_D(p_D)$ = the demand for electricity during day-time hours

and $q_N = q_D(p_N)$ = the demand for electricity during night-time hours

for any $p_D = p_N$ day-time demand is higher than night-time demand ($q_D(p_D) > q_N(p_D)$). The gross surplus during each period (area under the demand curve) is given by $S(q_i)$ and $\frac{\partial S_i}{\partial q_i} = p_i$.

Assume that the production of electricity is characterized by a simple fixed-proportions technology composed of a unit rental cost C_K for each unit of generating capacity (K) and a marginal operating cost C_E for each unit of electricity produced. We will assume that there are no economies of scale, recognizing that any budget balance constraints can be handled with second-best linear or non-linear prices. Demand in any period must be less than or equal to the amount of capacity installed so that $q_D \leq K$ and $q_N \leq K$.

The optimal prices are then given by solving the following program which maximizes net surplus subject to the constraints that output during each period must be less than or equal to the quantity of capacity that has been installed:

$$L^* = S(q_D) + S(q_N) - C_K K - C_E(q_D + q_N) + \lambda_D(K - q_D) + \lambda_N(K - q_N) \tag{12}$$

where λ_D and λ_N are the shadow prices on capacity The first order conditions are then given by:

$$p_D - C_E - \lambda_D = 0$$
$$p_N - C_E - \lambda_N = 0$$
$$\lambda_D + \lambda_N - C_K = 0$$

with complementary slackness conditions

$$\lambda_D(K - q_D) = 0$$
$$\lambda_N(K - q_N) = 0$$

There are then two interesting cases:

Case 1: Classic peak load pricing results:

$$P_D = C_E + C_k \quad (\lambda_D = C_K) \tag{13}$$
$$P_N = C_E \quad (\lambda_N = 0) \tag{14}$$
$$q_N < q_D \tag{15}$$

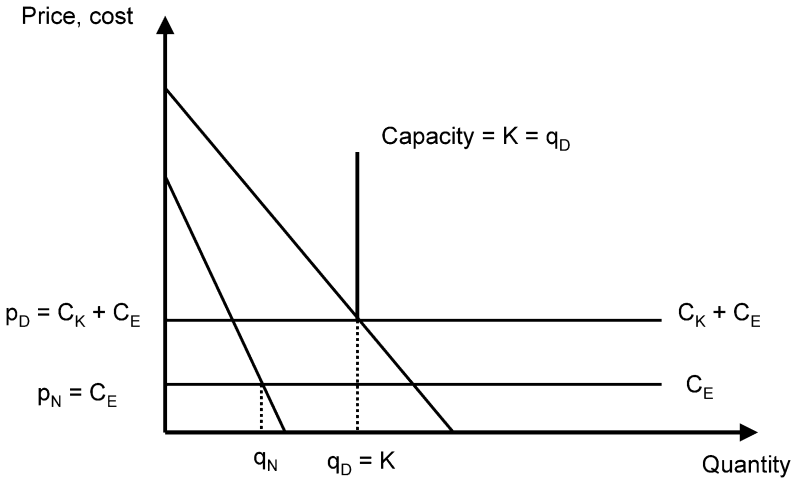


Figure 8. Peak-load pricing.

In this case, the optimal price during the peak period equals the sum of marginal operating costs and marginal capacity costs. In the off-peak period the optimal price equals only marginal operating costs. The result is depicted in Figure 8.

Case 2: Shifting peak case:

$$P_{\mathrm{D}} = C_{\mathrm{E}} + \lambda_{\mathrm{D}} \quad (\lambda_{\mathrm{D}} > 0) \tag{16}$$

$$P_{\mathrm{N}} = C_{\mathrm{E}} + \lambda_{\mathrm{N}} \quad (\lambda_{\mathrm{N}} > 0) \tag{17}$$

$$\lambda_{\mathrm{D}} + \lambda_{\mathrm{N}} - C_{\mathrm{K}} = 0 \tag{18}$$

$$q_{\mathrm{D}} = q_{\mathrm{N}} \tag{19}$$

Here, the optimal prices during the peak and off peak periods effectively share the marginal cost of capacity plus the marginal cost of producing electricity. The peak period price includes a larger share of the marginal cost of capacity than the off-peak price reflecting the differences between consumers' marginal willingness to pay between the two periods. This result is reflected in Figure 9.

The standard case is where $\lambda_{\mathrm{D}} > 0$ and $\lambda_{\mathrm{N}} = 0$. The peak price now equals the marginal capital and operating cost of the equipment and the off-peak price equals only the marginal operating costs. Investment in capacity K is made sufficient to meet peak demand ($\mathrm{K} = q_{\mathrm{D}} > q_{\mathrm{N}}$) and consumers buying power during the peak period pay all of the capital costs. Consumption during the day then carries a higher price than consumption at night. Does this imply that there is price discrimination at work here? The answer is no. Peak and off-peak consumption are essentially separate products and supply in both periods each pay their respective marginal supply costs. What is true, is

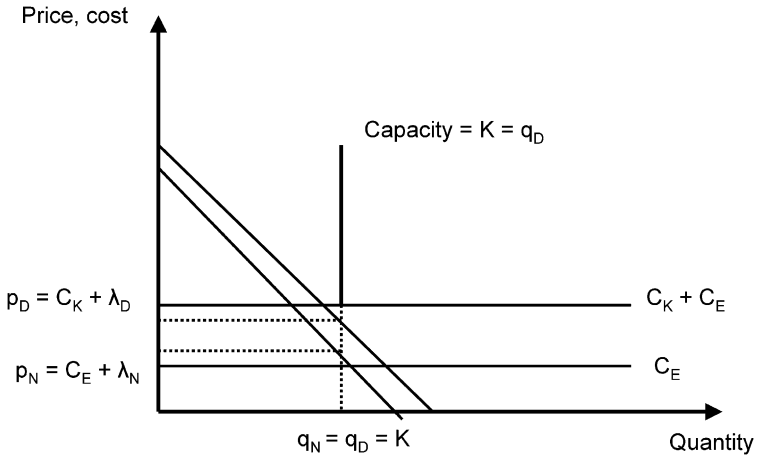


Figure 9. Peak-load pricing with sifting peak.

that the production of peak and off peak supplies are "joint products" that incur joint costs. That is, off-peak supply could not be provided so inexpensively if peak demand was not there to pay all of the capital costs.

The role of joint costs becomes evident when we look at the second potential solution of the simple problem above. This potential solution has the following properties:

$$q_{\mathrm{D}} = q_{\mathrm{N}}$$

and

$$\lambda_{\mathrm{D}} + \lambda_{\mathrm{N}} = C_{\mathrm{K}}$$

In this potential equilibrium, peak and off-peak consumption each bear a share of the capital or capacity costs. This situation arises when the peak and off-peak demands are so elastic that applying the simple peak load pricing rule that the peak demand pays for all capital costs and the off-peak demand for none, ends up shifting the peak demand to the off-peak (night) period. This problem was realized in practice in a number of countries that instituted simple peak load pricing models for electricity during the 1960s and 1970s. The off-peak price was set so low that many consumers installed electric heating equipment that on cold winter nights led electricity demand to reach its peak levels. The solution to the problem is to share the joint capacity costs in a way that reflects the relative valuations of peak and off-peak consumption (at the equilibrium prices) as displayed in Figure 9. The off-peak price still is lower than the peak price, but now all consumption pays a share of the network's capacity costs. Note as well, that the implementation of efficient prices now requires the regulator to have information about the elasticity of demand in different time periods.

There are numerous realistic complications that can and have been introduced into simple peak load pricing models such as the one above. Suppliers can choose among dif-

ferent production techniques with different (in the case of electricity) capital/fuel ratios. In addition, demand cannot be divided simply between "peak" and "off-peak." Rather the system is characterized by continuously varying demands that lie between some lower and upper bound. In this case, since some of the capacity is utilized relatively few hours each year, some during all hours and some for say half the hours of the year, it is economical to install a mix of "base load," "intermediate" and "peak load" capacity [Turvey (1968b), Crew and Kleinforfer (1976), Joskow (1976, 2006)] and by allowing prices to vary with marginal production costs, produce infra-marginal quasi rents to cover some of the costs of investments in production facilities. In addition, it turns out that even ignoring the "shifting peak" issue discussed above, when consumption is priced at marginal operating cost during most time periods, consumers during this hours make a contribution to the capital costs of the network because the marginal operating costs of the now diverse electricity production technologies on the network increases as the demand on the network increases. Enhancements of these models have also considered the stochastic attributes of demand and equipment (unplanned outages and planned maintenance requirements) to derive both optimal levels of reserve capacity and the associated optimal prices with and without real time price variations [Carlton (1977)].

The marginal cost of producing electricity varies almost continuously in real time. And when short run capacity constraints are reached the social marginal cost can jump to the valuation of the marginal consumer who is not served [the value of lost load or the value of unserved energy—Joskow and Tirole (2006)]. Many economists argue that electricity prices should vary in real time to convey better price signals (Borenstein, 2005). However, any judgment about which consumers should pay real time prices must take into account the transactions costs associated with recording consumption in real time, collecting and analyzing the associated data. It is generally thought that for larger customers the welfare gains from better pricing exceed the costs of installing and utilizing more sophisticated meters.

Variable demand, diverse technologies, reliability and real time pricing can all be integrated with the "budget balance" constraint considerations discussed earlier. The same basic second-best pricing results hold, though the relevant marginal costs are now more complicated as is the implementation of the budget balance constraint since when stochastic demand and supply attributes are introduced, the firm's revenues, costs and profits also become uncertain [Crew and Kleinforfer (1986)].

7. Cost of service regulation: response to limited information

The discussion of optimal pricing for a natural monopoly in the last section assumes that the regulator knows all there is to know about the regulated firm's costs and demand. In addition, the regulated firm does not act strategically by changing its managerial effort to increase costs or to distort the information the regulator possesses about its cost opportunities and the demand it faces for the services it provides in response to the incentives created by the regulatory mechanisms that have been chosen. In reality, regulators are not inherently well informed about the attributes of the firm's cost

opportunities, its demand, or its management's effort and performance. The regulated firm knows much more about these variables than does the regulator and, if given the opportunity, may have incentives to act strategically. The firm may provide incorrect information about the its cost, demand and managerial effort attributes to the regulator or the firm may respond to poorly designed regulatory incentives by reducing managerial effort, increasing costs, or reducing the quality of service. Much of the evolution of regulatory agencies and regulatory procedures in the U.S. during the last hundred years has focused on making it possible for the regulator to obtain better information about these variables and to use this information more effectively in the regulatory process. More recent theoretical and empirical research has focused on the development of more efficient regulatory mechanisms that reflect these information asymmetries and associated opportunities for strategic behavior as well as to better exploit opportunities for the regulator to reduce its information disadvantages.

I have chosen to begin the discussion of regulation when the regulator has limited information about the attributes of the firm and its customers with a discussion of traditional "cost of service" or "rate of return regulation" that has been the basic framework for commission regulation in the U.S. during most of the 20th century. The performance of this regulatory process (real and imagined) is the "benchmark" against which alternative mechanisms are compared. The "traditional" cost of service regulation model is frequently criticized as being very inefficient but the way it works in practice is also poorly understood by many of its critics. Its application in fact reflects efforts to respond to imperfect and asymmetric information problems that all regulatory processes must confront. Moreover, the application of modern "incentive regulation" mechanisms is frequently an addition to rather than a replacement for cost-of-service regulation [Joskow (2006)]. After outlining the attributes of cost of service regulation in practice I proceed to discuss the "Averch-Johnson model," first articulated in 1962 and developed extensively in the 1970s and 1980s, which endeavors to examine theoretically the efficiency implications of rate of return regulation and variations in its application.

7.1. Cost-of-service or rate-of-return regulation in practice

U.S. regulatory processes have approached the challenges created by asymmetric information in a number of ways. First, regulators have adopted a *uniform system of accounts* for each regulated industry. These cost-reporting protocols require regulated firms to report their capital and operating costs according to specific accounting rules regarding the valuation of capital assets, depreciation schedules, the treatment of taxes, operating cost categories, allocation of costs between lines of business and between regulated and unregulated activities, and the financial instruments and their costs used by the firm to finance capital investments. These reports are audited and false reports can lead to significant sanctions. Since most U.S. states and federal regulatory agencies use the same uniform system of accounts for firms in a particular industry, the opportunity to perform comparative analyses of firm costs and to apply "yardstick regulation" concepts also becomes a distinct possibility (more on this below). Regulatory agencies also have broad

power to seek additional information from regulated firms that would not normally be included in the annual reports under the uniform system of accounts; for example data on equipment outage and other performance indicia, customer outages, consumer demand patterns, etc., and to perform special studies such as demand forecasting and demand elasticity measurement. These data collection and analysis requirements are one way that U.S. regulators can seek to increase the quality of the information they have about the firms they regulate and reduce the asymmetry of information between the regulator and the firms that it regulates. Whether they use these data and authorities wisely is another matter.

Regulators in the U.S. and other countries have long known, however, that better data and analysis cannot fully resolve the asymmetric information problem. There are inherent differences between firms in terms of their cost opportunities and the managerial skill and effort extended by their managements and traditional accounting methods for measuring capital costs in particular may create more confusion than light. Accordingly, the regulatory process does not require regulators to accept the firms' reported and audited accounting costs as "just and reasonable" when they set prices. They can "disallow" costs that they determine are unreasonably through, for example, independent assessments of firm behavior and comparisons with other comparable firms [Joskow and Schmalensee (1986)]. Moreover, contrary to popular misconceptions, regulated prices are not adjusted to assure that revenues and costs are exactly in balance continuously. There are sometimes lengthy periods of "regulatory lag" during which prices are fixed or adjust only partially in response to realized costs the regulated firm shares in the benefits and burdens of unit cost increases or decreases [Joskow (1973, 1974), Joskow and Schmalensee (1986)]. And regulators may specify simple "incentive regulation" mechanisms that share variations in profitability between the regulated firm and its customers. These are generally called "sliding scale" regulatory mechanisms [Lyon (1996)], a topic that will be explored presently.

The "fixed point" of traditional U.S. regulatory practice is the *rate case* [Phillips (1993)]. The rate case is a public quasi judicial proceeding in which a regulated firm's prices or "tariffs" may be adjusted by the regulatory agency. Once a new set of prices and price adjustment formulas are agreed to with the regulator (and sustains any court challenges) they remain in force until they are adjusted through a subsequent rate case. Contrary to popular characterizations, regulated prices are not adjusted continuously as cost and demand conditions change, but rather are "tested" from time to time through the regulatory prices. Rate cases do not proceed on a fixed schedule but are triggered by requests from the regulated firm, regulators, or third-parties for an examination of the level or structure of prevailing tariffs [Joskow (1973)]. Accordingly, base prices are fixed until they are adjusted by the regulator through a process initiated by the regulated firm or by the regulator on its own initiative (perhaps responding to complaints from interested parties [Joskow (1972)]. This "regulatory lag" between rate cases may be quite long and has implications for the incentives regulated firms have to control costs (Joskow, 1974) and the distribution of surplus between the regulated firm and consumers.

A typical rate case in the U.S. has two phases. The first phase determines the firm's total *revenue requirement* or its total *cost of service*. It is convenient to think of the revenue requirements or cost of service as the firm's budget constraint. The second phase is the *rate design* or *tariff structure* phase. In this phase the actual prices that will be charged for different quantities consumed or to different types of consumers or for different products is determined. It is in the rate design phase where concepts of Ramsey pricing, non-linear pricing and peak load pricing would be applied in practice.

The firm's revenue requirements or cost of service has numerous individual components which can be grouped into a few major categories:

a. Operating costs (e.g. fuel, labor, materials and supplies)—OC
b. Capital related costs that define the effective "rental price" for capital that will be included in the firm's total "cost of service" for any given time period. These capital related charges are a function of:
   i. the value of the firm's "regulatory asset base" or its "rate base" (RAV)
   ii. the annual amount of depreciation on the regulatory asset base (D)
   iii. the allowed rate of return (s) on the regulatory asset base
   iv. income tax rate (t) on the firm's gross profits
c. Other costs (e.g. property taxes, franchise fees)—F

7.1.1. Regulated revenue requirement or total cost of service

The regulated firm's total *revenue requirements* or *cost of service* in year t is then given by

$$R_t = \mathrm{OC}_t + D_t + r(1+t)\mathrm{RAV} + F_t \tag{20}$$

These cost components are initially drawn from the regulated firm's books and records based on a uniform system of accounts adopted by the regulator. An important part of the formal rate case is to evaluate whether the firm's costs as reported on its books or projected into the future are "reasonable." The regulatory agency may rely on its own staff's evaluations to identify costs that were "unreasonable" or unrepresentative of a typical year, or the regulator may also rely on studies presented by third-party "intervenors" in the rate case [Joskow (1972)]. Interested third-parties are permitted to participate fully in a rate case and representatives of different types of consumers, a public advocate, and non-government public interest groups often participate in these cases, as well as in any settlement negotiations that are increasingly relied upon to cut the administrative process short. Costs that the regulatory agency determines are unreasonable are then "disallowed" and deducted from the regulated firm's cost of service.

There are a number of methods available to assess the "reasonableness" of a firm's expenditures. One type of approach that is sometimes used is a "yardstick" approach in which a particular firm's costs are compared to the costs of comparable firms and significant deviations subject to some disallowance [e.g. Jamasb and Pollitt (2001, 2003), Carrington, Coelli, and Groom (2002), Schleifer (1985)]. Such an approach has been

used to evaluate fuel costs, labor productivity, wages, executive compensation, construction costs and other costs. A related approach is to retain outside experts to review the firm's expenditure experience in specific areas and to opine on whether they were reasonably efficient given industry norms. The regulator may question assumptions about future demand growth, the timing of replacements of capital equipment, wage growth, etc. Finally, accountants comb through the regulated firm's books to search for expenditures that are either prohibited (e.g. Red Sox tickets for the CEO's family) or that may be of questionable value to the regulated firm's customers (e.g. a large fleet of corporate jets). These reasonableness review processes historically tended to be rather arbitrary and ad hoc in practice, but have become more scientific over time as benchmarking methods have been developed and applied. Since the regulated firm always knows more about its own cost opportunities and the reasons why it made certain expenditures than does the regulator, this process highlights the importance of thinking about regulation from an asymmetric information perspective.

From the earliest days of rate or return regulation, a major issue that has been addressed by academics, regulators and the courts is the proper way to value the firm's assets in which it has invested capital and how the associated depreciation rates and allowed rate of return on investment should be determined [Sharfman (1928), Phillips (1993), Bonbright (1961), Clemens (1950)]. A regulated firm makes investments in long-lived capital facilities. Regulators must determine how consumers will pay for the costs of these facilities over their economic lives. A stock (the value of capital investments) must be transformed into a stream of cash flows or annual rental charges over the life of the assets in which the regulated firm has invested in order to set the prices that the firm can charge and the associated revenues that it will realize to meet the firm's overall budget constraint.

The basic legal principle that governs price regulation in the U.S. is that regulated prices must be set at levels that give the regulated firm a reasonable opportunity to recover the costs of the investments it makes efficiently to meet its service obligations and no more than is necessary to do so.[15] One way of operationalizing this legal principle is to reduce it to the rule that the present discounted value of future cash flows that flow back to investors in the firm (equity, debt, preferred stock) should be at least equal to the cost of the capital facilities in which the firm has invested,. Where the discount rate is the firm's risk adjusted cost of capital "r." Let:

Π_t = cash flow in year t = Revenues − operating costs − taxes and other expenses

K_0 = the "reasonable" cost of an asset added by the utility in year t

r = the firm's after tax opportunity cost of capital

[15] *Federal Power Commission v. Hope Natural Gas Co., 320 U.S. 591, 602 (1944).*

The basic rule for setting prices to provide an appropriate return of and on an investment in an asset with a cost of K_0 made by the regulated firm can then be defined as:

$$K_0 \leq \sum_{1}^{n} \frac{\pi_t}{(1+r)^t} \tag{21}$$

where n is the useful/economic life of the asset. If this condition holds, then the regulated firm should be willing to make the investment since it will cover its costs, including a return on its investment greater than or equal to its opportunity cost of capital. If the relationship holds with equality then consumers are asked to pay no more than is necessary to attract investments in assets required to provide services efficiently. Since a regulated firm will typically be composed of many assets reflecting investments in capital facilities made at many different times in the past, this relationship must hold both on the margin and in the aggregate for all assets. However, it is easiest to address the relevant issues by considering a single-asset firm, say a natural gas pipeline company, with a single productive asset that works perfectly for n years and then no longer works at all (a one-horse-shay model).

There are many (infinite) different streams of cash flows that satisfy the NPV condition in (21) for a single asset firm that invests in the asset in year 1 and uses it productively until it is retired in year n. These cash flows can have many different time profiles. Cash flows could start high and decline over time. Cash flows could start low and rise over time. Cash flows could be constant over the life of the asset. Much of the historical debate about the valuation of the regulatory asset base, the depreciation rate and rate of return values to be used to turn the value of the capital invested by the regulated monopoly firm in productive assets into an appropriate stream of cash flows over time, has reflected alternative views about the appropriate time profile and (perhaps unintended) the ultimate level of the present discounted value of these cash flows. Unfortunately, this debate about asset valuation, depreciation and allowed return was long on rhetoric and short on mathematical analysis, had difficulty dealing with inflation in general and confused real and nominal interest rates in particular [Sharfman (1928), Clemens (1950, Chapter 7)]. The discussion and resulting regulatory accounting rules also assume that the regulated firm is a monopoly, does not face competition, and will be in a position to charge prices that recover the cost of the investment over the accounting life of the asset. Giving customers the option to switch back and forth from the regulated firms effectively imbeds a costly option into the "regulatory contract" and requires alternative formulas for calculating prices that yield revenues with an expected value equal to the cost of the investment [Pindyck (2004), Hausman (1997), Hausman and Myers (2002)].

A natural starting point for an economist is to rely on economic principles to value the regulated firm's assets. We would try to calculate a pattern of rental prices for the firm's assets that simulates the trajectory of rental prices that would emerge if the associated capital services were sold in a competitive market. This approach implies valuing assets at their competitive market values, using economic depreciation concepts, and taking

appropriate account of inflation in the calculation of real and nominal interest rates. Consider the following simple example:

Assume that a machine producing a homogeneous product depreciates (physically) at a rate d per period. You can think of this as the number of units of output from the machine declining at a rate d over time. Assume that operating costs are zero. Define the competitive rental value for a new machine at any time s by $v(s)$. Then in year s the rental value on an old machine bought in a previous year t would be

$$v(s)\mathrm{e}^{-d(s-t)}$$

since $(s-t)$ is the number of years the machine has been decaying.

The price of a new machine purchased in year t $[P(t)]$ is the present discounted value of future rental values. Let r be the firm's discount rate (cost of capital). Then the present value of the rental income in year s discounted back to year t is

$$\mathrm{e}^{-r(s-t)}v(s)\mathrm{e}^{-d(s-t)} = \mathrm{e}^{(r+d)t}v(s)\mathrm{e}^{-(r+d)s}$$

and the present value of the machine in year t is:

$$\begin{aligned}\mathrm{PDV}(t) &= \int_t^\infty \mathrm{e}^{(r+d)t}v(s)\mathrm{e}^{-(r+d)s}\,\mathrm{d}s \\ &= \text{competitive market price for a new machine in year } t \\ &= P(t) \qquad (22)\end{aligned}$$

Rewrite this equation as:

$$P(t) = \mathrm{e}^{(r+d)t}\int_t^\infty v(s)\mathrm{e}^{-(r+d)s}\,\mathrm{d}s \qquad (23)$$

and differentiate with respect to t

$$\mathrm{d}P(t)/\mathrm{d}t = (r+d)P(t) - v(t)$$

or

$$v(t) = (r+d)P(t) - \mathrm{d}P(t)/\mathrm{d}t$$

where $\mathrm{d}P(t)/\mathrm{d}t$ reflects exogenous changes in the price of new machines over time. These price changes reflect general inflation (i) and technological change (δ) leading to lower cost machines (or more productive machines). The changes in the prices of new machines affect the value of old machines because new machines must compete with old machines producing the same product.

$$\begin{aligned}v(t) &= (r+d)P(t) - (i-\delta)P(t) \\ &= (r+d-i+\delta)P(t) \qquad (24)\end{aligned}$$

The economic depreciation rate is then $(d-i+\delta)$ and the allowed rate of return consistent with it is given by r the firm's nominal cost of capital. Both are applied to the current competitive market value of the asset $P(t)$.

Equation (24) provides the basic formula for setting both the level and time profile of the capital cost component of user prices for this single-asset regulated firm assuming that there is a credible regulatory commitment to compensate the firm in this way over the entire economic life of the asset. Even though the value of the regulatory asset is effectively marked to market on a continuing basis, the combination of sunk costs and asset specificity considerations would require a different pricing arrangement if, for example, customers were free to turn to competing suppliers if changing supply and demand conditions made it economical to do so [Pindyck (2004), Hausman (1997), Hausman and Myers (2002)].

The earliest efforts to develop capital valuation and pricing principles indeed focused on "fair value" rate base approaches in which the regulated firm's assets would be revalued each year based on the consideration of "reproduction cost," and other "fair market value" methods, including giving some consideration to "original cost" [Troxel (1947, Chapters 12 and 13), Clemens (1950, Chapter 7), Kahn (1970, pp. 35–45)]. Implementing these concepts in practice turned out to be very difficult with rapid technological change and widely varying rates of inflation. Regulated firms liked "reproduction cost new" methods for valuing assets when there was robust inflation (as during the 1920s), but not when the nominal prices of equipment were falling (as in the 1930s). Moreover, "fair market value" rules led regulated firms to engage in "daisy chains" in which they would trade assets back and forth at inflated prices and then seek to increase the value of their rate bases accordingly. Methods to measure a firm's cost of capital were poorly developed [Troxel (1947, Chapters 17, 18, 19)]. Many regulated firm asset valuation cases were litigated in court. The guidance given by the courts was not what one could call crystal clear [Troxel (1947, Chapter 12)].

Beginning in the early 1920s, alternatives to the "fair value" concept began to be promoted. In a dissenting opinion in 1923,[16] Justice Louis Brandeis criticized the "fair valuation" approach. He proposed instead a formula that is based on what he called the prudent investment standard. Regulators would first determine whether an investment and its associated costs reflected "prudent" or reasonable decisions by the regulated firm. If they did, investors were to be permitted to earn a return of and on the "original cost" of this investment. The formula for determining the trajectory of capital related charges specified that regulators should use straight line depreciation of the original cost of the investment, value the regulatory asset base at the original cost of plant and equipment prudently incurred less the accumulated depreciation associated with it at any particular point in time, and apply an allowed rate of return equal to the firm's nominal cost of capital.

Consider a single asset firm with a prudent investment cost of K_0 at time zero. The Brandeis formula would choose an accounting life N for the asset. The annual depreciation was then given by $D_t = K_0/N$. The regulatory asset base in any year was then given by $\text{RAV}_t = K_0 - \sum_0^t D_t$. Then prices are set to produce net cash flows (after

[16] *Southwestern Bell Telephone Company v. Public Service Commission of Missouri 262 U.S. 276 (1923).*

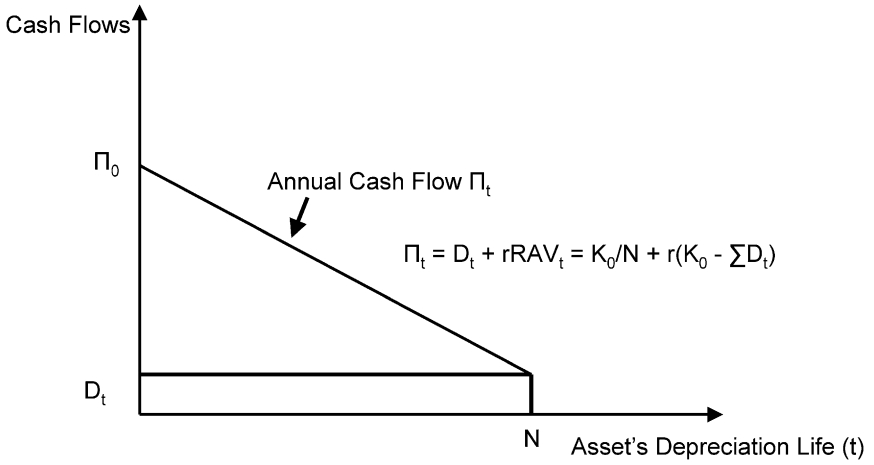


Figure 10. Depreciated original cost rate base.

operating costs, taxes and other allowable fees) based on the following net cash flow formula

$$\Pi_t = D_t + r\mathrm{RAV}_t = K_0/N + r\left(K_0 - \sum_0^t D_t\right) \tag{25}$$

which can be easily extended to multiple assets with different in-service dates and service lives. The cash flow profile for a single-asset firm is displayed in Figure 10. Brandeis argued that this approach would make it possible for regulators and the courts to "avoid the 'delusive' calculations, 'shifting theories," and varying estimates that the engineers use as they measure the reproduction costs and present values of utility properties." [Troxel (1947, p. 271)] while providing regulated firms with a fair return on the prudent cost of investments that they have made to support the provisions of regulated services.

The Brandeis formula is quite straightforward, and the prudent investment standard compatible with a regulatory system that guards against regulatory hold-ups of investors ex post. However, does it satisfy the NPV criterion discussed earlier and, in this way, provide an expected return that is high enough to attract investment, but not so high that it yield prices significantly higher than necessary to attract investment? It turns out that the Brandeis formula satisfies the NPV criterion (Schmalensee, 1989a). The present discounted value of cash flows calculated using the Brandeis formula (including an allowed rate of return that is equal to the regulated firm's nominal opportunity cost of capital) is exactly equal to the original cost of the investment; investors get a return of their investment and a return on their investment equal to their opportunity cost of capital. As Brandeis suggested, it provide a simple and consistent method for compen-

sating investors for capital costs and eliminates the uncertainties and opportunities for manipulation that characterized the earlier application of "fair valuation" concepts.

Beginning in the 1930s, regulators began to adopt and the courts began to accept the prudent investment/original cost approach and by the end of World War II it became the primary method for determining the capital charge component of regulated prices. In the *Hope* decision in 1944, the Supreme Court concluded that from a Constitutional perspective it was the "result" that mattered rather than the choice of a particular method and, in this way, getting the courts disentangled from deciding whether or not specific details of the regulatory formulas chosen by state and federal regulators passed Constitutional muster.

> "Under the statutory standard of 'just and reasonable it is the result reached not the method employed which is controlling."[17]
> "Rates which enable the company to operate successfully, to maintain its financial integrity, to attract capital, and to compensate its investors for the risk assumed certainly cannot be condemned as invalid, even though they might produce only a meager return on the so-called 'fair value' rate base."[18]

While the prudent investment/depreciated original cost standard satisfies the NPV criterion, and may have other attractive properties for attracting investment to regulated industries, it also has some peculiarities. These can be seen most clearly for the single asset company (e.g. a pipeline). The time pattern of capital charges has the property of starting at a particular level defined by the undepreciated (or barely depreciated) RAV equal to a value close to K_0 and then declining continuously over the life of the asset until it approaches zero at the end of its useful life (see Figure 10). However, there is no particular reason to believe that the annual capital charges defined by the formula *at any particular point in time*, reflect the "competitive" capital charges or rental rates that would emerge in a competitive market. For example, if we apply the economic depreciation and competitive market value RAV formula discussed earlier, if there is inflation but no technological change, the capital charges for the asset should increase at the rate of inflation over time rather than decrease steadily as they do with the Brandeis formula. In this case if we use the Brandeis formula, regulated prices start out too high and end up too low when the Brandeis formula is applied in this case. If the asset is replaced in year $N + 1$ and the Brandeis formula applied de novo to the new asset, the price for capital related charges will jump back to the value in year 1 (assuming zero inflation and no technological change) and then gradually decline again over time. Thus, while the Brandeis formula gives the correct NPV of cash flows to allow for recovery of a return of and on investment, it may also yield the wrong prices (rental charges associated with capital costs) at any particular point in time. This in turn can lead to the standard consumption distortions resulting from prices that are too high or too low.

[17] *Federal Power Commission v. Hope Natural Gas Co., 320 U.S. 591, 602 (1944).*
[18] Ibid at 605.

Moreover, because assets do not reflect their market value at any particular point in time, the Brandeis formula can and has led to other problems. Regulated prices for otherwise identical firms may be very different because the ages of their assets happen to be different from one another even if their market values are the same. An old coal-fired power plant may have a much higher market value than a new oil-fired power plant, but the prices charged to consumers of the regulated firm with the old coal plants will be low while those of the utility with the new oil-fired plant may be high. As an asset ages, the capital charges associated with it approach zero. For a single asset company, when this asset is replaced at an original cost reflecting current prices, the application of the Brandeis formula leads to a sudden large price increase (known as "rate shock") which creates both consumption distortions and political problems for regulators. Finally, when assets are carried at values significantly greater than their market values, it may create incentives for inefficient entry as well as transition problems when competition is introduced into formerly regulated industries. Who pays for the undepreciated portion of the new oil plant that has a low competitive market value and who gets the benefits from deregulating the old coal plant whose market value is much higher than its RAV when competition replaces regulated monopoly? These so-called "stranded cost" and "stranded benefit" attributes [Joskow (2000)] of the Brandeis formula have plagued the transitions to competition in telecommunications (e.g. mechanical switches that were depreciated too slowly in the face of rapid technological change) and electric power (e.g. costly nuclear power plants that were "prudent" investments when they were made).

It turns out that any formula for calculating the annual capital or rental charge component of regulated prices that has the property (a) the firm earns its cost of capital each period on a rate base equal to the depreciated original cost of its investments and (b) earns the book depreciation deducted from the rate base in each period, satisfies the NPV and investment attraction properties of the Brandeis formula [Schmalensee (1989a)]. There is nothing special about the Brandeis formula in this regard. Alternative formulas that have capital charges for an asset rise, fall or remain constant over time can be specified with the same NPV property. So, in principle, the Brandeis formula could be adjusted to take account of physical depreciation, technological change and inflation to better match both the capital attraction goals and the efficient pricing goals of good regulation.

Note that if a capital investment amortization formula of this type is used, the present discounted value of the firm's net cash flows using the firm's cost of capital as the discount rate should equal its regulatory asset base (RAV) or what is often referred to as its regulatory "book value" B. If investors value the firm based on the net present value of its expected future cash flows using the firm's discount rate then the regulated firm's market value M should be equal to its book value B at any point in time. Accordingly, a simple empirical test is available to determine whether the regulatory process is expected by investors to yield returns that are greater than, less than or equal to the firm's cost of capital. This test involves calculating the ratio of the firm's market value to its

book value:

$M/B = 1 \rightarrow$ Expected returns equal to the cost of capital

$M/B > 1 \rightarrow$ Expected returns greater than the cost ofcapital

$M/B < 1 \rightarrow$ Expected returns less than the cost of capital

In the presence of regulatory lag, we would not expect the M/B always to be equal to one. Moreover, as we shall discuss presently, there may be very good incentive reasons to adopt incentive regulatory mechanisms that, on average, will yield returns that exceed a typical firm's cost of capital. In fact, M/B ratios for regulated electric utilities have varied widely over time [Joskow (1989), Greene and Smiley (1984)] though during most periods of time they have exceeded 1. This is consistent with the observation I made earlier. Due to regulatory lag, a regulated firm's prices are not adjusted continuously to equal its actual costs of production. Deviations between prices and costs may persist for long periods of time [Joskow (1972, 1974)] and have significant effects on the regulated firm's market value [Joskow (1989)]. Accordingly, regulatory lag has both incentive effects and rent extraction effects that are often ignored in uniformed discussions of traditional cost of service regulation.

The final component of the computation of the capital charges that are to be included in regulated prices involves the calculation of the allowed rate of return on investment. Regulatory practice is to set a "fair rate of return" that reflects the firm's nominal cost of capital. Regulated firms are typically financed with a combination of debt, equity and preferred stock [Spiegel and Spulber (1994), Myers (1972a, 1972b)]. The allowed rate of return is typically calculated as the weighted average of the interest rate on debt, preferred stock and an estimate of the firm's opportunity cost of capital, taking the tax treatment of interest payments and the taxability of net income that flows to equity investors. So consider a regulated firm with the following capital structure:

Instrument	Average coupon rate	Fraction of capitalization
Debt	8.0%	50%
Preferred stock	6.0%	10%
Equity	N/A	40%

Then the firm's weighted average cost of capital (net of taxes) is given by

$$r = 8.0 * 0.5 + 6.0 * 0.1 + r_e * 0.4 \tag{26}$$

where r_e is the firm's opportunity cost of equity capital which must then be estimated. Rate cases focus primarily on estimating the firm's opportunity cost of equity capital and, to a lesser degree, determining the appropriate mix of debt, preferred stock, and equity that composes the firm's capital structure. A variety of methods have been employed to measure the regulated firm's cost of equity capital [Myers (1972a, 1972b)], including the so-called discounted cash flow model, the capital asset pricing model, and other "risk premium" approaches [Phillips (1993, pp. 383–412)]. At least in the U.S.,

the methods that are typically used to estimate the regulated firm's cost of capital are surprisingly unsophisticated given the advances that have been made in theoretical and empirical finance in the last thirty years.

With all of these cost components computed the regulator adds them together to determine the firm's "revenue requirement" or total "cost of service" R. This is effectively the budget balance constraint used by the regulator to establish the level and structure of prices—the firm's "tariff"—for the services sold by the regulated firm.

7.1.2. Rate design or tariff structure

In establishing the firm's tariff structure or rate design, regulators typically identify different "classes" of customers, e.g. residential, commercial, farm, and industrial, which may be further divided into further sub-classes (small commercial, large commercial, voltage level differentiation) [Phillips (1993, Chapter 10)]. Since regulatory statutes often require that prices not be "unduly discriminatory," the definition of tariff classes typically is justified by differences in the costs of serving the different groups. In reality, the arbitrariness of allocating joint costs among different groups of customers provides significant flexibility for regulators to take non-cost factors into account [Bonbright (1961), Clemens (1950, Chapter 11), Salinger (1998)]. For example, residential electricity customers require a costly low-voltage distribution system while large industrial customers take power from the network at higher voltages and install their own equipment to step down the voltage for use in their facilities. Accordingly, since the services provided to residential customers have different costs from those provided to large industrial customers it makes economic sense to charge residential and industrial customers different prices. At the same time, large industrial customers may have competitive alternatives (e.g. self-generation of electricity, shifting production to another state with lower prices) that residential customers do not have and this sets the stage for third-degree price discrimination. Many states have special rates for low-income consumers and may have special tariffs for particular groups of customers (e.g. steel mills), reflecting income distribution and political economy considerations. Historically, income redistribution and political economy considerations played a very important role in the specification of telephone services. Local rates were generally set low and long distance rates high, just the opposite of what theories of optimal pricing would suggest [Hausman, Tardiff, and Belinfante (1993), Crandall and Hausman (2000)] and prices in rural areas where set low relative to the cost of serving these customers compared to the price cost margins in urban areas. The joint costs associated with providing both local and long distance services using the same local telephone network made it relatively easy for federal and state regulators to use arbitrary allocations of these costs to "cost justify" almost any tariff structure that they thought met a variety of redistributional and interest group politics driven goals (Salinger, 1998). Non-linear prices have been a component of regulated tariffs for electricity, gas and telephone services since these services first became available. What is clear, however, is that the formal application of the theoretical principles behind Ramsey-Boiteux pricing, non-linear pricing, and peak-load pricing has been used infre-

quently by U.S. regulators, while these concepts have been used extensively in France since the 1950s [Nelson (1964)].

7.2. The Averch-Johnson model

What has come to be known as the Averch-Johnson or "A-J" model [Averch and Johnson (1962), Baumol and Klevorick (1970), Bailey (1973)] represents an early effort to capture analytically the potential effects of rate of return regulation on the behavior of a regulated monopoly. The A-J model begins with a profit-maximizing monopoly firm with a neoclassical production function $q = F(K, L)$ and facing an inverse market demand curve $p = D(q)$. The firm invests in capital (K) with an opportunity cost of capital r (the price of capital is normalized to unity and there is no depreciation) and hires labor L at a wage w. The monopoly's profits are given by:

$$\Pi = D(q)q - wL - rK$$

It is convenient to write the firm's revenues in terms of the inputs K and L that are utilized to produce output q. Let the firms total revenue $R = R(K, L)$ and then

$$\Pi = R(K, L) - wL - rK \tag{27}$$

The regulator has one instrument at it's disposal to control the monopoly's prices. It can set the firm's "allowed rate of return" on capital s at a level greater than or equal to the firm's opportunity cost of capital r and less than the rate of return r_m that would be earned by an unregulated monopoly. The firm's variable costs wL and its capital charges sK are passed through into prices continuously and automatically without any further regulatory review or delay. The regulator has no particular objective function and is assumed only to know the firm's cost of capital r. It has no other information about the firm's production function, it's costs or its demand. The rate of return constraint applied to the firm is then given by:

$$\big[R(K, L) - wL - sK\big] \leq 0 \quad \text{where } r < s < r_m$$

or rewriting

$$\Pi \leq (s - r)K \tag{28}$$

The regulated firm is then assumed to maximizes profits (1) subject to this rate of return constraint (2). Assuming that the rate of return constraint is binding and that a solution with $q > 0$, $K > 0$ and $L > 0$ exists, the firm's constrained maximization problem becomes:

$$\underset{(K,L,\lambda)}{\text{Max}}\, \Pi^* = R(K, L) - wL - rK - \lambda\big[R(K, L) - wL - sK\big] \tag{29}$$

where λ is the shadow price of the constraint. The first order conditions are

$$\frac{\partial \Pi^*}{\partial K} = R_K - r - \lambda(R_K - s) = 0 \tag{30}$$

$$\frac{\partial \Pi^*}{\partial L} = R_L - w - \lambda(R_L - w) = 0 \tag{31}$$

$$\frac{\partial \Pi^*}{\partial \lambda} = R(K, L) - wL - sK = 0 \tag{32}$$

where R_K and R_L are the marginal revenue products of capital and labor respectively ($R_i = \mathrm{MR}_q F_i$). We can rewrite these conditions as

$$R_K = r + [\lambda/(1+\lambda)](r - s)$$

$$R_L = w$$

and (using the second order conditions) $0 < \lambda < 1$. From the first order conditions we can derive the regulated firm's marginal rate of technical substitution of capital for labor as

$$\mathrm{MRT}_{KL} = F_K/F_L = r/w + \lambda/(1+\lambda)\big[(r-s)/w\big] \tag{33}$$

This leads to the primary A-J results. A cost minimizing firm would equate the marginal rate of substitution of capital for labor to the input price ratio. Accordingly, the regulated monopoly operating subject to a rate of return constraint does not minimize costs—input proportions are distorted. Indeed with $0 < \lambda < 1$, the distortion is in a particular direction. Since $\mathrm{MRT}_{KL} < r/w$ for the equilibrium level of output the regulated firm uses too much capital relative to labor. This is sometimes referred to as the *capital using bias* of rate of return regulation. Basically, the rate of return constraint drives a wedge between the firm's actual cost of capital and its effective net cost of capital after taking account of the net benefits associated with increasing the amount of capital used when there is a net return of $(s - r)$ on the margin from adding capital, other things equal.

During the 1970s, many variations on the original A-J model appeared in the literature to extend these results. The reader is referred to Baumol and Klevorick (1970), Klevorick (1973) and Bailey (1973) for a number of these extension. Among the additional results of interest are:

(a) The A-J firm does not "waste" inputs in the sense that inputs are hired but are not put to productive use [Bailey (1973)]. The firm produces on the boundary of its production function and there is no "X-inefficiency" or waste in that sense. The inefficiency is entirely in terms of inefficient input proportions.

(b) As the allowed rate of return s approaches the cost of capital r, the magnitude of the input distortion increases [Baumol and Klevorick (1970)].

(c) There is an optimal value s* for the allowed rate of return that balances the benefits of lower prices against the increased input distortions from a lower allowed rate of return [Klevorick (1971), Sheshinski (1971)]. However, to calculate the optimal rate of return the regulator would have to know the attributes of the firm's production function, input prices and demand, information that the regulator is assumed not to possess. If the regulator did have this information she could simply calculate the optimal input proportions and penalize the firm from deviating from them.

(d) Introducing "regulatory lag" into the model (in a somewhat clumsy fashion) reduces the magnitude of the input bias [Baumol and Klevorick (1970), Bailey and Coleman (1971), Klevorick (1973)]. This is the case because if prices are fixed between rate cases, the firm can increase its profits by reducing its costs [Joskow (1974)] until the next rate review when the rate of return constraint would be applied again. If rate of return reviews are few and far between the firm essentially becomes a residual claimant on cost reductions and has powerful incentives to minimize costs. In this case, rate of return regulation has incentive properties similar to "price cap" regulation with "resets" every few years. Price cap regulation will be discussed further below.

(e) Rate of return regulation of this type can affect the profitability of peak load pricing. In particular, under certain conditions peak load pricing may reduce the firm's capital/labor ratio and it could be more profitable for the firm not to level out demand variations. However, the A-J effect could go in the other direction as well.

A lot of ink was spent on the many papers that developed variations on the A-J model and to test its implications empirically during the 1970s and 1980s. The major conceptual innovation of this literature was to highlight the possibility that regulatory mechanisms could create incentives for regulated firms to produce inefficiently and, perhaps, to adopt organization forms (e.g. vertical integration) and pricing strategies (e.g. peak load pricing) that are not optimal. Moreover, these results depend upon an extreme asymmetry of information between the regulated firm and the regulator (or just the opposite if we assume that the regulator can set the optimal s^*). In the A-J type models, the regulator knows essentially nothing about the firm's cost opportunities, realized costs, or demand. It just sets an allowed rate of return and the firm does its thing. Imperfect and asymmetric information are important attributes of regulation from both a normative and a positive perspective. However, implicitly assuming the regulators have no information is an extreme case. Beyond this, there are significant deviations between the model's assumptions (as advanced through the literature) and how regulators actually regulate. Efforts to introduce dynamics and incentive effects through regulatory lag have been cumbersome within this modeling framework. Empirical tests have not been particularly successful [Joskow and Rose (1989)]. Moreover, the particular kind of inefficiency identified by the model (inefficient input proportions) is quite different from the kind of managerial waste and inefficiency that concerns policymakers and has been revealed in the empirical literature on the effects of regulation and privatization—X-inefficiency of various types arising from imperfections in managerial efforts to minimize the costs of production, leading to production inside the production frontier and not just at the wrong location on the production frontier.

8. Incentive regulation: theory

8.1. Introduction

It should be clear by now that regulators face a number of challenges in achieving the public interest goals identified at the end of Section 3. The conventional theories of optimal pricing, production and investment by regulated firms assume that regulators are completely informed about the technology, costs and consumer demand attributes facing the firms they regulate. This is clearly not the case in reality. Regulators have imperfect information about the cost opportunities and behavior of the regulated firm and the attributes of the demand for its services that it faces. Moreover, the regulated firm generally has more information about these attributes than does the regulator or third parties which may have incentives to provide the regulator with additional information (truthful or untruthful) about the regulated firm. Accordingly, the regulated firm may use its information advantage strategically in the regulatory process to increase its profits or to pursue other managerial goals, to the disadvantage of consumers [Owen and Brauetigam (1978)]. These problems may be further exacerbated if the regulated firm can "capture" the regulatory agency and induce it to give more weight to its interests [Posner (1974), McCubbins (1985), Spiller (1990), Laffont and Tirole (1993, Chapter 5)]. Alternatively, other interest groups may be able to "capture" the regulator and, in the presence of long-lived sunk investments, engage in "regulatory holdups" or expropriation of the regulated firm's assets. Higher levels of government, such as the courts and the legislature, also have imperfect information about both the regulator and the regulated firm and can monitor their behavior only imperfectly [McNollgast (Chapter 22 in this handbook)].

The evolution of regulatory practices in the U.S. reflects efforts to mitigate the information disadvantages that regulators must deal with, as well as broader issues of regulatory capture and monitoring by other levels of government and consumers. As already noted, these institutions and practices are reflected in laws and regulations that require firms to adhere to a uniform system of accounts, give regulators access to the books and records of the regulated firm and the right to request additional information on a case by case basis, auditing requirements, staff resources to evaluate this information, transparency requirements such as public hearings and written decisions, ex parte communications rules, opportunities for third parties to participate in regulatory proceedings to (in theory)[19] assist the regulatory agency in developing better information and reducing its information disadvantage, appeals court review, and legislative oversight processes. In addition, since regulation is a repeated game, the regulator (as well as legislators and appeals courts) can learn about the firm's attributes as it observes its behavioral and performance responses to regulatory decisions over time and, as a result,

[19] Of course, third parties may have an incentive to inject inaccurate information into the regulatory process as well.

the regulated firm naturally develops a reputation for the credibility of its claims and the information that it uses to support them. However, although U.S. regulatory practice focused on improving the information available to regulators, the regulatory mechanisms adopted typically did not utilize this information as effectively as they could have until relatively recently.

The A-J model and its progeny are, in a sense, the first crude analytical efforts to understand how, when regulators are poorly informed and have limited instruments at their disposal, the application of particular mechanisms to constrain the prices charged by a regulated firm may create incentives for a firm to respond in ways that lead to inefficiencies in other dimensions; in the AJ-type models to depart from cost-minimizing input proportions with a bias towards using more capital. However, in the A-J model the regulator has essentially no information about the regulated firm's costs or demand, there is no specification of the objectives of and incentives faced by the firm's managers that might lead the firm to exhibit inefficiencies in other dimensions, the instruments available to the regulatory are very limited, and indeed the choice of mechanisms by the regulator does not flow from a clear specification of managerial objectives and constraints.

More recent work on the theory of optimal incentive regulation deals with asymmetric information problems, contracting constraints, regulatory credibility issues, dynamic considerations, regulatory capture, and other issues that regulatory processes have been trying to respond to for decades much more directly and effectively [Laffont and Tirole (1993), Armstrong, Cowan, and Vickers (1994), Armstrong and Sappington (2003a, 2003b)]. This has been accomplished by applying modern theories of the firm, incentive mechanism design theory, auction theory, contract theory, and modern political economy in the context of adverse selection, moral hazard, hold-up and other considerations, to derive optimal (in a second best sense) mechanisms to achieve public interest regulatory goals. This has become a vast literature; some of which is relevant to actual regulatory problems and practice, though much of it is not.

Let us start with the simplest characterization of the nature of the regulator's information disadvantages. A firm's costs may be high or low based on inherent attributes of its technical production opportunities, exogenous input cost variations over time and space, inherent differences in the costs of serving locations with different attributes (e.g. urban or rural), etc. While the regulator may not know the firm's true cost opportunities she will typically have some information about them. The regulator's imperfect information can be summarized by a probability distribution defined over a range of possible cost opportunities between some upper and lower bound within which the regulated firms actual cost opportunities lie. Second, the firm's actual costs will not only depend on its underlying cost opportunities but also on the behavioral decisions made by managers to exploit these cost opportunities. Managers may exert varying levels of effort to get more (or less) out of the cost opportunities that the firm has available to it. The greater the managerial effort the lower will be the firm's costs, other things equal. However, exerting more managerial effort imposes costs on managers and on society. Other things equal, managers will prefer to exert less effort than more to increase their own satisfac-

tion, but less effort will lead to higher costs and more "x-inefficiency." Unfortunately, the regulator cannot observe managerial effort directly and may be uncertain about its quality and impacts on the regulated firm's costs and quality of service.

The uncertainties the regulator faces about the firm's inherent cost opportunities gives the regulated firm a strategic advantage. It would like to convince the regulator that it is a "higher cost" firm that it actually is, in the belief that the regulator will then set higher prices for service as it satisfies the firm's long-run viability constraint (firm participation or budget-balance constraint), increasing the regulated firm's profits, creating dead-weight losses from (second-best) prices that are two high, and allowing the firm to capture social surplus from consumers. Thus, the social welfare maximizing regulator faces a potential *adverse selection* problem as it seeks to distinguish between firms with high cost opportunities and firms with low cost opportunities while adhering to the firm viability or participation constraint.

The uncertainties that the regulator faces about the quantity and impact of managerial effort creates another potential problem. Since the regulator typically has or can obtain good information about the regulated firm's actual costs (i.e. it's actual expenditures), at least in the aggregate, one approach to dealing with the adverse selection problem outlined above would simply be to set (or reset after a year) prices equal to the firm's realized costs ex post. This would solve the adverse selection problem since the regulator's information disadvantage would be resolved by auditing the firm's costs.[20] However, if managerial effort increases with the firm's profitability, this kind of "cost plus" regulation may lead management to exert too little effort to control costs, increasing the realized costs above their efficient levels. If the "rat doesn't smell the cheese and sometimes get a bit of it to eat" he may play golf rather than working hard to achieve efficiencies for the regulated firm. Thus, the regulator faces a potential *moral hazard* problem associated with variations in managerial effort in response to regulatory incentives [Laffont and Tirole (1986), Baron and Besanko (1987b)].

Faced with these information disadvantages, the social welfare maximizing regulator will seek a regulatory mechanism that takes the social costs of adverse selection and moral hazard into account, subject to the firm participation or budget-balance constraint that it faces, balancing the costs associated with adverse selection and the costs associated with moral hazard. The regulator may also take actions that reduce her information disadvantages by, for example, increasing the quality of the information that the regulator has about the firm's cost opportunities.

Following Laffont and Tirole [(1993, pp. 10–19)], to illuminate the issues at stake we can think of two polar case regulatory mechanisms that might be applied to a monopoly firm producing a single product. The first regulatory mechanism involves setting a fixed price ex ante that the regulated firm will be permitted to charge going forward. Alternatively, we can think of this as a pricing *formula* that starts with a particular price and

[20] Of course, the auditing of costs may not be perfect and in a multiproduct context the allocation of accounting costs between different products is likely to reflect some arbitrary joint cost allocation decisions.

then adjusts this price for *exogenous* changes in input price indices and other exogenous indices of cost drivers. This regulatory mechanism can be characterized as a *fixed price* regulatory contract or a *price cap* regulatory mechanism. There are two important attributes of this regulatory mechanism. Because prices are fixed (or vary based only on exogenous indices of cost drivers) and do not respond to changes in managerial effort, the firm and its managers are the residual claimants on production cost reductions and the costs of increases in managerial effort (and vice versa). That is, the firm and its managers have the highest powered incentives fully to exploit their cost opportunities by exerting the optimal amount of effort [Brennan (1989), Cabral and Riordan (1989), Isaac (1991), Sibley (1989), Kwoka (1993)]. Accordingly, this mechanism provides optimal incentives for inducing managerial effort and eliminates the costs associated with managerial moral hazard. However, because the regulator must adhere to a firm participation or viability constraint, when there is uncertainty about the regulated firm's cost opportunities the regulator will have to set a relatively high fixed price to ensure that if the firm is indeed inherently high cost, the prices under the fixed price contract or price cap will be high enough to cover the firm's costs. Accordingly, while the fixed price mechanism may deal well with the potential moral hazard problem by providing high powered incentives for cost reduction, it is potentially very poor at "rent extraction" for the benefit of consumers and society, potentially leaving a lot of rent to the firm due to the regulator's uncertainties about the firm's inherent costs and its need to adhere to the firm viability or participation constraint. Thus, while a fixed price contract solves the moral hazard problem it incurs the full costs of adverse selection.

At the other extreme, the regulator could implement a "cost of service" contract or regulatory mechanism where the firm is assured that it will be compensated for all of the costs of production that it incurs. Assume for now that this is a credible commitment—there is no ex post renegotiation—and that audits of costs are accurate. When the firm produces it will then reveal whether it is a high cost or a low cost firm to the regulator. Since the regulator compensates the firm for all of its costs, there is no "rent" left to the firm as excess profits. This solves the adverse selection problem. However, this kind of cost of service recovery mechanism does not provide any incentives for the management to exert effort. If the firm's profitability is not sensitive to managerial effort, the managers will exert the minimum effort that they can get away with. While there are no "excess profits" left on the table, consumers are now paying higher costs than they would have to pay if the firm were better managed. Indeed, it is this kind of managerial slack and associated x-inefficiencies that most policymakers have in mind when they discuss the "inefficiencies" associated with regulated firms. Thus, the adverse selection problem can be solved in this way, but the costs associated with moral hazard are fully realized.

Accordingly, these two polar case regulatory mechanisms each has benefits and costs. One is good at providing incentives for managerial efficiency and cost minimization, but it is bad at extracting the benefits of the lower costs associated with single firm production for consumers when costs are subadditive. The other is good at rent extraction but leads to costs from moral hazard. Perhaps not surprisingly, the optimal regulatory

mechanism (in a second best sense) will lie somewhere between these two extremes. In general, it will have the form of a *profit sharing* contract or a *sliding scale* regulatory mechanism where the price that the regulated firm can charge is *partially* responsive to changes in realized costs and *partially* fixed ex ante [Schmalensee (1989b), Lyon (1996)]. As we shall see, by offering a menu of regulatory contracts with different cost sharing provisions, the regulatory can do even better than if it offers only a single profit sharing contract [Laffont and Tirole (1993)]. The basic idea here is to make it profitable for a firm with low cost opportunities to choose a relatively high powered incentive scheme and a firm with high cost opportunities a relatively low-powered scheme. Some managerial inefficiencies are incurred if the firm turns out to have high cost opportunities, but these costs are balanced by reducing the rent left to the firm if it turns out to have low cost opportunities.

We can capture the nature of the range of options in the following fashion. Consider a general formulation of a regulatory process in which the firm's revenue requirements "R" are determined based on a fixed component "a" and a second component that is contingent on the firm's realized costs "C" and where "b" is the sharing parameter that defines the responsiveness of the firm's revenues to realized costs.

$$R = a + (1 - b)C$$

Under a fixed price contract or price cap regulation:

$$a = C^*$$

where C^* is the regulators assessment of the "efficient" costs of the highest cost type

$$b = 1$$

Under cost of Service regulation:

$$a = 0$$

$$b = 0$$

Under profit sharing contract or sliding scale regulation (Performance Based Regulation)

$$0 < b < 1$$

$$0 < a < C^*$$

These different mechanisms then have the properties summarized in Table 1.

The challenges then are to find the optimal performance based mechanism given the information structure faced by the regulator and for the regulator to find ways to reduce its information disadvantages vis a vis the regulated firm and to use the additional information effectively. As we shall see, it is optimal for the regulator to offer a menu of contracts with different combinations of a and b that meet certain conditions driven by the firm participation constraint and an incentive compatibility constraint that leads

Table 1
Incentives vs. Rent Extraction

Mechanism	Managerial Incentives	Rent Extraction
Fixed Price	100%	0%
Cost of Service	0	100%
Performance Based	$0 < x < 100\%$	$0 < y < 100\%$

firms with low cost opportunities to choose a high powered scheme (b is closer to 1 and a is closer to the efficient cost level for a firm with low cost opportunities) and firms with high cost opportunities to choose a lower powered incentive scheme (a and b are closer to zero). The lower powered scheme is offered to satisfy the firm participation constraint, sacrificing some costs associated with moral hazard, in order to reduce the rents that must be left to the high cost as it is induced to exert the optimal amount of managerial effort. (So far, this discussion has ignored quality issues. Clearly if a regulatory mechanism focuses only on reducing costs and ignores quality it will lead to firm to provide too little quality. This is a classic problem with price cap mechanisms and will be discussed further below.)

8.2. *Performance Based Regulation typology*

As I have already indicated, there is a very extensive theoretical literature on incentive regulation, or as it is commonly called by policymakers, performance based regulation or PBR. The papers that comprise this literature reflect a wide range of assumptions about the nature of the information possessed by the regulator and the firm about costs, cost reducing managerial effort, demand and product quality, the attributes of the regulatory instruments available to the regulator, the risk preferences of the firm, regulatory capture by interest groups, regulatory commitment, flexibility, and other dynamic considerations. These alternative sets of assumption can be applied in both a single or multiproduct context. One strand of the literature initially focused primarily on adverse selection problems motivated by the assumption that regulators could not observe a firm's costs and ignoring the role of managerial effort [Baron and Myerson (1982), Lewis and Sappington (1988a, 1988b)]. Another strand of the literature focused on both adverse selection and moral hazard problems motivated by the assumption that regulators could observe a firm's realized cost ex post, had information about the probability distribution of a firm's cost ex ante, and that managerial effort did affect costs but that this effort was not observable by the regulator [Laffont and Tirole (1986)]. Over time, these approaches have evolved to cover a similar range of assumptions about these basic information and behavioral conditions and lead to qualitatively similar conclusions. Armstrong and Sappington (2003a, forthcoming) provides a detailed and thoughtful review and synthesis of this entire literature and I refer readers interested in a very detailed treatment of the full range specifications of incentive regulation problems to their paper. Here I will simply lay out a "typology" of how these issues have been developed

in the literature and then provide some simple theoretical examples to illustrate what I consider to be the literature's primary conclusions of potential relevance for regulation in practice.

What are the regulator's objectives? Much of the literature assumes that the regulator seeks to maximize a social welfare function that reflects the goal of limiting the rents that are transferred from consumers and taxpayers to the firm's owners and managers subject to a firm participation or breakeven constraint. Armstrong and Sappington (2003a, 2003b) articulate this by specifying an objective function $W = S + \alpha R$ where W is expected social welfare, S equals expected consumers' (including consumers as taxpayers) surplus, R equals the expected rents earned by the owners and managers of the firm (over and above what is needed to compensate them for the total costs of production and the disutility of managerial effort to satisfy the participation constraint), and where $\alpha < 1$ implies that the regulator places more weight on consumers surplus than on rents earned by the firm. That is, the regulator seeks to extract rent from the firm for the benefit of consumers, subject as always to a firm participation or break-even constraint. In addition, W will be reduced if excessive rents are left to the firm since this will require higher (second-best) prices and greater allocative inefficiency.

Laffont and Tirole (1988a, 1988b, 1993, 2000) create a social benefit from reducing the rents left to the firm in a different way. In their basic model, consumer welfare and the welfare of the owners and managers of the firm are generally weighted equally. However, one of the instruments available to the regulator is the provision of transfer payments to the firm which affect the rents earned by the firm. These transfer payments come out of the government's budget and carry a social cost resulting from the inefficiencies of the tax system used to raise these revenues. Thus, for every dollar of transfer payments given to the firm to increase its rent, effectively $(1 + k)$ dollars of taxes must be raised, where k reflects the inefficiency of the tax system. Accordingly, by reducing the transfers to the firm over and above what is required to compensate it for its efficient production costs and the associated managerial disutility of effort, welfare can be increased. This set-up which allows for the use of costly government transfer payments also leads to a nice dichotomy between incentive arrangements that effectively establish the formula for determining the firm's revenues in a way that deals with adverse selection and moral hazard problems in the context of asymmetric information and price setting which establishes the second-best prices for the services sold by the firm given consumer demand attributes and the regulator's knowledge of them. That is, regulators first establish compensation arrangements (define how the firm's budget constraint or "revenue requirements" will be defined) to deal as effectively as possible with adverse selection and moral hazard problems given the information structure assumed. The regulator separately establishes a second best price structure to deal with allocational efficiency considerations which may not cover all of the firm's costs, with the difference coming from net government transfers. In addition, Laffont and Tirole (1993) introduce managerial effort as a variable that affects costs and service quality. Managers have a disutility of effort and must be compensated for it. Accordingly, the utility of management also appears in the social welfare function.

What does the regulator know about the firm ex ante and ex post? In what follows I will use the term "cost" to refer to the firm's marginal costs and ignore fixed costs (or normalize them to zero). This allows me to ignore in the discussion of incentive issues in this section, the second-best pricing (rather than incentive) options available to deal with budget-balance constraints created by increasing returns since the major issues associated with these pricing problems have been discussed above and are not affected in important ways by introducing asymmetric information about the firm's costs and managerial effort into the analysis. Carrying these issues forward here would simply complicate the presentation of the key incentive regulation results of interest. Accordingly, in what follows the full information benchmark is marginal cost pricing with zero rents left for the firm. Armstrong and Sappington (2003a, 2003b) distinguish between fixed costs and marginal costs, what the regulator knows about each and allow the regulator to make non-distortionary lump sum transfers to the firm. In this context, if the regulator can only make distortionary transfer payments, the full information benchmark with linear prices is Ramsey-Boiteux pricing and otherwise it is the optimal non-linear prices given the regulator's information about consumer demand.

The literature that focuses on adverse selection builds on the fundamental paper by Baron and Myerson (1982). The regulator does not know the firm's cost opportunities ex ante but has information about the probability distribution over the firm's possible cost opportunities.[21] Nor can the regulator observe or audit the firm's costs ex post. The firm does know its own cost opportunities ex ante and ex post. The firm's demand is known by both the regulator and the firm. There is no managerial effort in these early models. Accordingly, the analysis deals with a pure adverse selection problem with no potential inefficiencies or moral hazard associated with inadequate managerial effort. With moral hazard alone only high-powered incentive mechanisms are optimal. The regulator in the presence of adverse selection literature then proceeds to consider asymmetric information about the firm's demand function, where the firm knows its demand but either the regulator does not observe demand ex ante or ex post or learns about demand only ex post [Lewis and Sappington (1988a), Riordan (1984)]. Combining asymmetric information about costs and demand, introducing a multidimensional characterization of asymmetric information, is then a natural extension of the regulation to respond to adverse selection literature [Lewis and Sappington (1988b, 1989), Dana (1993), Armstrong and Rochet (1999)].

In light of common regulatory practice, a natural extension of these models is to assume that the regulated firm's actual realized costs are observable ex post, at least with uncertainty. Baron and Besanko (1984) considers cases where a firm's costs are "audited" ex post, but the actual realized costs resulting from the audit are observable by the regulator with a probability less than one. The regulator can use this information to reduce the costs of adverse selection. Laffont and Tirole (1986, 1993) consider cases

[21] In models that distinguish between fixed and variable costs, the regulator may know the fixed costs but not the variable costs. See Armstrong and Sappington (2003a, 2003b).

where the firm's realized costs are fully observable by the regulator. However, absent the simultaneous introduction of an uncertain scope for cost reductions through managerial effort, the regulatory problem then becomes trivial—just set prices equal to the firm's realized costs. Accordingly, Laffont and Tirole (1986a, 1993) introduce managers of the firm who can choose the amount of cost reducing effort that they expend. Managerial effort is not observable by the regulator ex ante or ex post, but realized production costs are fully known to the regulator as is the managerial "production function" that transforms managerial effort into cost reductions and the managers' utility over effort function. The regulated firm fully observes managerial effort, the cost reducing effects of managerial effort, and demand. It also knows what managerial utility would be at different levels of effort. Armstrong and Sappington (2003a, 2003b) advance this analysis by considering cases where the regulated firm is uncertain about the operating costs that will be realized but knows that it can reduce costs by increasing managerial effort, though in a way that creates a moral hazard problem but no adverse selection problem. In the face of uncertainty over its costs, they consider cases where the firm may be either risk-neutral or risk averse.

The literature also examines situations in which the regulator is *captured* by an interest group and no longer seeks to maximize social welfare W. For example, the regulator may be bribed not to use or reveal information that would reduce the rents available to the firm [Laffont and Tirole (1993, Chapter 11)] and the regulator may effectively collude with the firm if she can be compensated in some way (monetary, future employment, jobs for friends and relatives) for doing so. The possibility of regulatory capture may affect the choice of the power of the incentive schemes used by the regulator. High powered incentive schemes are more susceptible to regulatory capture than are lower powered schemes [Laffont and Tirole (1993, pp. 57–58)]. To counteract the possibility of regulatory collusion, the analysis can also be expanded to include another level in which the government imposes an incentive scheme on the regulator to provide incentives to reveal and use all relevant information possessed by the regulator and more generally not to collude with interest groups.

What instruments are available to the regulator and how do the regulator and the regulated firm interact over time? Much of the incentive regulation literature is static. The regulator (or the government through the regulator) can offer a menu of prices (or fixed price contracts) with or without a fixed fee or transfer payment. The menu may contain prices that are contingent on realized costs (which can be thought of as penalties or rewards for performance) in those models where regulators observe costs ex post. Some of these instruments may be costly to utilize (e.g. transfer payments and auditing efforts). The more instruments the regulator has at its disposal and the lower the costs of using them, the closer the regulator will be able to get to the full information benchmark.

Of more interest are issues that arise as we consider the dynamic interactions between the regulated firm and the regulator and the availability and utilization of mechanisms that the regulator potentially has available to reduce its information disadvantage. It is inevitable that the regulator will learn more about the regulated firm as they interact over

time. So, for example, if the regulator can observe a firm's realized costs ex post, should the regulator use that information to reset the prices that the regulated firm receives [commonly known as a "ratchet"—Weitzman (1980)]? Or is it better for the regulator to commit to a particular contract ex ante, which may be contingent on realized costs, but the regulator is not permitted to use the information gained from observing realized costs to change the terms and conditions of the regulatory contract offered to the firm? Is it credible for the regulator to commit *not* to renegotiate the contract, especially in light of U.S. regulatory legal doctrines that have been interpreted as foreclosing the ability of a regulatory commission to bind future commissions?

Clearly, if the regulated firm knows that information about its realized costs can be used to renegotiate the terms of its contract, this will affect its behavior ex ante. It may have incentives to engage in less cost reduction in period 1 or try to fool the regulator into thinking it is a high cost firm so that it can continue to earn rents in period 2. Of if the regulated firm has a choice between technologies that involve sunk cost commitments, will the possibility of ex post opportunism or regulatory expropriation, perhaps driven by the capture of the regulator by other interest groups, affect its willingness to invest in the lowest cost technologies when they involve more significant sunk cost commitments (leading to the opposite of the A-J effect).

These dynamic issues have been examined more intensively over time and represent a merging of the literature on regulation with the literature on contracts and dynamic incentive mechanisms more generally [Laffont and Tirole (1988b, 1990a, 1993), Baron and Besanko (1987a), Armstrong and Vickers (1991, 2000), Armstrong, Cowan, and Vickers (1994)]. The impacts of regulatory lag of different durations [Baumol and Klevorick (1970), Klevorick (1973), Joskow (1974)] and other price adjustment procedures have been analyzed extensively as well [Vogelsang and Finsinger (1979), Sappington and Sibley (1988, 1990)].

8.3. Some examples of incentive regulation mechanism design

This section is based on Laffont and Tirole (1993, Chapter 2). We will examine the case of a regulated monopoly firm producing a single private good and which is restricted to charging linear prices. A specific firm's cost opportunities depend on the best technology and input prices that it has access to and which will characterize it's "type" denoted by β. The firm knows its type but the regulator is uncertain about the firm's type. We will begin with a two-type case where the firm can be either a low cost type denoted by β_L with probability v or a high cost type β_{H} with probability $1 - v$. The firm's management can exert effort e but managerial utility declines as effort increases. The firm's cost function is then given by:

$$C(q) = F + (\beta - e)q$$

Assume that F is known by the regulator and we normalize it to zero for simplicity. The regulator cannot observe β or e, but can observe the firm's actual production costs ex

post. Then the firm's marginal cost is given by

$$c = (\beta - e)$$

and the disutility of managers with respect to effort e is defined as

$$U = t - \psi(e)$$

where $\psi'(e) > 0$. This function is known to the firm and to the regulator, but the regulator cannot observe e or U directly.

Define:

$S(q) =$ gross consumers' surplus

$q = D(p) =$ market demand curve for the product

$P = P(q) =$ inverse market demand curve for the product

$R(q) = qP(q) =$ market revenue generated by the firm

Laffont and Tirole (1993) allow the government to make financial transfers to the firm with a social cost of λ per dollar transferred, so that to transfer one dollar to the firm costs the government (and society) $(1 + \lambda)$ dollars. To keep the accounting straight we adopt Laffont and Tirole's accounting convention. All revenues from sales of the product go to the government and then the government reimburses the firm for its actual production costs plus an additional transfer payment that is greater than or equal to zero. Thus, the firm's costs are covered (breakeven constraint is satisfied) and the (net) transfer payment t must be large enough at least to compensate the managers for the disutility of effort to satisfy the participation constraint. Then social welfare W is given by:

$$\begin{aligned} W &= V(q) - (1+\lambda)(C + t) + U \\ &= V(q) - (1+\lambda)(C + \psi(e)) - \lambda U \end{aligned} \tag{34}$$

where:

$$\begin{aligned} V(q) &= \big[S(q) - R(q)\big] + (1+\lambda)R(q) \\ &= S(q) + \lambda q P(q) \end{aligned}$$

The full information benchmark is then derived as follows:

$$\max_{(e,q)} W = S(q) + \lambda q P(q) - (1+\lambda)\big[(\beta - e)q + \psi(e)\big] - \lambda U \quad \text{s.t. } U \geq 0 \tag{35}$$

The first order conditions are:

$$U = 0 \text{ [no rent left to the firm/managers, but participation constraint is satisfied]} \tag{36}$$

$$\psi'(e) = q \text{ [marginal disutility of effort equals marginal cost savings from additional effort]} \tag{37}$$

$$P(q) + \lambda P(q) + \lambda q P'(q) = (1 + \lambda)(\beta - e) = (1 + \lambda)c$$

or

$$(p(q) - c)/p(q) = \left[\lambda/(1 + \lambda)\right](1/\eta) \text{ [Ramsey-Boiteux pricing]} \tag{38}$$

where η is the elasticity of demand for the product supplied by the regulated firm.

Condition (38) requires some explanation. It looks like the Ramsey-Boiteux pricing formula that we discussed earlier and, in a sense, it is. However, here λ is not the shadow price of the firm's budget constraint but rather the marginal cost of raising government revenues through the tax system and then distributing government revenues to the firm to cover its costs and a transfer payment to compensate managers for their disutility of effort. The optimal prices here serve a pure social allocation function that take into account the cost of using public funds to compensate the firm for its costs and managers for their disutility of effort. These "Ramsey-Boiteux prices" are equivalent to adding the optimal commodity taxes to the marginal cost of supplying these services. This is the essence of Laffont and Tirole's separation of or dichotomy between "incentives to deal with moral hazard and adverse selection" and "prices" to deal with consumption allocational considerations.

To summarized, with full information, the regulator would compensate the firm for its costs and the manager's disutility of effort leaving no rents to the firm (36). It would also require the managers of the firm to exert the optimal effort e^*, which in turn yields the optimal level of total and marginal costs (b). Let $q^*(c)$ denote the solution to (37) and (38) and call it the Ramsey-Boiteux output. Then $P(q^*(c))$ is the Ramsey-Boiteux price.

Now we consider the characteristics of (second-best) optimal regulatory mechanism when there is asymmetric information. Everything is common knowledge except the regulator cannot observe the firm's type β or the quantity of managerial effort e expended. In the most simple case, the regulator does know that the firm is either a high cost type with β_L with probability v or a high cost type β_{H} with probability $(1 - v)$. The attributes of the optimal regulatory mechanism are then derived by maximizing expected social welfare given the probability of each type subject to a firm viability constraint ($U \geq 0$) for each type and an incentive compatibility constraint that ensures that each type chooses the regulatory contract that is optimal given asymmetric information. Laffont and Tirole (1993) show that the binding incentive compatibility constraint is given by the low-cost type's rent which in turn is determined by the high cost type's marginal cost. Basically, the contract designed for the high cost type leaves no rent to the high-cost firm and its managers while the contract designed for the low cost type must leave enough rent to the low cost type so that it does not choose the contract designed for the high cost type. This rent is the difference in their realized marginal costs at the effort levels they choose given the contract they take up.

The expected welfare seen by the regulator is

$$\begin{aligned}\underset{(U_L,U_{\mathrm{H}},q_{\mathrm{H}},q_L,e_{\mathrm{H}},e_L)}{\operatorname{Max}} W &= v\big[V(q_L) - (1+\lambda)\big[(\beta_L - e_L)q + \psi(e_L)\big] - \lambda\Phi(e_{\mathrm{H}})\big] \\ &\quad + (1-v)\big[V(q_{\mathrm{H}}) - (1+\lambda)\big[(\beta_{\mathrm{H}} - e_{\mathrm{H}})q_{\mathrm{H}} + \psi(e_{\mathrm{H}})\big]\big]\end{aligned} \tag{39}$$

(subject to firm participation constraints and incentive compatibility constraints) where $\Phi(\cdot)$ is an increasing function of $e = \psi(e) - \psi(e - (\beta_{\mathrm{H}} - \beta_L))$. Maximizing expected welfare subject to the firm participation and incentive compatibility constraints yields the first order conditions are:

$$q_L = q^*(\beta_L - e_L) \tag{40}$$

$$\psi'(e_L) = q_L \tag{41}$$

$$q_{\mathrm{H}} = q^*(\beta_{\mathrm{H}} - e_{\mathrm{H}}) \tag{42}$$

$$\psi'(e_{\mathrm{H}}) = q_{\mathrm{H}} - \big[\lambda/(1+\lambda)\big]\big[v/(1-v)\big]\Phi'(e_{\mathrm{H}}) \tag{43}$$

First order conditions (41) and (42) are simply the Ramsey-Boiteux quantities given the realization of marginal cost and the associated Ramsey-Boiteux prices are optimal for each type. That is, they are the same as under full information. First order condition (41) shows that the optimal contract for the low-cost type will induce the low cost type to exert the optimal amount of effort as it would under full information. First order condition (43) shows that the effort exerted by the high cost type will be less than optimal. The firm participation constraint is also binding for the high cost type ($U = 0$) but not for the low-cost type ($U > 0$). Thus, while the low cost type chooses the optimal amount of effort, it gains an information rent $U > 0 = \Phi(\beta_{\mathrm{H}} - \beta_L)$. The reason that the effort of the high cost type is optimally distorted from the full information optimal level is to reduce the rent that must be left to the low cost type to satisfy the incentive compatibility constraint which is binding for the high cost type. Reducing e_{H} by a small amount has two effects. It reduces the disutility of effort and increases the cost of production. The net effect on the firm's unit cost, including managerial disutility of effort, is $1 - \psi(e_{\mathrm{H}})$. But this also reduces the rent that must be left to the low cost firm by $\Phi'(e_{\mathrm{H}})$ to satisfy the incentive compatibility constraint. So the expected increase in the net unit cost to the high cost firm are $(1-v)(1+\lambda)(1-\psi(e_{\mathrm{H}}))$ and the reduction in the unit cost of rent transfers to the low cost firm is $v\lambda\Phi'(e_L)$. The amount of the distortion in e_L is then chosen to equate these costs on the margin.

The optimal regulatory mechanism involves offering the regulated firm a choice between two regulatory contract options. One is a fixed price option that leaves some rent if the firm is a low-cost type but negative rent if it is a high cost type. The second is a cost-contingent contract that distorts the firm's effort if it is a high cost type but leaves it no rent. The high powered scheme is the most attractive to the low-cost type and the low-powered scheme is the most attractive to the high cost type. The expected cost of the distortion of effort if the firm is a high cost type is balanced against the expected cost of leaving additional rent top the firm if it is a low cost type—*the fundamental tradeoff between incentives and rent extraction*.

The two-type example can be generalized to a continuum of types [Laffont and Tirole (1993, pp. 137ff)]. Here we assume that β has a continuous distribution from some lower bound β_L to some upper bound β_H with a cumulative distribution $F(\beta)$ and a strictly positive density $f(\beta)$ where F is assumed to satisfy a monotone hazard rate condition and $F(\beta)/f(\beta)$ is non-decreasing in β. The regulator maximizes expected social welfare subject to the firm participation and incentive compatibility constraints as before and incentive compatibility requires a mechanism that leaves more rent to the firm the lower is its type β, with the highest cost type getting no rent, the lowest cost type getting the most rent and intermediate type's rent defined by the difference in their marginal costs. Similarly, the effort of the lowest cost type is optimal and the effort of the highest cost type is distorted the most, with intermediate types having smaller levels of distortion (and more rents) as β declines toward β_L. In the case of a continuous distribution of types, the optimality conditions are directly analogous to those for the two-type case.

$$q(\beta) = q^*\big(\beta - e(\beta)\big) \quad \text{[Ramsey Pricing]} \tag{44}$$

$$\Psi'(e(\beta)) = q(\beta) - \big[\lambda/(1+\lambda)\big]\big[F(\beta)/f(\beta)\big]\Psi''(e(\beta)) \tag{45}$$

Where (44) shows that Ramsey pricing is optimal given realized costs and (45) shows that effort is distorted as β increases to constrain the rents that are left to lower cost firms.

Laffont and Tirole (1993) show that these optimality conditions can be implemented by offering the firm a menu of linear contracts, which in their model are transfer or incentive payments in excess of realized costs (which are also reimbursed), of the form:

$$t(\beta, c) = a(\beta) - b(\beta)c$$

where a is a fixed payment, b is a cost contingent payment, and a and b are decreasing in β.

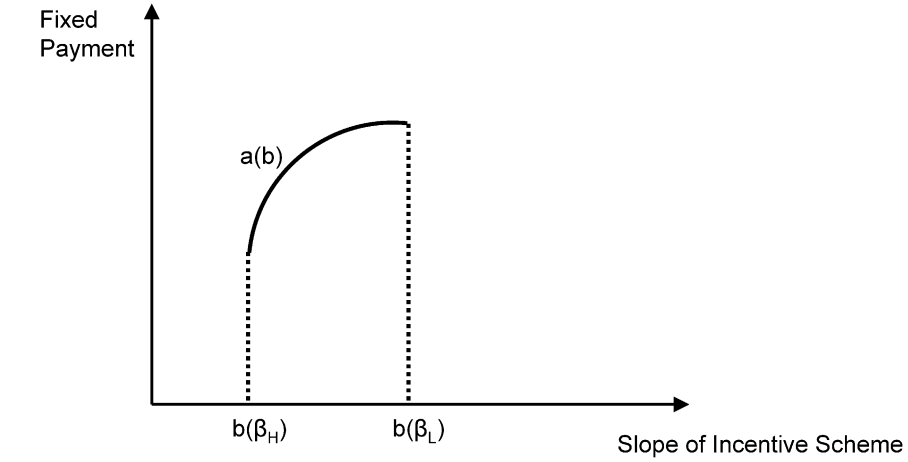
We can rewrite the transfer payment equation in terms of the gross transfer to the firm including the unit cost reimbursement:

$$R_f = a(\beta) - b(\beta)c + c = a(\beta) + \big(1 - b(\beta)\big)c \tag{46}$$

where $\mathrm{d}a/\mathrm{d}b > 0$ (for a given β a unit increase in the slope of the incentive payment must be compensated by an increase in the fixed payment to cover the increase in production costs) and $\mathrm{d}^2a/\mathrm{d}b^2 < 0$ (the fixed payment is a concave function of the slope of the incentive scheme).

Figure 11 displays this relationship. The lowest cost type chooses a fixed price contract with a transfer net of costs equal to U_L and the firm is the residual claimant on cost reducing effort ($b = 1$). As β increases, the transfer is less sensitive to the firm's realized costs (b declines) and the rent is lower (a declines).

Note that if one were to try empirically to relate the firms' realized costs to the power of the incentive scheme they had selected, a correlation between the power of the contract and the firm's realized costs would not tell us anything directly about the



Source: Laffont and Tirole (1993), Figure 1.5

Figure 11. Menu of incentive contracts.

incentive effects of higher-powered schemes in terms of inducing optimal effort and mitigating moral hazard problems. This is the case because the firms with the lower inherent costs will rationally choose the higher powered contracts. Assume that we had data for regulated firms serving different geographic regions (e.g. different states) which had different inherent cost opportunities (a range of possible values for β). If the regulators in each state offered the optimal menu of incentive contracts, the low β firms would choose high powered contracts and the high β firms would choose lower powered contracts. Accordingly, the effects of the mechanisms on mitigating the rents that would accrue to a low cost firm's information advantage from the effects on inducing optimal effort are not easily distinguished. I know of only one empirical paper that has endeavored to tackle this challenge directly [Gagnepain and Ivaldi (2002)] and it is discussed further below.

8.3.1. The value of information

This framework also provides us with insights into the value to the regulator of reducing her information disadvantage. Consider the two-type case. Let's say that the regulator is able to obtain information that increases her assessment of the probability that the firm is a low cost type from v to v_{H}. If the regulator's assessment of v increases there are two effects. The first effect is that the rent left to the low cost type falls. By increasing v, more weight in the social welfare function is placed on the realization of the firm being a low cost type and this increases the expected cost of rent transfers other things equal. Similarly, the optimal distortion induced in the high-cost type increases since less weight is placed on this realization in the expected welfare function. These

intuitive results carry through for the continuous type case. Overall, as the regulator's information becomes more favorable as defined in Laffont and Tirole (1993, pp. 76–81), the higher is welfare even though for a given realization of β we will observe firm's choosing (being offered) a lower powered incentive scheme. This latter result, does not appear to be generalizable to models where there are no government transfers and where revenues the firm earns from sales must be relied upon entirely to achieve both incentive goals (adverse selection and moral hazard) and allocational goals [Armstrong, Cowan, and Vickers (1994, pp. 39–43), Schmalensee (1989b)]. Without government transfers as an instrument, if the regulator's uncertainty about firm types declines she will choose a higher powered scheme, because the budget balance constraint effectively becomes less binding, allowing the regulator to tolerate more variation in a firm's realized net revenues.

One way in which regulators can effectively reduce their information advantage is by using competitive benchmarks or "yardstick regulation" in the price setting process. Schleifer (1985) shows that if there are $n > 1$ non-competing but otherwise identical firms (e.g. gas distribution companies in firms in different states), an efficient regulatory mechanism involves setting the price for each firm based on the costs of the other firms. Each individual firm has no control over the price it will be allowed to charge (unless the firms can collude) since it is based on the realized costs of $(n - 1)$ other firms. So, effectively each firm has a fixed price contract and the regulator can be assured that the budget balance constraint will be satisfied since if the firms are identical prices will never fall below their "efficient" realized costs. This mechanism effectively induces each firm to compete against the others. The equilibrium is a price that just covers all of the firm's efficient costs as if they competed directly with one another.

Of course, it is unlikely to be able to find a large set of truly identical firms. However, hedonic regression, frontier cost function estimation and related statistical techniques can be used to normalize cost variations for exogenous differences in firm attributes to develop normalized benchmark costs [Jamasb and Pollitt (2001, 2003), Estache, Rossi, and Ruzzier (2004)]. These benchmark costs can then be used by the regulator in a yardstick framework or in other ways to reduce its information advantage, allowing it to use high powered incentive mechanisms without incurring the cost of excessive rents that would accrue if the regulator had a greater cost disadvantage.

Laffont and Tirole (1993, pp. 84–86) offer a simple model that characterizes the issues as stake here. Let's say that the regulator is responsible for two non-competing firms ($i = 1, 2$) that each produce one unit of output supplied in separate geographic areas. Their costs are given by:

$$C^i = \beta^{\mathrm{a}} + \beta^i - e^i$$

Where β^{a} is an aggregate shock to both firms and β^i is an idiosyncratic shock that is independent of β^j and e^i is firm i's effort. As before the firm's rent is given by

$$U^i - \psi(e^i)$$

and the regulator can observe only realized costs. Each firm learns the realizations of its own shocks before choosing from the regulator contracts offered to it. Laffont and Tirole develop several cases:

Case 1: In the case of purely idiosyncratic shocks ($\beta^{\mathrm{a}} = 0$), the firms are unrelated and we are back to the standard case where they must be regulated separately.

Case 2: In the case of purely aggregate shocks ($\beta^{i} = 0$) the regulator can achieve the first best outcome by offering the firms only a fixed price contract based on their relative performance or "yardstick regulation." The transfer or incentive payment is then given by $t^{i} = \Psi(e^{*}) - (C^{i} - C^{j})$. Firm i maximizes $\{\Psi(e^{*}) - [(\beta^{\mathrm{a}} - e^{i}) - (\beta^{\mathrm{a}} - e^{j})]\} - \Psi(e^{i})$ and chooses $e^{i} = e^{*}$. Since the other firm is identical it also chooses e^{*}. Neither firm earns any rents and they both exert the optimal amount of effort (they are identical). By filtering aggregate uncertainty out of each firm's realized costs we can get to the first best.

Case 3: In the case of general shocks that cannot be separated into aggregate and idiosyncratic components, a mechanism can be designed that is based in part on relative performance that has superior welfare properties to the Laffont-Tirole menu of contracts

8.3.2. Ratchet effects or regulatory lag

So far, this analysis assumes the regulator establishes a regulatory contract once and for all. This assumption is important for the results because it is assumed that the regulator can observe the firm's realized costs ex post. If the regulator then used this information to reset the firm's prices, the firm would have a less powerful incentive to engage in cost reducing effort—a "ratchet" (Weitzman, 1980). More generally, as we discussed earlier, the behavior of a firm will depend on the information that its behavior reveals to the regulator ex post and how the regulator uses that information in subsequent regulatory reviews. The effects of this kind of interaction between the regulated firm and the regulator can be captured in the Laffont and Tirole model in a straightforward manner [Laffont and Tirole (1993, Chapter 9)].

Consider the two-type case again and ignore discounting. The low cost type will choose the high powered incentive contract and will earn a rent of $\Phi(e_{\mathrm{H}})$ until the regulator resets its prices to equal its realized costs at which time its rents will fall to zero. This is not incentive compatible. The low cost type would do better by exerting less effort in the first period, reducing its disutility of effort, leading its realized production costs to increase, effectively mimicking the observable production costs expected by the regulator for the high cost type (effectively leading the regulator to believe incorrectly that the low cost type is a high cost firm). The low cost firm still earns rents in period 1, but through a lower disutility of effort. Post-ratchet, the firm faces a fixed price set equal to its realized production costs in period 2 and can now exert optimal effort and earn rents again post-ratchet by reducing its production costs.

To restore incentive compatibility with a ratchet, the low-cost type would have to be given a larger rent in period 1, at least as large as the rent it can get in period 2 after mimicking the production costs of the high cost type in period 1. However, if the first

period rents are high enough, the high cost firm may find it attractive to choose the high powered incentive scheme in period 1 and then go out of business in period 2. Laffont and Tirole call this the "take the money and run" strategy.

These simple examples are obviously rather contrived. However, we can find examples of them in the real world. The regulatory mechanism utilized extensively in the U.K. since its utility sectors were privatized is effectively a fixed price contract (actually a price cap that adjusted for general movements in input prices and an assumed target rate of productivity growth—a so-called RPI-X mechanism as discussed further below) with a ratchet every five (or so) years when the level of the price cap is reset to reflect the current realized (or forecast) cost of service [Beesley and Littlechild (1989), Brennan (1989), Isaac (1991), Sibley (1989), Armstrong, Cowan, and Vickers (1994), Joskow (2005a, 2005b)]. It has been observed that regulated firms made their greatest cost reduction efforts during the early years of the cap and then exerted less effort at reducing costs as the review approached [OFGEM (2004a, 2004b)]. More generally, the examples make the important point that the dynamic attributes of the regulatory process and how regulators use information about costs revealed by the regulated firm's behavior over time have significant effects on the incentives the regulated firm faces and on its behavior [Gilbert and Newbery (1994)].

8.3.3. No government transfers

How do the basic results developed with the Laffont-Tirole framework change if no government transfers are permitted? Clearly, regulated prices alone must now serve to deal with adverse selection, moral hazard, and allocational issues. The dichotomy between prices and incentives no longer holds. However, the same basic attributes of incentive contracting continue to apply. Focusing on linear pricing, it is optimal for the regulator to offer a menu of cost contingent price options—cost sharing or sliding scale contracts—where the attributes of the menu are chosen to balance the firms budget (production cost plus incentive payment) in a way that trades off rent extraction, effort incentives, and allocation distortions subject to participation and incentive compatibility constraints [Laffont and Tirole (1993, pp. 151–153)]. The lowest price in the menu is a fixed price designed to be chosen by the low-cost opportunity firm. The price gives that firm high powered incentives to exert cost-reducing effort, but it also leaves the most rent to the firm and involves the greatest departure of price from marginal cost. As we move to higher cost types the price increases as does the sensitivity of the price level to changes in costs. Incentives for cost reducing effort decline as β increases, rents left to the firm fall, and prices are closer to the firm's realized marginal cost, though this cost is too high due to suboptimal effort.

8.4. Price regulation when cost is not observable

As noted earlier, the earliest modern theoretical work on incentive regulation [Baron and Myerson (1982)] assumed that the regulator could not observe costs at all, could

observe demand, and that there were no moral hazard problems.[22] The regulator cares about rent extraction and must adhere to a firm participation or viability constraint. In this context, the regulatory problem is an adverse selection problem and cost contingent contracts are not available instruments since costs are assumed not to be observable. I do not find this to be a particularly realistic characterization of regulation in many developed countries, but especially in developing countries, regulators often have difficulty getting credible cost information from the firms they regulate. Moreover, accurate and meaningful cost measurement may be very difficult for multiproduct firms that have joint costs. Accordingly, I will conclude this section with a brief discussion of this literature.

Let us begin with Baron and Myerson (1982), using the development in Laffont and Tirole (1993, pp. 155–158)]. Consider a firm that has cost

$$C = \beta q$$

where the regulator observes q, has a probability distribution over β, there is no moral hazard (e), and the firm receives revenues from sales at the regulated price P and a transfer payment t from the regulator. The firm's utility is now

$$U = t + P(q)q - \beta q$$

where $P(q)$ is the inverse demand function. Since cost is not observable, the regulator must rely on fixed price contracts (and accordingly if we added moral hazard the firm would have optimal cost-reducing incentives).

If the regulator had full information the optimal linear price would be the Ramsey-Boiteux price and the associated Ramsey-Boiteux output:

$$L = (p - \beta)/p = (\lambda/1 + \lambda)(1/\eta) \tag{47}$$

where λ is either the shadow cost of public funds with government transfers or the shadow price of the budget constraint when the firm must balance its budget from sales revenues and there are fixed costs to cover.

With asymmetric information of the kind assumed here, the regulator will offer a menu of fixed price contracts that are distorted away from the Ramsey-Boiteux prices given β. The distortion for a given β reflects the tradeoff between the allocational distortion from increasing prices further above marginal cost and the cost of leaving more rent to the firm subject to the firm viability and incentive compatibility constraints.

$$L = \big(p(\beta) - \beta\big)/p(\beta) = (\lambda/1 + \lambda)(1/\eta) + (\lambda/1 + \lambda)\big[(F(\beta)/(f(\beta)p(\beta)\big] \tag{48}$$

[22] Loeb and Magat (1979) propose a mechanism where the regulator can observe the firm's demand function and can observe price and quantity ex post. The regulator does not care about the distribution of the surplus. They propose a mechanism that offers the regulated firm a subsidy equal to the total consumer surplus ex post. The firm then has an incentive to set price equal to marginal cost to maximize its profits. If the regulator cares about the distribution of income (rent extraction) it could auction of this regulatory contract in a competition for the market auction. The mechanism then reduces to Demsetz's franchise bidding scheme. The latter raises numerous issues that are discussed above.

where $F(\beta)$ and $f(\beta)$ are as defined before and $\mathrm{d}[F(\beta)/(f\beta)]\mathrm{d}\beta \geq 0$. Prices clearly exceed the Ramsey-Boiteux level at all levels of β (compare (47) and (48)). Absent the ability to use a cost-contingent reimbursement mechanism, prices must be distorted away from their Ramsey levels to deal with rent extraction/adverse selection costs.

This analysis has been extended by Baron and Besanko (1984) to allow for random audits of the firm's costs by the regulator, again in the absence of moral hazard. The firm announces its costs and prices ex ante and there is some probability that the firm will be audited ex post. The result of the audit is a noisy measure of the firm's actual costs. After the audit the regulator can penalize the firm if it gets a signal from the audit that the firm's actual costs are greater than its announced costs. Absent moral hazard the optimal policy is to penalize the firm when the audit yields a measured cost that is low, signaling the regulator that the firm's costs are likely to be lower than the cost and associated price that the firm announced. (With moral hazard things are more complicated because one does not want to penalize the firm for cost-reducing effort.) The threat that the firm's announced costs will be audited reduces the price the firm charges, the rents it retains, and the allocational distortion from prices greater than costs.[23]

8.5. *Pricing mechanisms based on historical cost observations*

Vogelsang and Finsinger (1979) have developed a mechanism that relies on observations of a regulated firm's prices, output and profits to adjust the firm's prices over time. Their mechanism, as characterized by Laffont and Tirole [Laffont and Tirole (1993, pp. 162–163)], gives the firm a reward or bonus at each point in time defined by

$$B_t = a + (\Pi_t - \Pi_{t-1}) + (p_{t-1} - p_t)q_{t-1} \tag{49}$$

where $\Pi_t = (p_t q_t - C(q_t))$. Basically, the mechanism rewards price reductions up to a point. Think of p_t as starting at the monopoly price. If the firm leaves its price at this level it gets a bonus payment of a. If it reduces its price in the second period, q will increase, profits will fall from t to $t-1$, but total revenue will increase since $\mathrm{MR} > 0$ at the monopoly price. The increase in revenue from the second term will exceed the reduction in profits from the first term, increasing net profit under the bonus formula when the price falls from the pure monopoly level and the bonus will be higher than if the firm left its price at the monopoly level. The regulator is bribing the firm to lower its prices in order to reduce the allocative distortions from prices that are too high by rewarding it with some of the increases in infra-marginal consumers surplus resulting from lower prices. Finsinger and Volgelsang show that the firm has the incentive to continue to reduce its price until it reaches the Ramsey-Boiteux price. However, if cost reducing effort is introduced, the cost-contingent nature of this mechanism leads to too little cost reducing effort [Sappington (1980), Laffont and Tirole (1993, pp. 142–145)].

[23] Lewis and Sappington (1988a) extend this line of attached to assume that the firm has private information about demand rather than costs and extend the analysis [Lewis and Sappington (1988b)] to assume private information about both demand and costs.

9. Measuring the effects of price and entry regulation

Price and entry regulation may affect several interrelated performance indicia. These indicia include the level of prices, the structure of price charged to different groups of customers or for different products, prices for inputs paid by the regulated firm, the firm's realized costs of production, firm profits, research and development activity, the adoption of product and process innovations, and the distribution of economic rents between shareholders, consumers and input suppliers. To measure the effects of regulation one must first decide upon the performance norms against which regulatory outcomes are to be measured. Candidate benchmarks include characterizations or simulations of fully efficient outcomes, hypothetical unregulated/competitive outcomes, and outcomes resulting from the application of alternative regulatory mechanisms. The identification of benchmarks and, especially the use of alternative benchmarks for normative evaluation of the effects of regulatory mechanisms and processes, should be sensitive to the fact that fully efficient outcomes or perfectly competitive outcomes are unlikely to be achievable in reality. Accordingly, what Williamson (1996, pp. 237–238) refers to as a *remediableness* criterion should be applied in normative evaluations. That is, what is the best that can be done in an imperfect world?

Once the relevant benchmarks have been identified there are several different empirical approaches to the measurement of the effects of price and entry regulation on various performance indicia.

1. *Cross-sectional/Panel-data analysis:* These studies examine the performance indicia for firms serving different geographic areas and subject to different intensities or types of regulation, typically measured over a period of more than one year. For example studies may compare prices, costs, profits, etc. for similar firms serving customers in different states under different regulatory regimes for a period of one or more years. The classic study here is that of Stigler and Friedland (1962) where they examined differences in electricity prices between states with commission regulation and states without state commission regulation of electricity prices. Or the cross-sectional variation may be between states that use different types of regulatory mechanisms (traditional cost of service with or without PBR enhancements) or apply similar mechanisms more or less intensively [Mathios and Rogers (1989)]. The assumption in many of these studies is that the choice of whether to regulate or not is exogenous, so that cross-sectional data provide observations of "natural experiments" in the impacts of the effects of alternative regulatory mechanisms. However, several recent panel data studies recognize that the choice of regulatory instrument may be endogenous [e.g. Ai and Sappington (2002), Sappington and Ai (2005)].

   Natural or near-natural experiments that produce cross-section and time series variations in the nature or intensity of regulatory mechanisms can and have provided very useful opportunities to measure the effects of regulation, the effects of variations in the structure of regulatory mechanisms and the impacts of dereg-

ulation initiatives. However ensuring that one really has a meaningful natural experiment is always a challenge [Joskow (2005b)].

In principle, cross-country comparisons can be used in an equivalent fashion, though differences in accounting conventions, data availability, and basic underlying economic and institutional attributes make cross-country studies quite difficult. Nevertheless, there has been increasing use of cross-country data both to evaluate the effects of regulation and to provide data to develop performance benchmarks that can be used by regulators [Carrington, Coelli, and Groom (2002), Jamasb and Pollitt (2003)].

2. *Time series or "before and after" analysis:* These studies measure the effects of regulation by comparing various performance variables "before" and "after" a significant change in the regulatory environment. Much has been learned about the effects of price and entry regulation by comparing firm and industry behavior and performance under regulation with the changes observed after price and entry regulations are removed. Or it could be a shift from cost of service regulation to price cap regulation. Here the challenge is to control for other factors (e.g. input prices) that may change over time as well and the inconvenient fact that regulation and deregulation initiatives are often phased in over a period of time and not single well defined events.
3. *Structural models and policy simulations:* These studies specify and estimate the parameters of firm and/or industry demand, firm costs, and competitive interactions (if any) to compare actual observations on prices, costs, profits, etc. with simulated prices under alternative regulated and unregulated regimes. This is most straightforward in the case of legal monopolies. With the demand and cost functions in hand, the optimal prices can be derived and compared to the actual prices. Or industries where there are multiple firms competing based on regulated prices, the actual prices can be compared to either optimal prices or "competitive" prices, once the nature of competitive interactions have been specified. Related work measures the attributes of production functions and tests for cost minimization. Still other work uses models of consumer demand to measure the value of new products and, in turn, the social costs of regulatory delays in the introduction. Related work on process innovations can also be incorporated in a production function framework for similar types of analysis.

9.1. Incentive regulation in practice

There is an extensive literature that examines empirically the effects of price and entry regulation in sectors in which there are (or were) legal monopolies to serve specific geographic areas (e.g. electricity, gas distribution, water, telephone) as well as in sectors in which prices and entry were regulated but two or more firms were given the legal authority to compete in the market (e.g. airlines, trucking, railroads, automobile insurance, natural gas pipelines). Reviews of the pre-1990 literature on the measurement of the effects of regulation in both single and multi-firm settings can be found in Joskow

and Noll (1981), Berg and Tschirhart (1988), Joskow and Rose (1989), Winston (1993), (Joskow, 2005b). I will focus here on the more recent nascent literature that examines the effects of incentive or performance based regulation of legal monopolies.

Although the theoretical literature on incentive regulation is fairly recent, we can trace the earliest applications of incentive regulation concepts back to the early regulation of the gas distribution sector[24] in England in the mid-19th century [Hammond, Johnes, and Robinson (2002)]. A sliding scale mechanism in which the dividends available to shareholders were linked to increases and decreases in gas prices from some base level was first introduce in England in 1855 [Hammond, Johnes, and Robinson (2002, p. 255)]. The mechanism established a base dividend rate of 10%. If gas prices increased above a base level the dividend rate was reduced according to a sharing formula. However, if gas prices fell below the base level the dividend rate did not increase (a "one-way" sliding scale). The mechanism was made symmetric in 1867. Note that the mechanism was not mandatory and it was introduced during a period of falling prices [Hammond, Johnes, and Robinson (2002, pp. 255–256)]. A related profit sharing mechanism [what Hammond, Johnes, and Robinson (2002) call the "Basic Price System"] was introduced in 1920 that provided a minimum guaranteed 5% dividend to the firm's shareholders and shared changes in revenues from a base level between the consumers, the owners of the firm and the firm's employees. Specifically, this mechanism established a basic price p_b to yield a 5% dividend rate. This dividend rate was the minimum guaranteed to the firm. At the end of each financial year the firm's actual revenues (R) were compared to its basic revenues $R_b = p_b$ times the quantity sold. The difference between R and R_b was then shared between consumers, investors and employees, apparently subject to the constraint that the dividend rate would not fall below 5%. Hammond, Johnes, and Robinson (2002) use "data envelopment" or "cost frontier" techniques [Giannakis, Jamasb, and Pollitt (2004)] to evaluate the efficiency properties of three alternative gas distribution pricing mechanisms used in England based on data for 1937. While they find significant differences in performance associated with the different mechanisms, the linkage between the incentive structure of the different mechanisms and the observed performance is unclear. Moreover, the analysis does not appear to account for the potential endogeneity of the choice of regulatory mechanism applied to different firms.

In the early 20th century, economists took note of the experience with sliding scale mechanisms in England, but appear to have concluded that they were not well matched to the regulation of electricity and telephone service (and other sectors) where demand and technology were changing fast and future costs were very uncertain [Clark (1913)]. As already discussed, cost of service regulation (with regulatory lag and prudence reviews) evolved as the favored alternative in the U.S., Canada, Spain and other countries

[24] This is before the development of natural gas. "City gas" was manufactured from coal by local gas distribution companies. At the time there were both private and municipal gas distribution companies in operation in England.

with private (rather than state-owned) regulated monopolies and the experience in England during the 19th and early 20th centuries was largely forgotten by both regulators and students of regulation.

State public utility commissions began to experiment with performance based regulation of electric utilities in the 1980s. The early programs were targeted at specific components of an electric utility's costs or operating performance such as generation plant availability, heat rates, or construction costs [Joskow and Schmalensee (1986), Sappington et al. (2001)]. However, formal comprehensive incentive regulation mechanism have been slow to spread in the U.S. electric power industry [Sappington et al. (2001)], though rate freezes, rate case moratoria, price cap mechanisms and other alternative mechanisms have been adopted in many states, sometimes informally since the mid-1990s. Because of the diversity of these programs, the co-existence of formal and informal programs, and the simultaneous introduction of wholesale and retail competition and related vertical and horizontal restructuring initiatives (Joskow, 2000), it has been difficult to evaluate the impact of the introduction of these incentive regulation mechanisms in the electric power sector. Rose, Markowicz, and Wolfram (2004) examine aspects of the operating performance of regulated generating plants during the period 1981–1999 and find that the threat of the introduction of retail competition led to improvements in various indicia of generating plant performance.

Beginning in the mid-1980s a particular form of incentive regulation was introduced for the regulated segments of the privatized electric gas, telephone and water utilities in the U.K., New Zealand, Australia, and portions of Latin American as well as in the regulated segments of the telecommunications industry in the U.S.[25] The mechanism chosen was the "price cap" [Beesley and Littlechild (1989), Brennan (1989), Armstrong, Cowan, and Vickers (1994), Isaac (1991), Joskow (2006)]. In theory, a price cap mechanism is a high-powered "fixed price" regulatory contract which provides powerful incentives for the firm to reduce costs. Moreover, if the price cap mechanism is applied to a (properly) weighted average of the revenues the firm earns from each product it supplies, the firm has an incentive to set the second-best prices for each service [Laffont and Tirole (2000), Armstrong and Vickers (1991)].

In practice, price cap mechanisms apply elements of cost of service regulation, yardstick competition, high powered "fixed price" incentives plus a ratchet. Moreover, the regulated firm's ability to determine the structure of prices under an overall revenue cap is typically limited. Under price cap regulation the regulator sets an initial price p_o (or a vector of prices for multiple products). This price (or a weighted average of the prices allowed for multiple products) is then adjusted from one year to the next for changes in inflation (rate of input price increase or RPI) and a target productivity change factor "X." Accordingly, the price in period 1 is given by:

$$p_1 = p_o(1 + \text{RPI} - X) \tag{50}$$

[25] The U.S. is behind many other countries in the application of incentive regulation principles, though their use is spreading in the U.S. beyond telecommunications.

Typically, some form of cost-based regulation is used to set p_o. The price cap mechanism then operates for a pre-established time period (e.g. 5 years). At the end of this period a new starting price and a new X factor are established after another cost-of-service and prudence or efficiency review of the firm's costs. That is, there is a pre-scheduled regulatory-ratchet built into the system.

Several things are worth noting about price cap mechanisms since they have become so popular in the regulatory policy arena. A pure price cap without cost-sharing (a sliding scale mechanism) is not likely to be optimal given asymmetric information and uncertainty about future productivity opportunities. Prices would have to be set too high to satisfy the firm participation constraint and too much rent with be left on the table for the firm. The application of a ratchet from time to time that resets prices to reflect observed costs is a form of cost-contingent dynamic regulatory contract. It softens cost-reducing incentives but extracts more rents for consumers.

Although it is not discussed too much in the empirical literature, price cap mechanisms are typically focused on operating costs only, with capital cost allowances established through more traditional utility planning cost-of-service regulatory methods. In addition, it is widely recognized that a pure price cap mechanism provides incentives to reduce both costs and the quality of service. Accordingly, price cap mechanisms are increasingly accompanied either by specific performance standards and the threat of regulatory penalties if they are not met or formal PBR mechanisms that set performance standards and specify penalties and rewards for the firm for falling above or below these performance norms [OFGEM (2004b, 2004c, 2004d), Sappington (2003), Ai and Sappington (2002), Ai, Martinez, and Sappington (2004)].

A natural question to ask about price cap mechanisms is where does "X" (and perhaps p_o) come from [Bernstein and Sappington (1999)]? Conceptually, assuming that RPI is a measure of a general input price inflation index, X should reflect the difference between the expected or target rate of total factor productivity growth for the regulated firm and the corresponding productivity growth rate for the economy as a whole and the difference between the rate of change in the regulated firm's input prices and input prices faced by firms generally in the economy. That is, the regulated firm's prices should rise at a rate that reflects the general rate of inflation in input prices less an offset for higher (or lower) than average productivity growth and an offset for lower (or higher) input price inflation. However, the articulation of this conceptual rule still begs the question of how to calculate X in practice.

In practice, the computation of X has often been fairly ad hoc. The initial application of the price cap mechanism by the Federal Communications Commission (FCC) to AT&T's intercity and information services used historical productivity growth and added an arbitrary "customer dividend" to choose an X that was larger than the historical rate of productivity growth. In England and Wales and some other countries, benchmarking methods have come to be used to help to determine a value for X [Jamasb and Pollitt (2001, 2003)] in a fashion that is effectively an application of yardstick regulation. A variety of empirical methods have been applied to identify a cost efficiency frontier and to measure how far from that frontier individual regulated firms

lie. The value for X is then defined in such a way as to move the firms to the frontier over a pre-specified period of time (e.g. five years). These methods have recently been expanded to include quality of service considerations [Giannakis, Jamasb, and Pollitt (2004)].

The extensive use of periodic "ratchets" or "resets to cost" along with price cap mechanisms reflect the difficulties of defining an ideal long-term value for X and the standard tradeoffs between efficiency incentives, rent extraction and firm viability constraints. These ratchets necessarily dull incentives for cost reduction. Note in particular that with a pre-defined five year ratchet, a dollar of cost reduction in year one is worth a lot more than a dollar of cost reduction in year four since the cost savings are retained by the firm only until the next reset anniversary [OFGEM (2004b)].

Most of the scholarly research evaluating the effects of incentive regulation have focused on the telecommunications industry [Kridel, Sappington, and Weisman (1996), Tardiff and Taylor (1993), Crandall and Waverman (1995), Braeutigam, Magura, and Panzar (1997), Ai and Sappington (2002), Banerjee (2003)]. Ai and Sappington's study is the most recent and comprehensive. They examine the impact of state incentive regulation mechanism applied to local telephone companies between 1986 and 1999 on variables measuring network modernization, aggregate investment, revenue, cost, profit, and local service prices. The methodological approach involves the use of a panel of state-level observations on these performance indicia, state regulatory regime variables and other explanatory variables. Instrumental variables are used to deal with the endogeneity of the choice of regulatory regime and certain other explanatory variables so that the fixed-effects estimates are consistent. Ai and Sappington (2002) find that there is greater network modernization under price cap regulation, earnings sharing regulation, and rate case moratoria (effectively price cap regulation with $\text{RPI} + X = 0$), than under rate of return regulation. Variations in regulatory mechanisms have no significant effects on revenue, profit, aggregate investment, and residential prices, and except for rate case moratoria, on costs. Crandall and Waverman (1995) find lower residential and business prices under price cap regulation than under rate of return regulation but other forms of incentive regulation do not yield lower prices. Tardiff and Taylor (1993) use similar methods and find similar results to Ai and Sappington (2002). Braeutigam, Magura, and Panzar (1997) find lower prices under some types of price cap regulation but not under other form of incentive regulation.

Sappington (2003) reviews several studies that examine the effects of incentive regulation on the quality of retail telecommunications service in the U.S. These studies do not lead to consistent results and for many dimensions of service quality there is no significant effect of variations in the regulatory regime applied. Ai, Martinez, and Sappington (2004) also examine the effects of incentive regulation on service quality for a state-level panel covering the period 1991 through 2002. They find that incentive regulation is associated with significantly higher levels of service quality compared to rate of return regulation for some dimensions of service quality (e.g. installation lags, trouble reports, customer satisfaction) and significantly lower levels of service quality in other dimensions (e.g. delays in resolving trouble reports, percentage of installation

commitments met). Banerjee (2003) provides a related empirical analysis of the effects of incentive regulation on telephone service quality.

Systematic research on the effects of incentive regulation in other industries is limited. Newbery and Pollitt (1997) argue that the incentive regulatory mechanisms applied to electricity distribution companies in England and Wales during the first half of the 1990s led to significant efficiency improvements. Significant savings associated with the application of price cap and other incentive mechanisms to electricity distribution and transmission have also been noted by regulators in the U.K. [OFGEM (2004a, 2004b), Joskow (2006)]. Rudnick and Zolezzi (2001) examine the changes in several dimensions of productivity in the liberalized electricity sectors in Latin America during the 1990s and find significant improvements in these productivity indicia. Bacon and Besant-Jones (2000) provide a broader assessment of the effects of privatization, market liberalization and regulatory reform of electricity sectors in developing countries, indicating more mixed results. However, it is hard to know how much of these observed cost reductions is due to the incentive regulation mechanisms and how much to privatization. Estache and Kouasi (2002) examine the diverse performance effects of alternative governance arrangements on African water utilities and Estache, Guasch, and Trujillo (2003) analyze the effects of price caps and related incentive provisions on and the renegotiation of infrastructure contracts in Latin America.

Gagnepain and Ivaldi (2002) examine the effects of incentive regulatory policies on public transit systems in France using data on a panel of French municipalities over during the period 1985–1993. This is a particularly interesting paper because the empirical analysis is embedded directly in a structural model of optimal regulation a la Laffont and Tirole (1993) discussed above.

Since 1982 local public transport (buses, trams) in France has been decentralized to the municipalities. The municipalities own the rolling stock and infrastructure but contract out the operation of the systems to private operators in 80% of the cases (there are three private operators in the country which also provide other municipal services). Fares (P) do not produce enough revenue to cover the total costs incurred by the operators so there are transfer payments from the government to the operator to satisfy break-even constraints (the treatment of the costs of municipal-owned infrastructure in the analysis is a little unclear, but they are not paid directly by the operator).

Private operators are given either "cost-plus" (CP) contracts or "fixed price" (FP) contracts. The former cover observed costs and ex post deficits. The latter cover expected costs and expected deficits. In 1995, 62% of the operators had fixed price contracts and 25% cost-plus contracts. The rest were operated by the municipalities or are not in the data base for other reasons. The contracts have a duration of one year and municipalities apparently never switch operators during the sample period. The analysis focuses on the larger municipalities with more than 100,000 population (excluding Paris, Lyon, Marseilles, which were not included in the data set) and relies on a panel data set for 59 municipalities over 1985–1993 period on input costs, output, network infrastructure.

The paper includes the following empirical analyses: (a) estimates the parameters of a cost function (structural model) for urban public transport that treats the effects

of regulation on costs under asymmetric information as endogenous given the type of contract each system is place upon; (b) estimates the parameters of the distribution of the "labor inefficiency" parameter θ and the cost of effort function give assumptions about the form of these functions (e.g. a beta distribution for θ) and the cost function (Cobb-Douglas technology); (c) estimates the level of inefficiency θ_i and effort (e_i) of each urban transport system given the cost function's parameters, the estimated parameter of the cost of effort function and the regulatory contract they have been placed upon; (d) estimates the implied cost of public funds given the cost function parameters, the parameters of a demand function for public transport and the form of the contract each transport operator has been given assuming that the municipality sets the optimal fare (Ramsey-Boiteux) given demand, costs, and each municipality's cost of public funds λ and; (e) calculates the optimal regulatory contract [second-best under asymmetric information a la Laffont and Tirole (1993)] for each system given the cost of public funds and its inefficiency parameter and the welfare gains from doing so.

These analyses leads to several conclusions including: (a) there are economies of scale in urban transport; (b) there is a large variation in the efficiency parameters for different networks; (c) for the lowest θ group the cost distortions (difference in efficiency between a fixed price and a cost plus contract) are not significantly different between the FP and CP contracts, for the intermediate θ the difference in cost distortions is about 4%, and for the highest θ group there is a lot of inefficiency even with a FP contract, but FP contracts reduce costs significantly (mix of CP and FP contracts); (d) cost of public funds varies from 0.17 to 0.56 across municipalities and (e) optimal second best (Laffont-Tirole) contracts improve welfare significantly compared to cost plus contracts, but not compared to fixed price contracts.

Research measuring the effects of regulation and deregulation on the speed of introduction of new services and technologies in telecommunications also make it clear that dynamic considerations are extremely important from a social welfare perspective [Hausman (1997), Crandall and Hausman (2000)]. Hausman (1997) estimates the costs of FCC delays in the introduction of voice messaging service and cellular telephone service by estimating the structural parameters of consumer demand and the value to consumers of new goods. The basic method is to estimate the effect of the introduction of a new good on real consumer income and then to perform a counterfactual analysis to measure the costs foregone by regulatory delays in introducing the product. He finds that regulatory and court delays led voice messaging to be introduced 5 to 7 years later than it would have been without these delays. He finds as well [see also Hausman (2002, 2003)] that FCC regulatory delays led to cellular telephone being introduced 7 to 10 years later than would have been the case without these delays. The social costs of these delays are estimated to be about $6 billion and $30 billion in 1994 dollars respectively. Hausman (2002) also finds that other regulatory restrictions on mobile service competition led to significantly higher prices for mobile services.

Greenstein, McMaster, and Spiller (1995) examine the effects of incentive regulation on the deployment of digital technologies by local telephone carriers. Recent work by Thomas Hubbard has shown how new technologies adopted by post-deregulation

trucking firms have both served to improve service quality and to improve productivity and lower costs [Hubbard (2001, 2003)]. Regulation of prices and entry prior to deregulation in 1980 inhibited the diffusion of these kinds of technologies in a number of different ways, though the precise impact of regulation per se has not been measured. Rose and Joskow (1990) examine the diffusion of new electric generating technologies in the electric power sector. Goolsbee and Petrin (2004) find significant consumer benefits from the entry of direct broadcast satellite to compete with cable TV, but limited effects on the cable firms' marker power.

It is clear that the social costs of delaying product and process innovations can be very significant. Both theoretical and empirical research has probably focused too much on static welfare effects associated with the impacts of regulation on prices and costs in the short run and too little research has focused on the effects of regulation on the adoption and diffusion of product and process innovations.

10. Competitive entry and access pricing

The firms in many industries that have been subject to price and entry regulation have organizational structures that involve vertical integration between production of complementary services at different levels of the production chain. For example, in most countries, electric power companies historically evolved with governance structures where generation, transmission, distribution and retail marketing of electricity where vertically integrated (Joskow, 1997). However, there are also thousands of small municipal and cooperative distribution utilities that purchase power from third parties (typically proximate vertically integrated utilities) which they then resell to retail consumers to whom they provide distribution (delivery) service in their franchise areas. In many countries natural gas producers also own natural gas pipelines that transport natural gas from where it is produced to where it is distributed in local consumption areas. Telephone companies historically provided both local and intercity services and, in the U.S., the vertical integration extended into the production of telephone network equipment and customer premises equipment.

These industries likely evolved with these structures in response to economies of vertical integration [Joskow (1997)]. However, to the extent that the economies of vertical integration led to the integration of a production segment with natural monopoly characteristics with a production segment without natural monopoly characteristics, the effect of vertical integration is to extend the natural monopoly to the potentially competitive segments as well. For example, the transmission and distribution of electricity have natural monopoly characteristics. However, there are numerous generating plants in each region of the U.S., suggesting that the generation of power may be potentially competitive [Joskow and Schmalensee (1983), Joskow (1997)]. Vertical integration effectively extends the natural monopoly over transmission and distribution to generation when firms in the industry are vertically integrated, extending the boundaries of regulation and its complexities and potential imperfections. Alternatively, two or more vertically integrated segments may once have had natural monopoly characteristics as well as

economies of vertical integration, but technological change may have changed the characteristics of the underlying technology at one or more levels of the vertical chain to make it potentially competitive. For example, microwave, satellite, and radio technology, as well as the diffusion of cable television, have changed the economic attributes of both the supply and demand for intercity and local telecommunications services dramatically.

The bundling of multiple supply segments (or products), one or more of which does not have natural monopoly characteristics and is potentially competitive, into a single firm subject to price and entry regulation naturally leads to a number of questions and issues. Would better performance be achieved by separating the potentially competitive segments from the natural monopoly segments and removing price and entry regulation from the competitive segments? Are the benefits of potentially imperfect competition in these segments greater than the lost cost savings from vertical integration (if any)? Or should we allow the incumbent regulated firm to continue to offer both sets of services, but allow competitive entry into the potentially competitive segments so that entering firms can compete with the incumbent? If we take this approach when and how do we regulate and deregulate the prices charged by the incumbent for competitive services? How do we know that competitive entry will take place because lower cost suppliers have incentives to enter the market rather than inefficient entry resulting from price distortions resulting from decades of regulation? Should limits be placed on the ability of regulated firms to respond to competitive entry to guard against predatory behavior? Is structural separation necessary (divestiture) or is functional separation with line of business restrictions to deal with potential cross-subsidization of by regulated services by regulated services and behavior that disadvantages competitors sufficient to foster efficient competition? These issues are especially challenging in many regulated industries because access to the natural monopoly segments (e.g. the electric transmission network) is necessary for suppliers in the competitive segment (e.g. generating plants) to compete. Such networks are often referred to as "essential facilities" or "bottleneck facilities," though these terms have been abused in the antitrust policy context.

The terms and conditions under which competitive suppliers can gain access to the incumbent's monopoly network when the incumbent is also a competitor in the competitive segments has been the focus of considerable research in the last decade as previously regulated vertically integrated firms in several regulated industries are "restructured" to separate natural monopoly network segments from competitive segments and price and entry regulation relaxed in the competitive segments [Vickers (1995), Laffont and Tirole (2000), Baumol and Sidak (1994), Vogelsang (2003)]. If the access prices are set too low, inefficient entry may be encouraged. If access prices are set too high they will serve as a barrier to entry to competitors who are more efficient than the entrant or encourage inefficient bypass of the network to which access is sought. When prices for regulated services are partially based on realized costs, cost allocations between regulated and unregulated services becomes an issue as well since the incumbent may be able to subsidize the costs of providing competitive services by hiding some of them in the cost of service used for determining regulated prices. Access pricing issues

also arise when the incumbent network operator's business is restricted to regulated network services only, but the nature of the distortions is different as long as all competitors are treated equally.

10.1. One-way network access

Much of the access pricing literature initially evolved in the context of the development of competition in the supply of intercity communications services and the interconnection of competing intercity networks with regulated monopoly local telephone networks which originate and terminate intercity calls. I will focus on telecommunications examples here, following the development in Laffont and Tirole (2000). Conceptually similar issues arise in electricity and natural gas as well, though the technical details are different (Joskow, 2005a). There are two kinds of services. The first is provision of "local network" service which is assumed for now to be a natural monopoly and subject to price and entry regulation. The second service is intercity service which allows for transmission of voice and data signals between local networks in different cities, is supplied by the incumbent and is being opened to potential competitors. The incumbent is assumed to be vertically integrated into both the provision of local exchange services and the provision of intercity services and the prices for both services are assumed to be regulated. For a competitive intercity supplier to enter the market and compete with the incumbent it must be able to gain access to the local network in one city to originate calls and to gain access the local network in the other city to complete the calls. The entrant is assumed to provide its own intercity facilities to transport the calls between local networks but relies on the regulated monopoly incumbent to provide local connection services on the local origination and termination networks. These relationships are displayed in Figure 12.

Let:

q_o = quantity of local calls sold at price p_o

q_1 = quantity of incumbent's long distance calls at price p_1

q_2 = quantity of entrant's long distance calls at price p_2

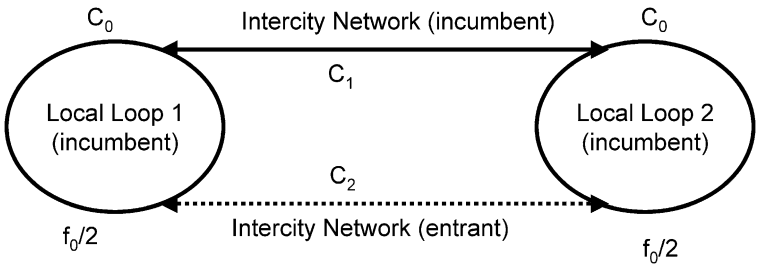


Figure 12. One-way access.

$Q = q_o + q_1 + q_2 =$ total calls

$f_o =$ fixed cost of the local network

$c_o =$ cost of originating or terminating a local call

$c_1 =$ incumbent's cost of a long distance call

$c_2 =$ entrant's cost of a long distance call

Convention: Every local or long distance call involves one origination and one termination on the local network. Each local network has a marginal cost per call of c_o so the marginal cost to use local networks to originate and complete a call is $2c_o$.

Incumbent's costs: $f_o + 2c_o(q_o + q_1 + q_2) + c_1 q_1$

Entrant's costs: $c_2 q_2 + aq_2 = c_2 q_2 + (p_2 - c_2)q_2$

where "a" is the *access price* the entrant must pay to the incumbent for using its local network facilities (one origination and one termination per long distance call).

Assume that the entrant has no market power so it sets its price for intercity calls equal to the marginal cost it incurs, including the price it is charged for access to the incumbent's local access. The entrant's price for long distance service p_2 must be (it passes along marginal costs with no additional markup):

$$p_2 = a + c_2$$

and

$$a = p_2 - c_2$$

(A useful way to think about this is that the incumbent "subcontracts" with the entrant to supply the entrant's long distance service at cost.)

The incumbent's profits on the provision of local, long distance and "access service" are then given by:

$$\begin{aligned}\pi(p_o, p_1, p_2) = {} & (p_o - 2c_o)q_o \\ & + (p_1 - c_1 - 2c_o)q_1 \\ & + (p_2 - c_2 - 2c_o)q_2 \\ & - f_o\end{aligned}$$

(where $p_2 - c_2 = a$).

Assume that $S_o(p_o)$ and $S_1(p_1, p_2)$ give the net consumers' surpluses for local and long distance and recall that the derivative of the net surplus with respect to a price is (minus) the corresponding quantity. Assume as well that the incumbent is regulated and is subject to a breakeven constraint. Then the optimal prices p_o, p_1, and p_2 (and "a") are given by:

$$\underset{(p_o, p_1, p_2)}{\text{Max}} \left\{ S_o(p_o) + S_1(p_1, p_2) + \pi(p_o, p_1, p_2) \right\} \tag{51}$$

$$\text{S.T. } \pi(p_o, p_1, p_2) \geq 0$$

This is just a familiar Ramsey-Boiteux pricing problem and yields the following familiar conditions where λ (> 0) is the shadow cost of the budget constraint and the η_i are price *superelasticities* that account for cross-price effects when there are goods that are substitutes or complements:

$$\frac{p_0 - 2C_0}{p_0} = \frac{\lambda}{1+\lambda}\frac{1}{\eta_0} \tag{52}$$

$$\frac{p_1 - C_1 - 2C_0}{p_1} = \frac{\lambda}{1+\lambda}\frac{1}{\eta_1} \tag{53}$$

$$\frac{p_2 - C_2 - 2C_0}{p_2} = \frac{\lambda}{1+\lambda}\frac{1}{\eta_2} \tag{54}$$

Note that if there were no fixed costs f_o, the optimal price would be equal to marginal cost and the access price "a" would equal $2c_o$. However, with fixed costs, the access price includes a contribution to these fixed costs. The optimal Ramsey-Boiteux access price $a = p_2 - c_2$ [Laffont and Tirole (1996, 2000, pp. 102–103)] then follows from the formula

$$a = 2C_0 + \lambda/(1+\lambda)(p_2/\eta_2)$$

This can be rewritten (Armstrong, Doyle, and Vickers, 1996) as

$$a = 2C_0 + \lambda/(1+\lambda)(p_2/\varepsilon_2) + \delta(p_1 - 2C_0 - C_1) \tag{55}$$

where

$$\delta = -\left[\frac{\partial q_1/\partial p_2}{\partial q_2/\partial p_2}\right]$$

is the change in the sales of the incumbent divided by the change in sales of the competitive entrant and ε_2 is the own-price elasticity of demand without accounting for cross-price effects. The Ramsey-Boiteux access price "a" is set above marginal cost and therefore contributes to the incumbent's fixed costs. It is composed of two components. The first is the standard Ramsey price equation. The second allows for the substitution between the incumbents sales of network services and its loss of retail sales to the competitive entrant.

According to Willig (1979) the appropriate access price is given by $a = p_1 - c_1$ or the difference between the incumbent's retail price for long distance calls and the marginal or avoided cost of supplying these calls (see also Baumol and Sidak, 1994, Baumol, Ordover, and Willig, 1997). This rule is often referred to as the Efficient Component Pricing Rule or ECPR. It has been argued that this rule has a number of desirable features including: (a) potential entrants can enter profitably if and only if they are more efficient than the incumbent; since $a + c_2 = p_1 - (c_1 - c_2)$, only a cost advantage will lead to entry; (b) entry is neutral regarding operating profit for the incumbent since it still gets the same profits on sales of "access" as it does on retail sales. Incumbent does

not have incentive to destroy entrant; (c) entry does not interfere with existing cross-subsidies and is not "unfair" to the incumbent; (d) if entrants do not have lower costs there will be no entry and (e) if entrants have lower costs the incumbent will be driven from the retail market and will supply only "access."

Laffont and Tirole (1996, 2000) and others point out a number of problems with the ECPR. They include: (a) ECPR is a "partial rule" in the sense that it does not tell us how p_1 should be set optimally. It takes p_1 as given. However, given that regulators are unlikely to have set second-best efficient prices as competition emerges, this may be a very practical real world approach; (b) ECPR is implied by Ramsey-Boiteux pricing only under a restrictive set of assumptions. These assumptions are equivalent to assuming that there is full symmetry between the incumbent and the potential entrant in the sense that they have equal costs of providing intercity services ($c_1 = c_2$), that they face symmetrical demands in the intercity (competitive) segment, and that the entrants have no market power. In this case since $a = p_2 - c_2$, the combination of $p_1 = p_2$ and $c_1 = c_2$ implies that $a = p_1 - c_1$ which is the efficient component pricing rule; (c) ECPR gives the wrong access price for competitive services that are differentiated products rather than being identical to the product produced by the incumbent.

To illustrate this last point, assume that the competitors have the same costs by different demands

$$q_1 = a_1 - bp_1 + dp_2$$
$$q_2 = a_2 - bp_2 + dp_1$$

where $a_1 > a_2$ (brand loyalty/less elastic demand) and $b > d$.

In this case is can be shown that the access price should be lower than ECPR:

$$a < p_1 - c_1 \quad \text{and} \quad p_1 > p_2$$

The reason is that the optimal price for the incumbent is higher than the optimal price for the entrant because the incumbent has a less elastic demand:

$$p_1 > p_2$$

The access price (for an intermediate good) must be lower to keep p_2 from rising above its optimal level. In this case, if the incumbent has lower costs than the entrant the access price should be higher than ECPR. The logic is the same in reverse. The optimal prices are $p_1 < p_2$.

The ECPR is also not efficient if entrants have market power. If the entrants have market power they will mark up the access price when they set the retail price leading to a classic double marginalization problem [Tirole (1988, Chapter 4)]. When there is competitive entrant with market power, the optimal access prices is the ECPR level minus the competitor's unit markup m:

$$a = p_1 - c_1 - m$$

Finally, the entrant may be able and willing inefficiently to bypass the incumbent's network if the access price is greater than its own cost of duplicating the network.

The regulator can respond to this problem by setting access charge lower than the incumbent's entry cost. But this increases the incumbent's access deficit and the incumbent would than have to increase prices further for captive customers. In principle the regulator could charge low access price and then levy an effective excise tax on the competitor's sales to cover the access deficit, but such an instrument may not be available to the regulator.

These considerations suggest that setting the optimal access price requires consideration of many other aspects of the industrial organization of the potentially competitive sector which may not be consistent with the assumptions that lead to the ECPR. The optimal access prices reflect standard Ramsey pricing considerations, the relationship between wholesale access prices and the incumbent's retail sales, differentiated product and double marginalization (vertical market power) considerations, and imperfect competition in the potentially competitive segment. Setting the optimal access prices clearly places very significant information and computational burdens on regulators.

Laffont and Tirole (1996, 2000) suggest that a superior approach to setting access prices is to apply a *global price cap* to the regulated firm that includes "network access" as one of the products included in the price cap. If the weights in the price cap formula are set "properly" (equal to the realized quantities from optimal pricing) then the regulated firm will have the incentive to price all of the services covered, including pricing "access" to the network at the optimal Ramsey-Boiteux prices that take all of the relevant costs and super-elasticities into account. They recognize, however, that finding the optimal quantities also creates a significant information and computational burden on regulators. In addition, applying a price cap mechanism in this way may enhance incentives for the incumbent to adopt a predatory pricing strategy that leads to an access price that is set too low, with the lost net revenue partially recouped in the short run by increasing prices for other regulated services and in the long run by inducing competitors to exit the market. Accordingly, they suggest that a global price cap be combined with a rule that the access price can be no lower than the difference between the incumbent's retail price and its avoided cost or its "opportunity cost" of sales lost to the incumbent ($a \leq p_1 - c_1$).

10.2. Introducing local network competition

In 1996, the U.S. Congress determined that competition should be opened up for providing local network services as well as intercity services. That is, it adopted a set of policies that allowed competitors to offer local telephone services in competition with the incumbent Local Exchange Carriers (ILECs). The Competitive Local Exchange Carriers (CLECs) could compete by building their own facilities (as a cable television network might be able to do) or by leasing the facilities owned by the ILEC. The argument was that while there were opportunities for facilities-based competition at the local network level, there were likely to be components of the local network that had natural monopoly characteristics (e.g. the "last mile" from the local exchange to the end-user's premises). At the very least, it would take time for facilities based competitors to build

out a complete network. This is not the place to go into the complex issues associated with local service competition, but this policy required regulators to set regulated prices at which competitors could gain access to the local loop. Accordingly, a brief discussion of the issues associated with the regulated pricing of "unbundled network elements" is in order (Laffont and Tirole, 2000, Crandall and Hausman, 2000).

Following the Telecommunications Act of 1996, the FCC required ILECs to offer to lease pieces of their networks (network elements) to CLECs. In addition to requiring interconnection of networks so that all termination locations could be reached on any network, the FCC concluded that it would also require ILECs to lease individual network elements to CLECs. The FCC decomposed ILEC networks into a complete set of "Network Elements" and required the ILECs to lease each and every element to CLECs requesting service "at cost" (see Hausman, 1999, Crandall and Hausman, 2000). For example, RCN built its own network providing cable TV, telephone and high-speed internet service in portions of Boston and Brookline. It is interconnected to Verizon's local telephone network so that RCN subscribers can reach on-net and off-net locations and vice versa. If RCN also wanted to offer service to potential subscribers in, say, Cambridge, but building its own network there is uneconomical, it could then lease all of the network elements on Verizon's Cambridge's network at "cost-based" wholesale prices, and begin offering service there as if it were its own network. At that time, the FCC thought that this was the best way to promote local service competition. This policy leads to a number of questions, only two of which I will discuss briefly here.[26]

What is the right regulated price for network elements? The Federal Communications Commission (FCC) used an engineering model of a local telephone network and estimates of the current cost of equipment and maintenance to build an "optimal network" and then to estimate the "forward looking long run incremental costs" for each network element (TELRIC). This approach has a number of shortcomings: (a) The underlying engineering model is at best an imperfect representation of real telephone networks; (b) the cost calculations fail properly to take into account economic depreciation of equipment and lead to current cost estimates that are biased downward [see Hausman (1999, 2003)]; (c) the cost calculations fail to take into account the interaction between the sunk cost nature of telecom network investments and uncertainty over future demand, equipment prices, and technical change. This also leads to a significant underestimate of the true economic costs of short-term leasing arrangements. The FCC leasing rule effectively includes an imbedded option. The CLEC can take the service for a year and then abandon it if a cheaper alternative emerges or continue buying at wholesale until a better alternative does emerge (if the ILECs instead could sell the

[26] Other questions include: If RCN is simply buying service on Verizon's network at wholesale prices (including connects and disconnects, network maintenance, etc.) and then reselling these services under it's brand name, what is the social value added from making this competition possible? "Retail service" costs are very small. If an ILEC must lease any and all of its facilities to competitors "at cost," how does this affect its incentives to invest on its network in general, and in particular, to invest in new technologies for which it must compete with other firms?

network elements to the CLECs at their installation cost rather than offer the service on short-term leases, this would solve this problem) [Hausman (1999), Hausman and Myers (2002), Pindyck (2004)]; (d) wholesale network element prices determined by the TELRIC rules are substantially below the ILECs actually regulated costs. This is not surprising since the regulated costs are based on traditional "depreciated original cost ratemaking" techniques and reflect historical investments that were depreciated too slowly. This creates stranded cost problems for the ILECs and potential distortions in demand for "new network elements" rather than equivalent "old network elements." Unlike the situation in electricity and natural gas sector reforms, where regulators have recognized and made provisions for stranded costs recovery, this issues has largely been ignored in the U.S. in the case of telecom reform; and (e) these rules reduce ILEC incentives to invest in uncertain product service innovations. The FCC rules ignore the cost of "dry holes." CLECs can buy new successful services "at cost," compete with the ILEC for customers for these services, and avoid paying anything for ILEC investments that are unsuccessful [Crandall and Hausman (2000), Pindyck (2004)].

All of these considerations suggest that TELRIC underprices network elements. Moreover, competitive strategies of CLECS may be driven more by imperfections in FCC pricing rules than by their ability to offer cheaper/better products. Nevertheless, so far there has been only limited successful CLEC competition except for business customers in central cities.

10.3. Two-way access issues

Opening up the local loop to competition raises another set of interesting pricing issues that arise when there are two (or more) bottleneck networks which (a) need to interconnect (cooperate) with one another to provide retail services and (b) may compete with one another for customers. Such situations include (a) overlapping LECs which require interconnection; (b) internet networks which exchange data traffic; (c) credit card associations, and (d) international telephone calls where networks at each end must exchange traffic. These situations create a set of "two-way access" pricing problems. What are the most efficient access pricing arrangements to support interconnection between the two (or more) networks and what institutional arrangements should be relied upon to determine threes prices? Regulation, cooperation, non-cooperate competitive price setting? There are a number of policy concerns: (a) cooperation may lead to collusion to raise prices at the level where the networks compete (e.g. retail calling); (b) non-cooperative access pricing may fail to account properly for impacts on other networks; and (c) inefficient access prices may increase entry barriers and soften competition [Laffont and Tirole (2000); Laffont, Rey, and Tirole (1998a, 1998b)].

The literature on two-way access pricing is closely related to the growing and much broader literature on "two-sided markets" [Rochet and Tirole (2003, 2004)], though he precise definition of what is included within the category "two-sided markets" is somewhat ambiguous. However, many markets where network platforms are characterized by network externalities are "two-sided" in the sense that the value of the network plat-

forms depends on getting buyers and/or sellers on both sides of the market to use them effectively through pricing arrangements and market rules. The value of a credit card to consumers depends on its broad acceptance by retailers. The value of a telephone network depends on the number of consumers who can be reached (to call or be called) on it or through interconnection with other telephone networks. The value of a bank's ATM network to its depositors depends on their ability to use other ATM networks to get cash from their bank accounts.

A discussion of the literature on two-sided markets in beyond the scope of this chapter. However, to identify the issues at stake I will briefly discuss the nature of the access pricing issues that arise absent price regulation when multiple networks serve consumers who in turn value reaching or being reached by consumers connected to the other networks. While these kinds of problems may be solved by regulation, the more typical solution is for the network participants and the networks to negotiate access pricing arrangements and market rules to deal with the potential inefficiencies created by network externalities and market power. I will follow Laffont and Tirole (2000) to identify some of the issues at stake in this literature.

Consider a situation where we have a city served by two local telephone networks which we can call the A and B networks (Weiman and Levin, 1994). Customers are connected to one network or the other and all customers need to be able to call any other customer whether they are on the same network or not. Laffont and Tirole's (2000) analysis of this situation adopts two "conventions." (a) The calling company's network pays a (per minute) termination (access) charge "a" to the termination company's network and can bill the caller for this charge. The receiving customer does not pay a termination charge for the call (this is known as a "caller pays" system; and (b) retail prices are unregulated so that networks are free to charge whatever they conclude is profit maximizing for sales to final consumers. The question is whether competition is likely to lead to efficient outcomes absent any regulatory rules.

Assume that there are 100 consumers each connected to a separate independent network who call each other. Each network sets it own access charge for terminating calls to it. The originating network incurs marginal cost c to get the call to the network interface and then a termination charge a_i to the receiving network. Assume that each originating network sets a retail price equal to c plus the average of the termination charges of the 99 networks to which it interconnects (no price differences based on the location of the termination network). In this case, the impact of an increase in a_i on the average termination price originating callers pay to call network i is very small. In this case each network has an incentive to charge high access charges because the perceived impact on the volume of calls that it will receive is small. All networks set access fees too high and the average access fee passed along to consumers by the calling networks is too high. This in turn leads to high retail prices and too little calling.

This result is most striking for a large number of networks and no network specific price discrimination. However, Laffont and Tirole (2000) show that similar results emerge when there are only two networks which have market power. Consider the case of international calls. There are two monopolies (one in each country) and each

sets a termination charge that applies to calls received from the other. Since each is a monopoly whatever the access charge is chosen by the other, this will get "marked up" by the local monopoly leading to a double marginalization problem [Laffont and Tirole, (2000, Box 5.1)].

It should be obvious as well that if two networks which compete intensely (Bertrand) with one another at retail *cooperated* in setting their respective access prices that they could agree on high access prices, increasing the perceived marginal cost at retail and the associated retail prices. The monopoly profits would then reside at the wholesale (access business) level rather than the retail level. Basically, it is profitable for each network to increase its rivals costs so that market prices rise more than do the firm's costs (Ordover, Saloner, and Salop, 1990). Accordingly, both non-cooperative and co-operative access pricing can lead to excessive retail prices. In the context of a simple duopoly model with competing firms selling differentiated products, Laffont and Tirole (2000) derive the access prices that would result if the firms compete Bertrand, derive the Ramsey-Boiteux prices for this demand and cost structure and show that the access prices that result from Bertrand competition are two high. Indeed, the socially optimal access/termination charge lies below the marginal cost of termination while the (imperfectly) competitive access price lies above the marginal cost of termination. When fixed costs are added to the model, the relationship between the competitive prices and the second-best optimal prices is ambiguous. The results can be further complicated by introducing asymmetries between the competing firms. Various extensions of these models of non-cooperative and cooperative access pricing have recently appeared in the literature. While one might make a case for regulation of access prices in this context, computing the optimal access prices in a two-way access situation would be extremely information intensive and subject to considerable potential for error.

11. Conclusions

For over 100 years economists and policymakers have refined alternative definitions of natural monopoly, developed a variety of different regulatory mechanisms and procedures to mitigate the feared adverse economic consequences of natural monopoly absent regulation, and studied the effects of price and entry regulation in practice. The pendulum of policy toward real and imagined natural monopoly problems has swung from limited regulation, to a dramatic expansion of regulation, to a gradual return to a more limited scope for price and entry regulation. Natural monopoly considerations became a rationale for extending price and entry regulation to industries that clearly did not have natural monopoly characteristics while technological and other economic changes have erased or reduced the significance of natural monopoly characteristics that may once have been a legitimate concern. However, the adverse effects of economic regulation in practice led scholars and policymakers to question whether the costs of imperfect regulation were greater than the costs of imperfect markets. These developments in turn have led to the deregulation of many industries previously subject to price and entry

regulation, to a reduction in the scope of price and entry regulation in several other industries, and to the application of better performance-based regulatory mechanisms to the remaining core natural monopoly segments of these industries.

After the most recent two decades of deregulation, restructuring, and regulatory reform, research on the regulation of the remaining natural monopoly sectors has three primary foci. First, to develop, apply and measure the effects of incentive regulation mechanisms that recognize that regulators have imperfect and asymmetric information about the firms that they regulate and utilize the information regulators can obtain in effective ways. Second, to develop and apply access and pricing rules for regulated monopoly networks that are required to support the efficient expansion of competition in previously regulated segments for which the regulated networks continue to be an essential platform to support this competition. Third, to gain a better understanding of the effects of regulation on dynamic efficiency, in terms of the effects of regulation on the development and diffusion of new services and new supply technologies. These targets of opportunity are being addressed in the scholarly literature but have been especially slow to permeate U.S. regulatory institutions. Successfully bringing this new learning to the regulatory policy arena is a continuing challenge.

References

Ai, C., Martinez, S., Sappington, D.E. (2004). "Incentive regulation and telecommunications service quality". Journal of Regulatory Economics 26 (3), 263–285.

Ai, C., Sappington, D. (2002). "The impact of state incentive regulation on the U.S. telecommunications industry". Journal of Regulatory Economics 22 (2), 133–160.

Armstrong, M., Rochet, J.-C. (1999). "Multi-dimensional screening: a users guide". European Economic Review 43, 959–979.

Armstrong, M., Sappington, D. (2003a). "Recent developments in the theory of regulation". In: Armstrong, M., Porter, R. (Eds.), Handbook of Industrial Organization, vol. III. Elsevier Science Publishers, Amsterdam, in press.

Armstrong, M., Sappington, D.M. (2003b). "Toward a synthesis of models of regulatory policy design with limited information". Mimeo.

Armstrong, M., Sappington, D. (2006). "Regulation, competition and liberalization". Journal of Economic Literature 44, 325–366.

Armstrong, M., Vickers, J. (1991). "Welfare effects of price discrimination by a regulated monopolist". Rand Journal of Economics 22 (4), 571–580.

Armstrong, M., Vickers, J. (2000). "Multiproduct price regulation under asymmetric information". Journal of Industrial Economics 48, 137–160.

Armstrong, M., Cowan, S., Vickers, J. (1994). Regulatory Reform: Economic Analysis and British Experience. MIT Press, Cambridge, MA.

Armstrong, M., Doyle, C., Vickers, J. (1996). "The access pricing problem: a synthesis". Journal of Industrial Economics 44 (2), 131–150.

Averch, H., Johnson, L.L. (1962). "Behavior of the firm under regulatory constraint". American Economic Review 52, 1059–1069.

Bacon, J.W., Besant-Jones, J. (2000). "Global electric power reform, privatization and liberalization of the electric power sector in developing countries". World Bank, Energy and Mining Sector Board Discussion Paper Series, Working Paper No. 2, June.

Bailey, E.E. (1973). Economic Theory of Regulatory Constraint. Heath and Company, Lexington Books, Lexington, D.C.

Bailey, E.E., Coleman, R.D. (1971). "The effect of lagged regulation in an Averch-Johnson model". Bell Journal of Economics 2, 278–292.

Bain, J.S. (1956). Barriers to New Competition. Harvard University Press, Cambridge, MA.

Banerjee, A. (2003). "Does incentive regulation cause degradation of telephone service quality?" Information Economics and Policy 15, 243–269.

Baron, D., Besanko, D. (1984). "Regulation, asymmetric information and auditing". Rand Journal of Economics 15 (4), 447–470.

Baron, D., Besanko, D. (1987a). "Commitment and fairness in a dynamic regulatory relationship". Review of Economic Studies 54 (3), 413–436.

Baron, D., Besanko, D. (1987b). "Monitoring, moral hazard, asymmetric information and risk sharing in procurement contracting". Rand Journal of Economics 18 (4), 509–532.

Baron, D., Myerson, R. (1982). "Regulating a monopolist with unknown costs". Econometrica 50 (4), 911–930.

Baumol, W., Bailey, E., Willig, R. (1977). "Weak invisible hand theorems on the sustainability of prices in multiproduct monopoly". American Economic Review 67 (3), 350–365.

Baumol, W., Klevorick, A.K. (1970). "Input choices and rate of return regulation: an overview of the discussion". Bell Journal of Economics and Management Science 1 (2), 169–190.

Baumol, W., Ordover, J., Willig, R. (1997). "Parity pricing and its critics: a necessary condition for the provision of bottleneck services to competitors". Yale Journal on Regulation 14 (1), 145–164.

Baumol, W., Sidak, G. (1994). "The pricing of inputs sold to competitors". Yale Journal on Regulation 11 (1), 171–202.

Baumol, W.J., Bradford, D.F. (1970). "Optimal departures from marginal cost pricing". American Economic Review 60, 265–283.

Baumol, W.J., Panzar, J., Willig, R.D. (1982). Contestible Markets and the Theory of Industry Structure. Harcourt Brace Javanovich, New York.

Beesley, M., Littlechild, S. (1989). "The regulation of privatized monopolies in the United Kingdom". Rand Journal of Economics 20 (3), 454–472.

Bernstein, J.I., Sappington, D.M. (1999). "Setting the X-factor in price cap regulation plans". Journal of Regulatory Economics 16, 5–25.

Berg, S.V., Tschirhart, J. (1988). Natural Monopoly Regulation: Principles and Practice. Cambridge University Press, Cambridge.

Boiteux, M. (1960). "Peak load pricing". Journal of Business 33, 157–179. Translated from the original in French published in 1951.

Boiteux, M. (1971). "On the management of public monopolies subject to budget constraint". Journal of Economic Theory 3, 219–240. Translated from the original in French and published in Econometrica in 1956.

Bonbright, J.C. (1961). Principles of Public Utility Rates. Columbia University Press, New York.

Borenstein, S. (2005). "Time-varying retail electricity prices: theory and practice". In: Griffin, Puller (Eds.), Electricity Deregulation: Choices and Challenges. University of Chicago Press, Chicago.

Braeutigam, R. (1989). "Optimal prices for natural monopolies". In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. II. Elsevier Science Publishers, Amsterdam.

Braeutigam, R., Magura, M., Panzar, J. (1997). "The effects of incentive regulation on local telephone service rates". Northwestern University, mimeo.

Brennan, T. (1989). "Regulating by capping prices". Journal of Regulatory Economics 1 (2), 133–147.

Brown, S.J., Sibley, D.S. (1986). The Theory of Public Utility Pricing. Cambridge University Press, Cambridge.

Cabral, L., Riordan, M. (1989). "Incentives for cost reduction under price cap regulation". Journal of Regulatory Economics 1 (2), 93–102.

Carlton, D. (1977). "Peak load pricing with stochastic demand". American Economic Review 67, 1006–1010.

Carlton, D., Perloff, J. (2004). Modern Industrial Organization, 4th edn. Addison-Wesley, Boston, MA.
Carrington, R., Coelli, T., Groom, E. (2002). "International benchmarking for monopoly price regulation: the case of Australian gas distribution". Journal of Regulatory Economics 21, 191–216.
Christiansen, L.R., Greene, W.H. (1976). "Economies of scale in U.S. electric power generation". Journal of Political Economy 84, 655–676.
Clark, J.M. (1911). "Rates for public utilities". American Economic Review 1 (3), 473–487.
Clark, J.M. (1913). "Frontiers of regulation and what lies beyond". American Economic Review 3 (1), 114–125.
Clemens, E.W. (1950). Economics of Public Utilities. Appleton-Century-Crofts, New York.
Cowing, T.G. (1974). "Technical change and scale economies in an engineering production function: the case of steam electric power". Journal of Industrial Economics 23, 135–152.
Crandall, R.W., Hausman, J.A. (2000). "Competition in U.S. telecommunications services: effects of the 1996 legislation". In: Peltzman, S., Winston, C. (Eds.), Deregulation of Network Industries. Brookings Institution Press, Washington, D.C.
Crandall, R.W., Waverman, L. (1995). Talk is Cheap: The Promise of Regulatory Reform in North American. Brookings, Washington, D.C.
Crawford, G. (2000). "The impact of the 1992 cable act on consumer demand and welfare: a discrete-choice, differentiated products approach". Rand Journal of Economics 31, 422–450.
Crew, M.A., Kleinforfer, P.R. (1976). "Peak load pricing with a diverse technology". Bell Journal of Economics 7, 207–231.
Crew, M.A., Kleinforfer, P.R. (1986). The Economics of Public Utility Regulation. MIT Press, Cambridge, MA.
Dana, J. (1993). "The organization and scope of agents: regulating multiproduct industries". Journal of Economic Theory 59 (2), 288–310.
Demsetz, H. (1968). "Why regulate utilities". Journal of Law and Economics 11 (1), 55–65.
Dreze, J. (1964). "Contributions of French economists to theory and public policy". American Economic Review 54 (4), 2–64.
Ely, R. (1937). Outlines of Economics. MacMillan, New York.
Estache, A., Kouasi, E. (2002). "Sector organization, governance and the inefficiencies of African water utilities". World Bank Policy Research Working Paper No. 2890, September.
Estache, A., Guasch, J.-L., Trujillo, L. (2003). "Price caps, efficiency payoffs, and infrastructure contract renegotiation in Latin America". Mimeo.
Estache, A., Rossi, M.A., Ruzzier, C.A. (2004). "The case for international coordination of electricity regulation: evidence from the measurement of efficiency in South America". Journal of Regulatory Economics 25 (3), 271–295.
Evans, D.S. (1983). Breaking Up Bell: Essays in Industrial Organization and Regulations. North-Holland, New York.
Farrer, T.H. (1902). The State in Relation to Trade. Macmillan, London.
Faulhaber, G.R. (1975). "Cross-subsidization: pricing in public utility enterprises". American Economic Review 65, 966–977.
Fiorina, M. (1982). "Legislative choice of regulatory forums: legal process or administrative process". Public Choice 39, 33–36.
Fraquelli, G., Picenza, M., Vannoni, D. (2004). "Scope and scale economies from multi-utilities: evidence from gas, water and electricity combinations". Applied Economcs 36 (18), 2045–2057.
Gagnepain, P., Ivaldi, M. (2002). "Incentive regulatory policies: the case of public transit in France". Rand Journal of Economics 33, 605–629.
Gasmi, F., Laffont, J.J., Sharkey, W.W. (2002). "The natural monopoly test reconsidered: an engineering process-based approach to empirical analysis in telecommunications". International Journal of Industrial Organization 20 (4), 435–459.
Giannakis, D., Jamasb, T., Pollitt, M. (2004). "Benchmarking and incentive regulation of quality of service: an application to the U.K. distribution utilities". Cambridge Working Papers in Economics CWEP 0408, Department of Applied Economics, University of Cambridge.

Gilbert, R., Newbery, D. (1994). "The dynamic efficiency of regulatory constitutions". Rand Journal of Economics 26 (2), 243–256.

Gilligan, T.W., Marshall, W.M., Weingast, B.R. (1989). "Regulation and the theory of legislative choice: the interstate commerce act of 1887". Journal of Law and Economics 32, 35–61.

Gilligan, T.W., Marshall, W.J., Weingast, B.R. (1990). "The economic incidence of the interstate commerce act of 1887: a theoretical and empirical analysis of the short-haul pricing constraint". Rand Journal of Economics 21, 189–210.

Glaeser, M.G. (1927). Outlines of Public Utility Economics. MacMillan, New York.

Goldberg, V.C. (1976). "Regulation and administered contracts". Bell Journal of Economics 7, 426–448.

Goolsbee, A., Petrin, A. (2004). "The consumer gains from direct broadcast satellites and the competition with cable television". Econometrica 72 (2), 351–381.

Greene, W.H., Smiley, R.H. (1984). "The effectiveness of utility regulation in a period of changing economic conditions". In: Marchand, M., Pestieau, P., Tulkens, H. (Eds.), The Performance of Public Enterprise: Concepts and Measurement. Elsevier, Amsterdam.

Greenstein, S., McMaster, S., Spiller, P. (1995). "The effect of incentive regulation on infrastructure modernization: local exchange companies' deployment of digital technology". Journal of Economics & Management Strategy 4, 187–236.

Hadlock, C.J., Lee, D.S., Parrino, R. (2002). "Chief executive officer careers in regulated environments: evidence from electric and gas utilities". Journal of Law and Economics 45, 535–564.

Hammond, C.J., Johnes, G., Robinson, T. (2002). "Technical efficiency under alternative regulatory regimes". Journal of Regulatory Economics 22 (3), 251–270.

Hausman, J.A. (1997). "Valuing the effects of regulation on new services in telecommunications". Brookings Papers on Economics Activity: Microeconomics 1–54.

Hausman, J.A. (1998). "Taxation by telecommunications regulation". NBER/Tax Policy and the Economy 12 (1), 29–49.

Hausman, J.A. (1999). "The effects of sunk costs in telecommunications regulation". In: Alleman, J., Noam, E. (Eds.), Real Options: The New Investment Theory and its Applications for Telecommunications Economics. Kluwer Academic, Norwell, MA.

Hausman, J.A. (2002). "Mobile telephone". In: Cave, M.E. et al. (Eds.), Handbook of Telecommunications Economics. Elsevier Science.

Hausman J.A. (2003). "Regulated costs and prices of telecommunications". In: Madden, G. (Ed.), Emerging Telecommunications Networks. Edward Elgar Publishing.

Hausman, J.A., Myers, S.C. (2002). "Regulating the United States railroads: the effects of sunk costs and asymmetric risk". Journal of Regulatory Economics 22, 287–310.

Hausman, J.A., Tardiff, T., Belinfante, A. (1993). "The effects of the breakup of AT&T on telephone penetration in the United States". American Economic Review 83, 178–184.

Hendricks, W. (1977). "Regulation and labor earnings". Bell Journal of Economics 8, 483–496.

Hubbard, T. (2001). "Contractual form and market thickness in trucking". Rand Journal of Economics 32 (2), 369–386.

Hubard, T. (2003). "Information, decisions and productivity: on board computers and capacity utilization in trucking". American Economic Review 94 (4), 1328–1353.

Hughes, T.P. (1983). Networks of Power: Electrification in Western Society 1880–1930. Johns Hopkins University Press, Baltimore, MD.

Isaac, R.M. (1991). "Price cap regulation: a case study of some pitfalls of implementation". Journal of Regulatory Economics 3 (2), 193–210.

Jamasb, T., Pollitt, M. (2001). "Benchmarking and regulation: international electricity experience". Utilities Policy 9, 107–130.

Jamasb, T., Pollitt, M. (2003). "International benchmarking and regulation: an application to European electricity distribution utilities". Energy Policy 31, 1609–1622.

Jarrell, G.A. (1978). "The demand for state regulation of the electric utility industry". Journal of Law and Economics 21, 269–295.

Joskow, P.L. (1972). "The determination of the allowed rate of return in a formal regulatory hearing". Bell Journal of Economics and Management Science 3, 633–644.
Joskow, P.L. (1973). "Pricing decisions of regulated firms". Bell Journal of Economics and Management Science 4, 118–140.
Joskow, P.L. (1974). "Inflation and environmental concern: structural change in the process of public utility price regulation". Journal of Law and Economics 17, 291–327.
Joskow, P.L. (1976). "Contributions to the theory of marginal cost pricing". Bell Journal of Economics 7 (1), 197–206.
Joskow, P.L. (1989). "Regulatory failure, regulatory reform and structural change in the electric power industry". Brookings Papers on Economic Activity: Microeconomic 125–199.
Joskow, P.L. (1997). "Restructuring, competition and regulatory reform in the U.S. electricity sector". Journal of Economic Perspectives 11 (3), 119–138.
Joskow, P.L. (2000). "Deregulation and regulatory reform in the U.S. electric power industry". In: Peltzman, S., Winston, C. (Eds.), Deregulation of Network Industries. Brookings Institution Press, Washington, D.C.
Joskow, P.L. (2005a). "Transmission policy in the United States". Utilities Policy 13, 95–115.
Joskow, P.L. (2005b). "Regulation and deregulation after 25 years". International Review of Industrial Organization 26, 169–193.
Joskow, P.L. (2006). "Incentive regulation in theory and practice". NBER Regulation Project, mimeo. (http://econ-www.mit.edu/faculty/download_pdf.php?id=1220.)
Joskow, P.L., Noll, R.G. (1981). "Regulation in theory and practice: an overview". In: From, G. (Ed.), Studies in Public Regulation. MIT Press, Cambridge, MA.
Joskow, P.L., Noll, R.G. (1999). "The Bell doctrine: applications in telecommunications, electricity and other network industries". Stanford Law Review 51 (5), 1249–1315.
Joskow, P.L., Rose, N.L. (1985). "The effects of technological change, experience and environmental regulation on the costs of coal-burning power plants". Rand Journal of Economics 16 (1), 1–27.
Joskow, P.L., Rose, N.L. (1989). "The effects of economic regulation". In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. II. North-Holland, Amsterdam.
Joskow, P.L., Rose, N.L., Wolfram, C.D. (1996). "Political constraints on executive compensation: evidence from the electric utility industry". Rand Journal of Economics 27, 165–182.
Joskow, P.L., Schmalensee, R. (1983). Markets for Power. MIT Press, Cambridge, MA.
Joskow, P.L., Schmalensee, R. (1986). "Incentive regulation for electric utilities". Yale Journal on Regulation 4, 1–49.
Joskow, P.L., Tirole, J. (2005). "Retail electricity competition". Rand Journal of Economics (in press). (http://econ-www.mit.edu/faculty/download_pdf.php?id=918.)
Joskow, P.L., Tirole, J. (2006). "Reliability and competitive electricity markets". Rand Journal of Economics (in press). (http://econ-www.mit.edu/faculty/download_pdf.php?id=917.)
Kahn, A.E. (1970). The Economics of Regulation: Principles and Institutions, volume I. Wiley, New York.
Katz, M., Shapiro, C. (1986). "Technology adoption in the presence of network externalities". Journal of Political Economy 94, 822–841.
Kaysen, C., Turner, D. (1959). Antitrust Policy: An Economic ad Legal Analysis. Harvard University Press, Cambridge, MA.
Klemperer, P. (2002). "What really matters in auction design". Journal of Economic Perspectives 16, 169–189.
Klevorick, A.K. (1971). "The optimal fair rate of return". Bell Journal of Economics 2, 122–153.
Klevorick, A.K. (1973). "The behavior of the firm subject to stochastic regulatory review". Bell Journal of Economics 4, 57–88.
Kolbe, L., Tye, W. (1991). "The Duquesne opinion: how much 'Hope' is there for investors in regulated firms?" Yale Journal on Regulation 8 (1), 113–157.
Kolko, G. (1965). Railroads and Regulation 1877–1916. Princeton University Press, Princeton.
Kridel, D., Sappington, D., Weisman, D. (1996). "The effects of incentive regulation in the telecommunications industries: a survey". Journal of Regulatory Economics 18, 269–306.

Kwoka, J. (1993). "Implementing price caps in telecommunications". Journal of Policy Analysis and Management 12 (4), 722–756.
Laffont, J.-J. (1999). "Competition, information and development". In: Annual World Bank Conference on Development Economics 1998. The World Bank, Washington, D.C.
Laffont, J.-J., Rey, P., Tirole, J. (1998a). "Network competition: I. Overview and nondiscriminatory pricing". Rand Journal of Economics 29, 1–37.
Laffont, J.-J., Rey, P., Tirole, J. (1998b). "Network competition: II. Price discrimination". Rand Journal of Economics 29, 38–56.
Laffont, J.-J., Tirole, J. (1986). "Using cost observations to regulate firms". Journal of Political Economy 94 (3), 614–641.
Laffont, J.-J., Tirole, J. (1988a). "Auctioning incentive contracts". Journal of Political Economy 95 (5), 921–937.
Laffont, J.-J., Tirole, J. (1988b). "The dynamics of incentive contracts". Econometrica 56 (5), 1153–1176.
Laffont, J.-J., Tirole, J. (1990a). "Adverse selection and renegotiation in procurement". Review of Economic Studies 57 (4), 597–626.
Laffont, J.-J., Tirole, J. (1990b). "Optimal bypass and cream-skimming". American Economic Review 80 (4), 1041–1051.
Laffont, J.-J., Tirole, J. (1993). A Theory of Incentives in Regulation and Propcurement. MIT Press, Cambridge, MA.
Laffont, J.-J., Tirole, J. (1996). "Creating competition through interconnection: theory and practice". Journal of Regulatory Economics 10 (3), 227–256.
Laffont, J.-J., Tirole, J. (2000). Competition in Telecommunication. MIT Press, Cambridge, MA.
Levy, B., Spiller, P. (1994). "The institutional foundations of regulatory commitment: a comparative analysis of telecommunications". Journal of Law, Economics and Organization 10 (2), 201–246.
Lewis, T., Sappington, D.M. (1988a). "Regulating a monopolist with unknown demand". American Economic Review 78 (5), 986–998.
Lewis, T., Sappington, D.M. (1988b). "Regulating a monopolist with unknown demand and cost functions". Rand Journal of Economics 18 (3), 438–457.
Lewis, T., Sappington, D. (1989). "Regulatory options and price cap regulation". Rand Journal of Economics 20 (3), 405–416.
Loeb, M., Magat, W. (1979). "A decentralized method for utility regulation". Journal of Law and Economics 22 (2), 399–404.
Lowry, E.D. (1973). "Justification for regulation: the case for natural monopoly". Public Utilities Fortnightly November 8, 1–7.
Lyon, T. (1996). "A model of the sliding scale". Journal of Regulatory Economics 9 (3), 227–247.
Marshall, A. (1890). Principles of Economics, 8ht edn. MacMillan, London. (1966).
Mathios, A.D., Rogers, R.P. (1989). "The impact of alternative forms of state regulation of AT&T direct-dial, long-distance telephone rates". Rand Journal of Economics 20, 437–453.
McCubbins, M.D. (1985). "The legislative design of regulatory structure". American Journal of Political Science 29, 721–748.
McCubbins, M.D., Noll, R.G., Weingast, B.R. (1987). "Administrative procedures as instruments of corporate control". Journal of Law, Economics and Organization 3, 243–277.
McDonald, F. (1962). Insull. University of Chicago Press, Chicago.
Megginson, W., Netter, J. (2001). "From state to market: a survey of empirical studies of privatization". Journal of Economic Literature 39, 321–389.
Mullin, W.P. (2000). "Railroad revisionists revisited: stock market evidence from the progressive era". Journal of Regulatory Economics 17 (1), 25–47.
Myers, S.C. (1972a). "The application of finance theory to public utility rate cases". Bell Journal of Economics and Management Science 3 (1), 58–97.
Myers, S.C. (1972b). "On the use of β in regulatory proceedings". Bell Journal of Economics and Management Science 3 (2), 622–627.

National Civic Federation (1907). Municipal and Private Operation of Public Utilities, volume I. National Civic Federation, New York.

Nelson, J.R. (1964). Marginal Cost Pricing in Practice. Prentice-Hall, Englewood-Cliffs, N.J.

Newbery, D.M., Pollitt, M.G. (1997). "The restructuring and privatisation of Britain's CEGB: was it worth it?" Journal of Industrial Economics 45 (3), 269–303.

Noll, R.G. (1989). "Economic perspectives on the politics of regulation". In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. II. North-Holland, Amsterdam.

Office of Gas and Electricity Markets (OFGEM) (2004a). "Electricity distribution price control review: policy document". March, London, UK.

Office of Gas and Electricity Markets (OFGEM) (2004b). "Electricity distribution price control review: final proposals". 265/04, November, London.

Office of Gas and Electricity Markets (OFGEM) (2004c). "NGC system operator incentive scheme from April 2005: initial proposals". December, London.

Office of Gas and Electricity Markets (OFGEM) (2004d). "Electricity transmission network reliability incentive scheme: final proposals". December, London.

Owen, B., Brauetigam, R. (1978). The Regulation Game: Strategic Use of the Administrative Process. Ballinger Publishing Company, Cambridge, MA.

Ordover, J.A., Saloner, G., Salop, S.C. (1990). "Equilibrium vertical foreclosure". American Economic Review 80, 127–142.

Palmer, K. (1992). "A test for cross subsidies in local telephone rates: do business customers subsidize residential customers?" Rand Journal of Economics 23, 415–431.

Panzar, J.C. (1976). "A neoclassical approach to peak load pricing". Bell Journal of Economics 7, 521–530.

Peltzman, S. (1989). "The economic theory of regulation after a decade of deregulation". Brookings Papers on Economic Activity: Microeconomics 1–60.

Phillips, C.F. Jr. (1993). The Regulation of Public Utilities: Theory and Practice. Public Utilities Report, Inc., Arlington, VA.

Pindyck, R. (2004). "Pricing capital under mandatory unbundling and facilities sharing". December, mimeo.

Pindyck, R., Rubinfeld, D. (2001). Microeconomics, 5th edn. Prentice-Hall, Upper Saddle River, N.J.

Posner, R.A. (1969). "Natural monopoly and regulation". Stanford Law Review 21, 548–643.

Posner, R.A. (1971). "Taxation by regulation". Bell Journal of Economics and Management Science 2, 22–50.

Posner, R.A. (1974). "Theories of economic regulation". Bell Journal of Economics 5, 335–358.

Posner, R.A. (1975). "The social cost of monopoly and regulation". Journal of Political Economy 83, 807–827.

Prager, R.A. (1989a). "Using stock price data to measure the effects of regulation: the interstate commerce act and the railroad industry". Rand Journal of Economics 20, 280–290.

Prager, R.A. (1989b). "Franchise bidding for natural monopoly". Journal of Regulatory Economics 1 (2), 115–132.

Prager, R.A. (1990). "Firm behavior in franchise monopoly markets". Rand Journal of Economics 12, 211–225.

Ramsey, F. (1927). "A contribution to the theory of taxation". Economic Journal 37, 47–61.

Riordan, M. (1984). "On delegating price authority to a regulated firm". Rand Journal of Economics 15 (1), 108–115.

Rochet, J.C., Tirole, J. (2003). "Platform competition in two-sided markets". Journal of the European Economic Association 1 (4), 990–1029.

Rochet, J.C., Tirole, J. (2004). "Two-sided markets: an overview". Institute d'Economie Industrielle, March, mimeo.

Rose, N.L. (1987). "Labor rent sharing and regulation: evidence from the trucking industry". Journal of Political Economy 95, 1146–1178. December.

Rose, N.L., Joskow, P.L. (1990). "The diffusion of new technology: evidence from the electric utility industry". Rand Journal of Economics 21 (3), 354–373.

Rose, N., Markiewicz, K., Wolfram, C. (2004). "Does competition reduce costs? Reviewing the impact of regulatory restructuring on U.S. electric generation efficiency". MIT CEEPR Working Paper 04-018. (http://web.mit.edu/ceepr/www/2004-018.pdf.)

Rudnick, H., Zolezzi, J. (2001). "Electric sector deregulation and restructuring in Latin America: lessons to be learnt and possible ways forward". IEEE Proceedings Generation, Transmission and Distribution 148, 180–184.

Salinger, M.E. (1984). "Tobin's q, unionization, and the concentration-profits relationship". Rand Journal of Economics 15, 159–170.

Salinger, M.E. (1998). "Regulating prices to equal forward-looking costs: cost-based prices or price-based cost". Journal of Regulatory Economics 14, 149–163.

Sappington, D.M. (1980). "Strategic firm behavior under a dynamic regulatory adjustment process". Bell Journal of Economics 11 (1), 360–372.

Sappington, D.M. (2003). "The effects of incentive regulation on retail telephone service quality in the United States". Review of Network Economics 2 (3), 355–375.

Sappington, D., Ai, C. (2005). "Reviewing the impact of incentive regulation on U.S. telephone service quality". Utilities Policy 13 (3), 201–210.

Sappington, D., Sibley, D. (1988). "Regulating without cost information: the incremental surplus subsidy scheme". International Economic Review 31 (2), 297–306.

Sappington, D., Sibley, D. (1990). "Regulating without cost information: further observations". International Economic Review 31 (4), 1027–1029.

Sappington, D., et al. (2001). "The state of performance based regulation in the U.S. electric utility industry". Electricity Journal, 71–79.

Schleifer, A. (1985). "A theory of yardstick competition". Rand Journal of Economics 16 (3), 319–327.

Schmalensee, R. (1979). The Control of Natural Monopolies. Lexington Books, Lexington, MA.

Schmalensee, R. (1981). "Output and welfare implications of monopolistic third-degree price discrimination". American Economic Review 71, 242–247.

Schmalensee, R. (1989a). "An expository note on depreciation and profitability under rate of return regulation". Journal of Regulatory Economics 1 (3), 293–298.

Schmalensee, R. (1989b). "Good regulatory regimes". Rand Journal of Economics 20 (3), 417–436.

Sharfman, I.L. (1928). "Valuation of public utilities: discussion". American Economic Review 18 (1), 206–216.

Sharkey, W.W. (1982). The Theory of Natural Monopoly. Cambridge University Press, Cambridge.

Sheshinski, E. (1971). "Welfare aspects of regulatory constraint". American Economic Review 61, 175–178.

Sibley, D. (1989). "Asymmetric information, incentives and price cap regulation". Rand Journal of Economics 20 (3), 392–404.

Sidak, G., Spulber, D. (1997). Deregulatory Takings and the Regulatory Contract. Cambridge University Press, Cambridge.

Spence, M. (1975). "Monopoly, quality and regulation". Bell Journal of Economics 6 (2), 417–429.

Spiegel, Y., Spulber, D. (1994). "The capital structure of regulated firms". Rand Journal of Economics 25 (3), 424–440.

Spiller, P. (1990). "Politicians, interest groups and regulators: a multiple principal agent theory of regulation". Journal of Law and Economics 33 (1), 65–101.

Steiner, P. (1957). "Peak loads and efficient pricing". Quarterly Journal of Economics 71 (4), 585–610.

Stigler, G.J. (1971). "The theory of economic regulation". Bell Journal of Economics and Management Science 2, 3–21.

Stigler, G.J., Friedland, C. (1962). "What can regulators regulate: the case of electricity". Journal of Law and Economics 5, 1–16.

Sutton, J. (1991). Sunk Costs and Market Structure. MIT Press, Cambridge, MA.

Tardiff, T., Taylor, W. (1993). Telephone Company Performance Under Alternative Forms of Regulation in the U.S. National Economic Research Associates.

Teeples, R., Glyer, D. (1987). "Cost of water delivery systems: specific and ownership effects". Review of Economics and Statistics 69, 399–408.

Tirole, J. (1988). The Theory of Industrial Organization. MIT Press, Cambridge, MA.
Troxel, E. (1947). Economics of Public Utilities. Rineheart & Company, New York.
Turvey, R. (1968a). "Peak load pricing". Journal of Political Economy 76, 101–113.
Turvey, R. (1968b). Optimal Pricing and Investment in Electricity Supply: An Essay in Applied Welfare Economics. MIT Press, Cambridge, MA.
Vickers, J. (1995). "Competition and regulation in vertically related markets". Review of Economic Studies 62 (1), 1–17.
Vickers, J., Yarrow, G. (1991). "Economic perspectives on privatization". Journal of Economic Perspectives 5, 111–132.
Vogelsang, I. (2003). "Price regulation of access to telecommunications networks". Journal of Economic Literature 41, 830–862.
Vogelsang, I., Finsinger, J. (1979). "A regulatory adjustment process for optimal pricing of multiproduct firms". Bell Journal of Economics 10 (1), 151–171.
Weiman, D.F., Levin, R.C. (1994). "Preying for monopoly? The case of southern Bell Telephone, 1894–1912". Journal of Political Economy 102 (1), 103–126.
Weingast, B.R., Moran, M.J. (1983). "Bureaucratic discretion or congressional control? Regulatory policy-making at the Federal Trade Commission". Journal of Political Economy 91, 765–780.
Weitzman, M. (1980). "The ratchet principle and performance incentives". Bell Journal of Economics 11 (1), 302–308.
Weitzman, M.A. (1983). "Contestable markets: an uprising in the theory of industry structure: comment". American Economic Review 73 (3), 486–487.
Williamson, O.E. (1976). "Franchise bidding for natural monopolies: in general and with respect to CATV". Bell Journal of Economics 7 (1), 73–104.
Williamson, O.E. (1985). The Economic Institutions of Capital: Firms, Markets and Contracting. Free Press, New York.
Williamson, O.E. (1996). The Mechanisms of Governance. Oxford University Press, New York.
Willig, R. (1978). "Pareto-superior non-linear outlay schedules". Bell Journal of Economics 9 (1), 56–69.
Willig, R. (1979). "The theory of network access pricing". In: Trebing, H. (Ed.), Issues in Public Utility Regulation. Michigan State University Press, East Lansing, MI.
Winston, C. (1993). "Economic deregulation: days of reckoning for microeconomists". Journal of Economic Literature 31 (3), 1263–1289.
Winston, C., Peltzman, S. (2000). Deregulation of Network Industries. Brookings Institution Press, Washington, D.C.
Zupan, M. (1989a). "Cable franchise renewals: do incumbent firms behave opportunistically". Rand Journal of Economics 20 (4), 473–482.
Zupan, M. (1989b). "The efficacy of franchise bidding schemes for CATV: some systematic evidence". Journal of Law and Economics 32 (2), 401–456.

Chapter 17

EMPLOYMENT LAW

CHRISTINE JOLLS*

School of Law, Yale University, and National Bureau of Economic Research

Contents

* Thanks to Louis Kaplow, Daniel Klaff, Cass Sunstein, and participants at the March 13, 2004, conference for contributors to the *Handbook of Law and Economics* for helpful comments and discussions, and to Dina Mishra and Kenneth Moon for exceptional research assistance.

Handbook of Law and Economics, Volume 2
Edited by A. Mitchell Polinsky and Steven Shavell

DOI: 10.1016/S1574-0730(07)02017-8

Abstract

Legal rules governing the employer-employee relationship are many and varied. Economic analysis has illuminated both the efficiency and the effects on employee welfare of such rules, as described in this chapter. Topics addressed below include workplace safety mandates, compensation systems for workplace injuries, privacy protection in the workplace, employee fringe benefits mandates, targeted mandates such as medical and family leave, wrongful discharge laws, unemployment insurance systems, minimum wage rules, and rules requiring that employees receive overtime pay. Both economic theory and empirical evidence are considered.

Keywords

Employment law, workplace safety, workplace privacy, benefits mandates, wrongful discharge law, overtime pay rules

JEL classification: J08, J18, J38, J80, K00, K31, K32

In a modern economy, individuals usually rely on paid work to meet their basic material needs. The present chapter is concerned with the economic analysis of laws governing this ubiquitous employer-employee relationship. Such laws have proliferated in both number and scope across nations, and economic analysis of these laws' desirability and effects has accordingly attracted significant attention.

Within a pervasively unionized economy, the body of what is referred to in the United States as "labor law" plays an important role in regulating the employer-employee relationship. This body of law governs the practices and treatment of labor unions. In countries with only moderate levels of unionization, however, the central responsibility for legal regulation of the employer-employee relationship falls to what is referred to in the United States as "employment law." Employment law governs the treatment of individual employees regardless of their union status. Areas of regulation include workplace safety and privacy, employee fringe benefits, workplace leave, job security, and the payment of wages. The present chapter is concerned with the economic analysis of these employment law rules.

An obvious but critical starting point in the economic analysis of employment law is that legal regulation of the treatment of employees typically takes place against the backdrop of a market relationship. This market relationship often imposes significant limits on the prospects for using employment law purely for the purpose of transferring power, wealth, or other entitlements to employees—although employment law is often enacted with such motives as the law's declared purpose. In the area of mandated leave from employment, for instance, if employment law seeks to better employees' situation by specifying minimum entitlements to leave from work, it is possible that the end result will be to worsen employees' situation as wages or employment levels adjust in response to the new legal requirements. Because of the way in which the market constrains the prospects for using employment law purely to effect transfers of resources, the economic analysis of employment law in this chapter gives primary emphasis to market failures in the employer-employee relationship. In the presence of a market failure, legal intervention through employment law may both enhance efficiency and make employees better off.

This chapter situates the major areas of employment law within this market-failure analytic framework; importantly, it also identifies certain areas in which legal intervention may help targeted employees even in the absence of market failure (Sections 6 and 9 below). At the broadest level, the chapter seeks to describe both theoretically and empirically the degree to which major forms of legal regulation of the employer-employee relationship may enhance efficiency and make employees better off. Section 1 of the chapter briefly presents the basic framework used throughout much of the chapter. Sections 2–10 consider specific areas of employment law, including workplace safety regulation, privacy protection in the workplace, fringe benefits mandates, targeted mandates such as medical and family leave, wrongful-discharge laws, minimum wage requirements, and rules requiring overtime pay. A separate chapter of this *Handbook* (Chapter 18, by John Donohue, 2007) considers antidiscrimination requirements in employment law and other domains such as housing and education.

1. Framework

As suggested above, the analysis in this chapter rests on the assumption that employment law operates to regulate market relationships between employers and employees. Obviously, it is possible to organize an economy in which wages and conditions of employment are set not by private actors in employment markets but instead by a central planner; such non-market settings would naturally require a different analysis.

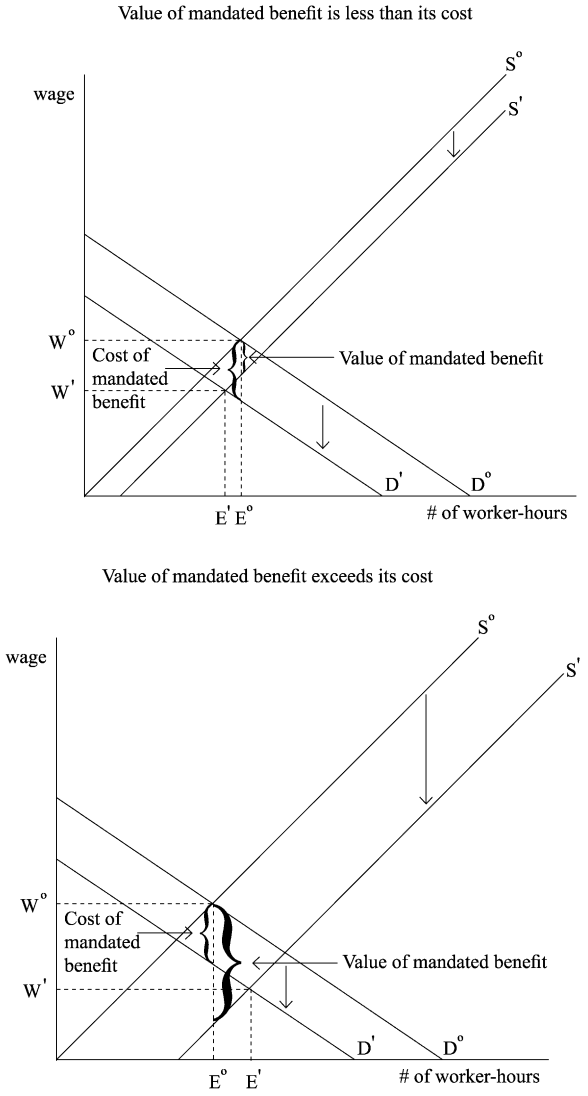
Within a market setting, when wages and conditions of employment are determined by the interaction of labor supply and labor demand, legal intervention not tied to an identified market failure will often, although not always, reduce both efficiency and employee welfare (Section 1.1). However, a number of important market failures may characterize the employer-employee relationship (Section 1.2).

1.1. Employment law in the absence of market failure

Most of the rules of employment law considered in this chapter reflect the imposition of a legally prescribed term governing the conditions of work into the parties' employment contract. This subsection describes the typical effects of such rules in the absence of market failure. Section 6 below shows how the analysis differs for rules that target the work conditions of discrete, identifiable subgroups of employees. Note that the framework described here does not apply to the rules establishing minimum wage and overtime pay requirements or to the rules governing discharge of employees because these forms of legal regulation directly operate on the wage or employment level in the regulated employment markets rather than on the conditions of work; these forms of legal regulation are covered separately in Sections 7–10 below.

In speaking of legally prescribed terms within a contract, it is important to distinguish between mandatory and default terms. Mandatory terms are ones that cannot be changed by the parties even if they express a desire to do so. For instance, if the Occupational Safety and Health Act (OSHA), which regulates workplace safety in the United States, requires that a particular safety measure be adopted, employers and employees may not avoid the requirement simply by agreeing through contract that it does not apply. Default terms, by contrast, are terms that apply unless the parties reach a contrary agreement. Most employment law rules are mandatory rules, and the analysis to follow assumes that the rule in question is a mandatory one. The specific employment law rules discussed in the remainder of this chapter are all mandatory ones.

The legally prescribed terms that employment law specifies for the employer-employee relationship typically require employers to provide something of value—a safer workplace; privacy; certain fringe benefits; leaves from work in specified circumstances—to employees. Accordingly, within a simple labor supply and demand framework with no market failure of any sort, these rules will produce a downward shift in the labor supply curve by the amount of the value of the mandated benefit and a downward shift in the labor demand curve by the amount of the cost of the mandated benefit (Summers, 1989). These effects are depicted graphically in Figure 1.



Notes: Employees are assumed to be demanded and paid in accordance with their marginal revenue product of labor (rather than, for example, earning efficiency wages), and the mandated benefit is assumed to be a variable-cost rather than a fixed-cost one.

Figure 1. Labor market effects of legally mandated benefits in the absence of market failure.

Within this simple framework, it is obvious that the effects of a particular employment law rule on efficiency and employee welfare turn on the relative magnitude of the labor supply and labor demand shifts. If the downward shift in the labor supply curve (S° to S' in Figure 1) is less than the downward shift in the labor demand curve, then the wage will fall by more than the value of the legally mandated benefit to employees in the employment market in question, and employment (and, with it, efficiency and employee welfare) will fall (Figure 1, top panel). If, by contrast, the downward shift in the labor supply curve is larger than the downward shift in the labor demand curve, then the wage will fall by less than the value of the legally mandated benefit to employees, and employment (and with it, efficiency and employee welfare) will rise (Figure 1, bottom panel). But if that were the case, then, given the assumed lack of any form of market failure, employers would have offered the benefit without any need for a legal mandate.[1]

Thus, in the absence of market failure, employment law rules will generally reduce both efficiency and employee welfare. However, employment markets may fail for a variety of reasons. Section 1.2 below describes the most commonly discussed market failures in the employer-employee relationship.

1.2. Market failures in the employer-employee relationship

As described above, market failures provide an organizing paradigm for the analysis of many major areas of employment law. A number of market failures may occur in the employer-employee relationship, as described below.

1.2.1. Information failures

Of central importance in the employment setting are possible information failures. Information failures occur when some market participants lack information that bears upon their decisions in that market. In employer-employee relationships, both parties may suffer from information failures.

1.2.1.1. Employee-side information failures Some aspects of the employment relationship are likely to be relatively transparent to employees. Wages, for instance, are an aspect of the relationship about which employees will usually have good information. By contrast, the magnitude of the risk of long-term occupational disease is something about which employees will often not be well informed.

Employee-side information failures may be modeled in two distinct ways. One possibility is that employees with limited information are aware of their informational limits

[1] This analysis assumes that there is no binding legally specified minimum wage in the employment market in question. At least in the United States, the minimum wage is sufficiently low in most jurisdictions that it is not binding in most employment markets. With a binding minimum wage (meaning the market wage is near or below the legally prescribed minimum wage) the effects of legally mandated terms would be felt in employment levels rather than wages.

and take rational steps in response to those limits. As described below, this type of modeling assumption is often adopted in analyzing employer-side information failures. A second possibility is that employees with limited information either are not aware of their informational limits or do not change their behavior in response to such awareness. This is the usual modeling assumption adopted in analyzing employee-side information failures and is the assumption used in this chapter for analyzing such information failures.

If employees are unaware of some aspect of the employment relationship that would affect their willingness to supply labor, then observed labor supply will differ from employees' "true" willingness to supply labor. In the case of workplace safety, for instance, employees may lack adequate information about risks and harms and, as a result, may oversupply labor at a given wage rate. In terms of Figure 1 above, the labor supply curve is shifted toward S', and employees behave *as if* their workplace were covered by an employment term fostering safety even though it is not. In such cases, the information failure means that some transactions that take place are inefficient—because the cost of supplying labor (measured by S° in Figure 1) in these transactions exceeds the marginal revenue product of labor (measured by D°); it also means that employees are worse off than they would be in a well-functioning market—because they are engaging in some transactions in which the cost of supplying labor exceeds the wage they earn. Employee-side information failures are discussed further in Sections 2 (workplace safety) and 4 (workplace privacy) below.

1.2.1.2. Employer-side information failures Employers will frequently have only imperfect information about the attributes of their employees. This is particularly likely to be true at the point of hiring but may be true later on as well. A large literature in labor economics examines the consequences of such employer-side imperfect information for labor markets (e.g. Greenwald, 1986; Gibbons and Katz, 1991).

As noted above, the usual modeling assumption in the case of an employer-side information failure is that employers are aware of the limits on their information and respond rationally to these limits. Thus, the employers' information problem is typically modeled as a situation of adverse selection. Employer-side information failures and the resulting adverse selection problems are discussed further in Sections 5 (fringe benefits mandates) and 7 (wrongful discharge law) below.

1.2.2. Monopsony

A second failure in employment markets is monopsony power. (Because this chapter is focused on employment law rules that protect employees and does not examine labor law, the chapter does not discuss the alternative scenario of market power or monopoly on the employee side.) If an employer is a monopsonist in the market for a particular type of employee, then instead of taking the wage as given as in a competitive labor market, the individual employer will face an upward-sloping labor supply curve. This employer will choose its employment level E to maximize $R(E)-w(E)E$, where $R(E)$

is the employer's revenue and $w(E)$ is the labor supply curve. Under the associated first-order condition for the employer, it is clear that the wage under monopsony falls short of the marginal revenue product of labor $R'(E)$. This outcome is inefficient, as well as detrimental to employee welfare, because some employees who would produce more value than the cost of their labor are not hired. Section 9 below notes the familiar argument that minimum wage laws may respond to such monopsony-based market failure—a point that, while of theoretical interest, is generally believed to have limited practical importance.

1.2.3. Externalities

A third potential failure in employment markets arises from the external effects of some decisions by market participants. If, for instance, an employee is killed or injured on the job, it is not only the employee (and possibly employer) who may suffer harm. Family members will typically suffer, though employees may usually take such effects into account. Systems of social support will often be affected as well. Although externalities are obviously a classic form of market failure, they have received less attention than information failures and monopsony in the existing literature on the economics of employment law and, thus, receive relatively limited attention below.

1.2.4. Employee-side cognitive bias

Information failures, market power, and externalities are the traditional forms of market failure within conventional economic analysis. A separate set of potential market failures, however, arises from the possibility that employees will exhibit distorted labor supply decisions as a result of various forms of cognitive bias.

Of natural relevance to employment law is a substantial literature showing that many individuals exhibit optimism bias, adjudging their personal probability of facing bad outcomes to be lower than the average probability of facing such outcomes (Weinstein, 1980). Many people, for instance, believe that their chances of having an automobile accident are significantly lower than the average person's chances of experiencing this event (DeJoy, 1989), although of course these beliefs cannot all be correct, for if everyone were below "average," then the average would be lower. There is also evidence that people underestimate their absolute as well as relative (to other individuals) probability of experiencing negative events such as automobile accidents (Arnould and Grabowski, 1981, pp. 34–35; Camerer and Kunreuther, 1989, p. 566).

From a modeling standpoint, optimism bias among employees is similar to the employee-side information failures discussed in Section 1.2.1 above. Parallel to the case of such information failures, the most natural assumption, and the one that will be utilized in this chapter, is that employees with optimism bias either are not aware of the bias or do not change their behavior in response to the bias.[2] Thus, as with in-

[2] Akerlof and Dickens (1982), by contrast, analyze workplace safety under the assumption that employees are aware of their tendency to believe (in Akerlof and Dickens's model, because of cognitive dissonance) that

formation failures, if employees exhibit optimism bias in relation to some aspect of the employment relationship that affects their willingness to supply labor, then observed labor supply will differ from employees' true willingness to supply labor. In the case of workplace safety, for instance, optimism bias may lead even employees who have full information about the general risks and harms in their workplace to underestimate the probability that they personally will experience negative outcomes, and, as a result, they may oversupply labor. The effects of this were discussed in Section 1.2.1 above. Optimism bias is discussed further in Sections 2 (workplace safety) and 4 (workplace privacy) below.

2. Workplace safety mandates

This section and the sections that follow analyze the major forms of legal regulation of the employer-employee relationship, beginning with the legal regulation of workplace safety. A few of the topics included in the sections to follow are already treated in some depth in the *Handbook of Labor Economics* or the *Handbook of Public Economics* and, thus, are mentioned only briefly here, with cross-references to the longer treatments in the existing *Handbook* volumes.[3] Throughout, work published in law review, as opposed to economics journal, format is described at greater length, on the theory that the typical degree of background and detailed exposition provided in a law review article creates a form of barrier to entry that is not present when reference is made to work published in economics journal format.

Because many individuals spend a substantial fraction of their waking hours at work and often encounter risks in the course of work, workplace safety is a central issue of public policy. Many features of employment law are concerned with enhancing workplace safety. Legal regulation in this area includes both direct mandates of safe work conditions, as under the Occupational Safety and Health Act (OSHA) in the United States, and indirect channels for improving workplace safety through the employer incentives created by mandated compensation for workplace injuries. This section and Section 3 consider these two basic approaches in turn.[4]

The general starting point for economic analysis of workplace safety regulation is the observation that in the absence of market failure, less safe working conditions should

the workplace is safer than it is. Because of such awareness, employees in their model take action in response to their biased tendencies.

[3] In areas not already covered in other *Handbook* chapters, the treatment offered below seeks to describe and synthesize the most influential economic analyses in each major area of employment law, rather than to catalogue in a comprehensive manner all existing economic analyses in each area.

[4] In addition to workplace safety mandates and mandated compensation for workplace injuries, general tort law may have some effect on workplace safety incentives; however, such effects are limited by the fact that, at least in the United States, workers' compensation programs (discussed in Section 3) preempt most tort liability for workplace injuries. See Chapter 2, by Steven Shavell (2007), in this *Handbook* for further discussion of the economic analysis of tort law.

be fully compensated by higher wages—an application of the theory of equalizing wage differentials.[5] Viscusi (1978) and Viscusi and O'Connor (1984), among others, present evidence of adjustments in wages in response to workplace risks. The evidence of risk-based adjustments in wages, however, is limited in two important ways. First, without some independent way of monetizing the cost of a higher-risk job, at most the empirical evidence can tell us that wages move in a particular *direction* in relation to risk; it cannot tell us whether wages adjust *by the right amount* (given the employee's underlying preferences) in light of workplace risks. Movement in one direction or the other is ultimately a fairly weak test of the theory of equalizing wage differentials. The second limitation is that sources of credible identification of empirical effects are extremely difficult to find in this area because employees who select into different types of jobs with different levels of associated risk may differ along important dimensions that are not observed by the analyst; and moreover, the jobs into which employees select may also differ along dimensions that are not observed by the analyst. If either individual or job characteristics that are unobservable by the analyst are correlated with job risks, then correlations between risks and wages may be entirely spurious. Employment law, not satisfied that equalizing wage differentials fully address the issue of workplace safety, has chosen to regulate workplace safety in both the direct and indirect ways noted above.

2.1. Theoretical analysis of workplace safety mandates

The most direct form of workplace safety regulation is workplace safety mandates. These mandates require particular workplace practices intended to enhance workplace safety. In the United States, as previously noted, OSHA imposes a range of such mandates, which are described in detail in Smith (1976) and Mendeloff (1979).

Within economic analysis, workplace safety mandates are typically justified on the ground that either information failures or optimism bias on the part of employees leads them to oversupply labor at a given wage rate. Frequently employees will not be aware of the risks of a particular workplace, and even employees who have good information about the general risks of their workplace may overoptimistically assume that those risks do not apply to them personally.

As noted in Section 1.2 above, an oversupply of labor because of information failures, optimism bias, or both means that some inefficient transactions are taking place and that employees are worse off than they would be in a well-functioning market. Workplace injuries may also have important externality effects, but the informational and cognitive problems have been central in the existing literature. In the presence of such problems, workplace safety mandates can in theory improve both efficiency and employee welfare by eliminating the inefficient and detrimental (to employees) transactions described in Section 1.2. If observed labor supply under a mandate matches the "true" willingness to supply labor (because the workplace is safe, consistent with employees' belief), then

[5] Brown (1980) offers a general treatment of the topic of equalizing wage differentials.

only efficient transactions will occur, and employees will accept employment only to the extent that the wage equals or exceeds their "true" cost of supplying labor. Of course, in the real world the mandated levels of workplace safety may not match employees' beliefs or may be exorbitantly expensive; responding to even a genuine employee-side informational or cognitive problem does not guarantee that efficiency and employee welfare will rise under a workplace safety mandate. Ultimately, the effects of a workplace safety mandate are an empirical question.[6]

2.2. *Empirical analysis of workplace safety mandates*

As just noted, the central question about workplace safety mandates concerns their empirical effects on wages, employment levels, and, of course, workplace safety. A large empirical literature has attempted to identify such effects.[7] This literature has focused on workplace injuries (in the sense of immediate negative health effects), as distinguished from longer-term health risks from work; this focus is reflected in the discussion below. As the discussion below will make clear, the overall body of empirical evidence suggests modest (at best) effects of workplace safety mandates on observed levels of workplace safety, and, presumably because of the limited evidence of effects on the basic level of workplace safety, wage and employment effects of workplace safety mandates have generally not been studied. Most of the empirical evidence on the effects of workplace safety mandates comes from the United States, and that focus is reflected in the discussion below.

In sharp contrast to the workplace injury compensation systems noted in Section 3 below, workplace safety mandates in the United States operate at the federal level, and, as a result of OSHA's status as a federal law, state law variation generally cannot be used to identify OSHA's effects.[8] Limited exceptions to this statement about OSHA are Ruser and Smith (1988) and Morantz (2005). Ruser and Smith examine the effects of the initiation of a records-check procedure in some but not all states on the level of reported workplace injuries; they find a lower level of reported injuries with the records-check procedure in place, but they are unable to determine from their data whether this results

[6] Not addressed in this discussion is the possibility that workplace safety mandates may affect the degree of precautionary behavior by employees. Rea (1981) offers related discussion, although his analysis is primarily focused on compensation systems for workplace injuries (the topic of the next section) rather than on workplace safety mandates.

[7] For a comprehensive survey of this literature through the early 1990s, see Smith (1992).

[8] Much empirical work in employment law, including the work described in several of the sections below as well as much work in the employment discrimination area (e.g., Neumark and Stock, 1999; Jolls, 2004a; Jolls and Prescott, 2007), exploits variation in legal innovation across states to identify the effects of legal rules. With change over time but no cross-state variation—as with a federal law such as OSHA—it is often difficult to disentangle the effects of the law's enactment from other unobserved changes occurring at the same time. Indeed, as Smith (1992) notes, even a simple national-level before-after comparison of injury rates found in relation to OSHA's enactment is not possible because OSHA significantly changed the manner of collecting data on workplace safety.

from a true reduction in injury rates or (the interpretation they emphasize) an increase in the frequency of underreporting because under the records-check procedure underreporting would reduce the likelihood of costly safety inspections. Ruser and Smith's study is largely concerned with effects on reporting levels rather than the more basic issue of effects on workplace safety. Meanwhile, Morantz's (2005) work examines how injury rates among construction workers vary with federal versus state *enforcement* of OSHA's substantive provisions and finds significantly lower injury rates in states with federal enforcement. Although the injury rate data do not go back far enough to allow Morantz to examine changes over time with the move from federal to state enforcement (so that, in contrast to Ruser and Smith, identification is based on cross-state variation only, rather than variation across both states and time), Morantz's empirical approach does attempt to control comprehensively for other cross-state differences, apart from federal versus state enforcement, that could be influencing injury rates.

Other empirical work on OSHA has examined variation in its enforcement over time or across industries (rather than across states) as a source of identification of the law's effects on workplace safety. In light of the low financial penalties for the typical OSHA violation and the low likelihood of OSHA inspections (Viscusi, 1979; Weil, 1996), many have questioned whether OSHA is likely to have any effect on workplace safety at all, regardless of variation in enforcement within the low observed ranges of penalties and inspection frequencies. Viscusi (1979) and Bartel and Thomas (1985), for instance, find limited or no relationship between measures of OSHA enforcement and injury rates, while Viscusi (1986) finds at best a modest relationship.[9]

Scholz and Gray (1990) have suggested, however, that the limited effects of OSHA on injury rates found in many empirical studies reflect the overly generalized approach of these studies. Scholz and Gray focus on the firms most likely to be affected by OSHA—large, frequently inspected firms—and find significant effects of OSHA enforcement on injury rates in their sample. Unlike earlier studies, Scholz and Gray's study uses plant-level rather than industry-level data on OSHA enforcement levels and injury rates. However, a recent study by Gray and Mendeloff (2005) using a methodology similar to that used in Scholz and Gray (1990) found that the significant effects of OSHA on injury rates lasted only through the early 1990s and were not apparent in data for most of the 1990s.

While the central variable of policy interest in studying OSHA's effects is the workplace injury rate, several studies have examined the relationship between the level of OSHA enforcement and the degree of firms' observed compliance with OSHA's requirements. Bartel and Thomas (1985) and Weil (1996), for instance, find evidence of significant effects of enforcement on compliance with OSHA's requirements, notwithstanding the general evidence, noted above, of very limited levels of OSHA enforcement. By contrast, Weil (2001) finds only limited effects of enforcement activity on compliance in the construction industry.

[9] See Viscusi (1983) for discussion of additional studies of OSHA's effects on injury rates from the period just after OSHA's enactment.

Ultimately, the existing empirical work provides relatively little evidence of effects of OSHA on injury rates, with slightly more evidence of compliance effects. The lack of variation across states has, not surprisingly, impeded efforts to pin down the effects of OSHA's workplace safety mandates more definitively. Although the theory suggesting the likelihood of market failure in the workplace safety context seems strong, direct empirical evidence on the effects of OSHA has, at least thus far, not provided a clear basis for concluding that this law has enhanced efficiency and employee welfare or, even more basically, that the law has decreased workplace injuries. As described below, the empirical picture on compensation systems for workplace injuries—the other major employment law mechanism for improving workplace safety—is at least somewhat more positive.

3. Compensation systems for workplace injuries

Employment law seeks to enhance workplace safety not only through direct safety mandates, as under OSHA, but also through the deterrent effects afforded by legally mandated systems of compensation for those injured on the job. In the United States, state workers' compensation programs are a major source of such legally mandated compensation. Injured workers in the United States may also be eligible for compensation through the federal disability insurance program, but the absence of any experience-rating component in this program means that it does not create any particular deterrent effects for employers. The disability insurance program, which is funded through compulsory payroll deductions by employers, is nonetheless a significant additional source of compensation for employees injured at work and, thus, is usually included in analyses of compensation systems for workplace injuries.

Both the theory and the empirical evidence relating to workers' compensation systems are developed in Chapter 33 of the *Handbook of Public Economics* (Krueger and Meyer, 2002), and the federal disability insurance program is also extensively discussed in that chapter; thus the reader is referred to Krueger and Meyer's treatment for further discussion of the economic analysis of these programs. In brief, Krueger and Meyer conclude that the empirical evidence suggests longer periods out of work with more generous workers' compensation benefits (a finding with ambiguous welfare consequences, as noted by Krueger and Meyer); they also note the work by Gruber and Krueger (1991) suggesting no statistically significant reduction in aggregate employment levels from increases in the generosity of workers' compensation programs (a finding suggestive of the absence of significant welfare costs from these programs, although not of affirmative efficiency gains from the programs). Meanwhile, with respect to the federal disability insurance program, Krueger and Meyer emphasize the unresolved question of what causes the observed major changes in the size of the population receiving federal disability insurance payments, a topic of ongoing research (e.g., Autor and Duggan, 2003).

4. Workplace privacy mandates

A very rapidly growing area of employment law—particularly with the increasing presence of computers and the associated monitoring possibilities in the workplace—is workplace privacy. Issues of workplace privacy span a broad array of domains. Questions include the degree to which employers may videotape or audiotape their employees' activities in the workplace, the restrictions (if any) on drug and alcohol testing of employees, the degree to which employers may engage in various forms of monitoring of employees' activities on computers, and the limits on employers' disclosure of medical and other personal information about employees. Workplace privacy mandates require employers to conduct activities of this sort in specified ways, if at all. This section briefly highlights the central issues in the economic analysis of such workplace privacy mandates.

4.1. Theoretical analysis of workplace privacy mandates

From an economic standpoint, the possible market failures in the workplace privacy setting parallel the main potential market failures in the workplace safety context. One potential market failure is an employee-side information failure; for some forms of privacy-related employer behavior, the employee may simply have no idea that the behavior is occurring. Employees may assume that their privacy is protected at work, just as they may assume that their workplace is safe. Examples in the privacy context include video or audio monitoring and the monitoring of employees' computers—practices of which employees may be entirely unaware. In other settings, by contrast, an employee may know when a particular invasion of privacy occurs; an example is urine testing for drug use—an employer practice that obviously cannot be implemented without the employee's knowledge.

Even employees who have accurate information about an employer's general practices in relation to workplace privacy may be led by optimism bias to underestimate the likelihood that the at-issue employer behavior will be undertaken in relation to, or have a negative effect upon, them as opposed to other employees. The analysis is again parallel to the workplace safety context, in which employees may underestimate their personal likelihood of workplace risks. Note that in the workplace privacy context, the risks of market failure related to optimism bias will tend to be greater when there is more uncertainty in the employer's policy. For instance, employees may underestimate the likelihood that they will be subjected to a drug test under an employer policy permitting drug testing of employees involved in a workplace accident, simply because employees may tend to underestimate the likelihood that they will be involved in such an accident. By contrast, if an employer policy provides that employees will be subjected to ongoing video or audio monitoring as a matter of course, then optimism bias is likely to be less important—though optimistically biased employees could still underestimate the likelihood that the monitoring will detect a prohibited behavior on their part.

Parallel to the discussion of workplace safety in Section 2.1 above, a workplace privacy mandate may address an employee-side information failure or cognitive bias by eliminating the gap between observed labor supply and the "true" willingness to supply labor (because privacy is now protected at the workplace, consistent with employees' expectation). Of course, as above, whether any given workplace privacy mandate actually has positive effects on efficiency and employee welfare is ultimately an empirical question.

4.2. Empirical analysis of workplace privacy mandates

Very little empirical evidence currently exists on the effects of workplace privacy mandates. However, in the case of the United States, the existence of variation in these mandates across states and time suggests the potential value of empirical inquiry in this area. Whether wages and employment levels would ever move to a discernible degree in response to workplace privacy mandates is unclear, but effects on observable outcomes other than wages and employment levels may be easier to detect. Jacobson (2003), for instance, presents evidence of better safety outcomes in safety-sensitive occupations in states that enacted legislation clarifying the permissibility of drug testing for safety-sensitive positions than in states that did not enact such legislation. There is some suggestion of preexisting trends in the enacting states, so the results should be viewed with some caution, but Jacobson also finds improved safety outcomes after the imposition of federally mandated drug testing for certain safety-sensitive positions. Jacobson's findings suggest that some forms of workplace privacy mandates (especially in the drug testing context) may have the unfortunate effect of worsening safety outcomes. Additional empirical work on the effects of workplace privacy mandates is likely to appear as these mandates become increasingly common.

5. Fringe benefits mandates

Employers may be responsible not only for improved workplace conditions—including safety and privacy—but also for the provision of important fringe benefits. Fringe benefits are usually understood to include such benefits as health insurance and pensions. The present section focuses on the legal regulation of health insurance, the great majority of which is provided through the employment relationship (although this is less true outside the United States). Pension-related mandates are not considered in this chapter because there is very little existing literature on the effects of such mandates on wages, employment levels, or other labor market outcomes.[10]

[10] The central law governing pensions in the United States is the Employee Retirement Income Security Act, a federal law that broadly displaces potential state-level regulation of pensions and, thus, leaves little opportunity for credible identification (through state-level variation in legal innovation) of the effects of legal regulation of pensions. See Ippolito (1988) for further discussion of the effects of the pension-related mandates imposed at the federal level.

With respect to health insurance mandates, the Medicare program in the United States requires compulsory payroll deductions by employers and offers government-provided health insurance to (primarily) retirement-aged individuals. Under this program, individuals see a deduction from their paychecks during their working years and then are entitled to health insurance financed by the federal government during their retirement.[11] The Medicare system is discussed in Chapter 50 of the *Handbook of Labor Economics* (Currie and Madrian, 1999) and Chapter 31 of the *Handbook of Public Economics* (Cutler, 2002), and the reader is referred to those volumes for further discussion of this system. Because the Medicare system is covered in those volumes, the focus of the present section is other types of health insurance mandates.

Both federal and state law in the United States play a role in structuring the system of health insurance mandates considered in this section. The federal role here is two-fold. First, the Employee Retirement Income Security Act (ERISA) broadly "preempts," or renders inapplicable, all state-level health insurance mandates insofar as employer-provided health insurance is concerned, unless the employer-provided insurance is procured from an insurance company. Thus, any employer that self-insures—as many now do—is not subject to any health insurance mandates imposed at the state level. Second, federal law imposes a few limited mandates on health insurance plans. One such mandate is contained in the Consolidated Omnibus Budget Reconciliation Act (COBRA), which requires that, above a certain employer-size threshold, employees be allowed to continue purchasing health insurance through their former employer for up to 18 months after leaving that employer. A second source of federal health insurance mandates is the Health Insurance Portability and Accountability Act (HIPAA), which imposes a variety of requirements on employers and other health insurance providers, including mandated coverage for preexisting conditions.[12]

State-level health insurance mandates, to the extent that their application is not preempted by ERISA, also impose a variety of requirements on health insurance plans; for instance, some state laws impose continuation coverage mandates similar to the mandate in COBRA (see Gruber and Madrian, 1994). State-level regulation is discussed briefly below where its mandates overlap with federal requirements, but, because of the preemption issue noted above, the focus of this section is on federal health insurance mandates.[13]

[11] By contrast, the Medicaid program in the United States is targeted to the needy and is not linked to the employment relationship in any way.

[12] Moreover, recently enacted HIPAA regulations impose a series of privacy-related requirements on employers concerning the disclosure of medical information; those requirements may be analyzed under the framework set forth in Section 4 (workplace privacy) above. An additional federal law regulating health insurance is the Mental Health Parity Act. For the reasons given in Section 6 below, this law is best viewed as a targeted mandate and, thus, is discussed in Section 6 rather than here.

[13] Economic analyses of the effects of various state-level health insurance mandates include Gruber (1994b), Kaestner and Simon (2002), and Simon (2005). Chapter 31 in the *Handbook of Public Economics* (Cutler, 2002) also contains discussion of these mandates.

5.1. Theoretical analysis of fringe benefits mandates

Several potential market failures, including employee-side information failures and cognitive biases, may be relevant to analysis of the health insurance mandates noted just above, but perhaps the most obvious labor market failure in this context involves employer-side imperfect information. (For discussion of the broader issue of provision of health insurance through, rather than independently of, the employment relationship, see Currie and Madrian (1999).) For instance, an individual employer opting unilaterally to offer continuation coverage of the sort required by COBRA might disproportionately attract employees who are less likely to remain employed (especially those who simultaneously have concerns about their future health) and, thus, who are particularly focused on the issue of continuation coverage. In broad terms, the problem of adverse selection here is similar to the problem—described in detail in Section 7.1 below—faced by an individual employer opting to offer protection against discharge from employment without just cause. In both cases, offering the benefit may make the employer disproportionately attractive to less desirable employees.[14]

In the health insurance context, if the value to employees of a particular form of health insurance coverage (such as continuation coverage) exceeds the cost of providing such coverage, then mandating such coverage may efficiently respond to the adverse selection problem just described. Of course, in the case of continuation coverage, employers in an ideal world might offer not this type of coverage but, instead, employee-specific packages of wages and health insurance tailored to each employee's individual costs and needs (Gruber and Madrian, 1994). However, as Gruber and Madrian note, the barriers to such an approach—whether achieved voluntarily or through legal regulation—in the real world are clear, and thus mandated continuation coverage, as under COBRA, may be a good second-best solution. Empirical evidence on the effects of both mandated continuation coverage and mandated coverage for preexisting conditions is noted in the next subsection.

5.2. Empirical analysis of fringe benefits mandates

Empirical evidence on the effects of continuation coverage mandates on a variety of employment-related outcomes is discussed in Chapter 50 of the *Handbook of Labor Economics* (Currie and Madrian, 1999); the discussion there of the studies by Gruber and Madrian (1994, 1995, 1996, 1997) of both COBRA and state-level continuation coverage mandates is especially relevant. On balance, continuation coverage mandates appear to increase both separation and retirement from work by employees; effects on overall wage and employment levels (which would permit conclusions about the welfare effects of continuation coverage mandates) presumably were not large enough to be studied empirically.

[14] Aghion and Hermalin (1990) offer general discussion of the problem of adverse selection in employer-employee and other contracting releationships.

With respect to preexisting conditions coverage mandates, Chapter 31 of the *Handbook of Public Economics* (Cutler, 2002) discusses the effects of various state-level health insurance mandates, including preexisting conditions coverage mandates, on health insurance coverage rates. In addition, in the time since this volume of the *Handbook of Public Economics* was published, studies by Kaestner and Simon (2002) and Simon (2005) have also examined the effects of state-level health insurance mandates, including preexisting conditions coverage mandates, while recent work by Sanz-de-Galdeano (2006) studies the effects of HIPAA, which (as noted above) includes a preexisting conditions coverage mandate. The body of existing empirical work suggests that these mandates have limited or no effects on insurance coverage rates, wage and employment levels, or other outcomes.[15]

6. Targeted mandates

While the employment law mandates discussed in Sections 2 through 5 above are mandates predominantly directed to employees as a whole rather than to any discrete subgroup of employees, other employment law mandates are targeted to particular demographic subgroups. Indeed, many employment law mandates—including certain aspects of employment discrimination law as well as employment law rules such as mandated workplace leave (emphasized below)—are targeted in this way. The direct effects of these targeted mandates will be felt primarily or exclusively by demographic subgroups of employees, such as individuals with disabilities, women, or members of particular racial groups, rather than by employees as a whole (Gruber, 1994a).

Of course, no benefits mandate may be entirely untargeted; many mandates, including at least some of those discussed above, may tend to benefit some demographic groups within a given employment market more than others within this market. Thus, the difference between general and targeted mandates is probably best viewed as a difference in degree rather than a difference in kind.

The importance of distinguishing between general and targeted mandates arises from the fact that the subgroups to which targeted mandates are directed are generally *groups whose wages and employment opportunities are legally required to match those of the nontargeted employees*. In the absence of such requirements (or if such requirements were not binding),[16] the targeted subgroup could simply be treated as a separate labor

[15] Note that although the primary focus of the empirical literature on health insurance mandates is on outcomes other than overall wage and employment levels, insofar as overall wage and employment levels are concerned, the framework described in Section 1.1 above may require some adjustment when analyzing the effects of health insurance mandates. This is so because the cost of health insurance is likely to vary with the number of employees rather than, as depicted in Figure 1 above, with the number of total worker-hours; in other words, the distribution of worker-hours over employees may matter greatly in the health insurance context. Cutler and Madrian (1998) offer further discussion of this point.

[16] More precisely, if restrictions on wage differentials between the two groups were not binding. See Jolls (2000) for further discussion.

market, and the same basic sort of analysis as was discussed in Sections 2 through 5 above would continue to apply. However, in the presence of binding restrictions on differential treatment across groups, the analysis of targeted mandates proves to be quite different from the analysis employed until this point in the chapter. As described more fully below, employment law mandates targeted to demographic subgroups of employees constitute an important exception to the general focus in this chapter on market failures as a necessary condition for employment law regulation to enhance employee welfare. If the goal of a targeted mandate is to enhance the welfare of targeted employees—as it often is—then the desired result may obtain wholly apart from any sort of market failure—a point obscured by the common focus in the literature on the standard mandated benefits framework even when analyzing targeted mandates.[17] In light of this observation, and for the sake of analytic clarity, the present section will analyze targeted mandates on the assumption of no market failure of any sort. The analysis below could readily be adjusted to incorporate market failure in combination with a targeted rather than general mandate. As above, empirical analysis follows the theoretical analysis.

6.1. Theoretical analysis of targeted mandates

With a targeted mandate and no market failure of any sort, labor supply with the mandate in place will shift exclusively or disproportionately for the targeted employees, rather than (as in Section 1.1 above) for employees as a whole. Meanwhile, on the employer side, the total marginal revenue product of labor (reflecting, among other things, the cost of the mandated benefit) will shift exclusively or disproportionately for the targeted group. If, for instance, employers are legally required to provide leave from work following the birth of a child, as they are in many countries, then both the willingness to supply labor and the total marginal revenue product of labor are likely to shift disproportionately for female employees.[18]

Because of the differential effects of a targeted mandate across groups of employees, it is important for purposes of the analysis of such mandates to separate out two distinct labor markets: the market for employees targeted by the mandate and the market for the remaining employees. Each market will have its own labor supply and demand functions (although, as described below, the demand functions will end up being the same if restrictions on wage and employment differentials are binding). And, because the demand for employees of one type will depend, among other things, on the demand

[17] See Jolls (2000) for further discussion of this point.

[18] For expositional ease, the analysis below will assume that the mandated benefit has value only to the targeted group and imposes costs only in connection with that group. However, the conclusions offered below would remain qualitatively unchanged if the mandate had some value beyond the targeted group (although less than the value to the targeted group); the same is true if the mandate imposed some cost beyond the targeted group (although less than the cost for the targeted group).

for employees of the other type, it will no longer be possible to represent everything of interest on a single, two-dimensional labor supply and demand diagram, as in Figure 1.

The following notation will be used below:

E_t = employment level of targeted employees;

W_t = wage level of targeted employees;

E_n = employment level of nontargeted employees;

W_n = wage level of nontargeted employees;

$C(> 0)$ = per-worker-hour cost of providing mandated benefit to targeted employees;

V = per-worker-hour value of mandated benefit to targeted employees;

M = per-worker-hour marginal revenue product of labor from production.

Prior to the imposition of a targeted mandate, the wages earned by targeted and nontargeted employees will be given by $W_t = W_n = M(E_t + E_n)$. Meanwhile, labor supply for the two groups of employees prior to the imposition of a targeted mandate may be written as follows:

$$E_t = S_t(W_t)$$
$$E_n = S_n(W_n)$$

Let $(W_t^\circ, W_n^\circ, E_t^\circ, E_n^\circ)$ with $E_t^\circ > 0$ and $E_n^\circ > 0$ denote an interior solution to this system.

After the imposition of a targeted mandate, labor supply for the two groups of employees will be given by the following equations:

$$E_t = S_t(W_t + V); \quad (1)$$
$$E_n = S_n(W_n). \quad (2)$$

However, labor demand after the imposition of the mandate will depend on the degree to which restrictions on wage and employment differentials are binding. One possibility is that restrictions on wage differentials are not binding (and restrictions on employment differentials are or are not binding); in this case, as noted above, a targeted mandate may be analyzed within the basic type of framework used in Sections 2 through 5 above.[19] A second possibility is that restrictions on wage differentials are binding while restrictions on employment differentials are not; in this case targeted employees will effectively be more expensive to employ (as they must earn the same wage but cost the employer more), and employers will thus tend to reduce their hiring of targeted employees. The most interesting possibility, and the one that will be examined in the remainder of this subsection, occurs when restrictions on both wage and employment differentials

[19] See Jolls (2000) for further discussion.

are binding. In this case there must not be any difference between the wages or employment opportunities of targeted and nontargeted employees within a given employment market; employers must pay each type of employee the same wage and must demand each type in proportion to its willingness to supply labor at that wage.

With binding restrictions on wage and employment differentials, the common wage W for the two groups of employees after the imposition of the mandate will be given by:

$$W = [E_t/(E_t + E_n)][M(E_t + E_n) - C] + [E_n/(E_t + E_n)]M(E_t + E_n).$$

Rewriting:

$$W = M(E_t + E_n) - [E_t/(E_t + E_n)]C. \tag{3}$$

The equation in (3), together with the labor supply equations in (1) and (2), yields a system of three equations in three unknowns; let (W^*, E_t^*, E_n^*) denote a solution to this system.

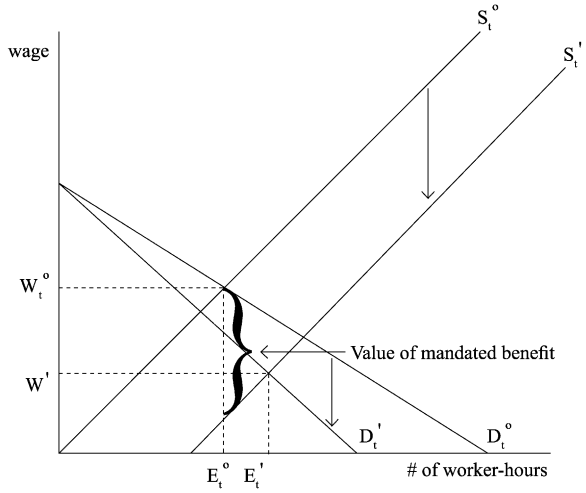
As (3) shows, a targeted mandate will affect labor demand both as a result of the mandate's effect on the marginal revenue product of labor from production through changes in $E_t + E_n$ and as a result of the direct cost C of the mandated benefit. With respect to the former effect, it is straightforward to show that because of the downward pressure on labor demand as a result of the cost of the mandated benefit, the employment level of nontargeted employees will fall with the mandate: $E_n^* < E_n^\circ$.[20] Intuitively, the mandate, by requiring employers to incur costs for targeted employees that nontargeted employees will have to share as a consequence of the binding restrictions on wage and employment differentials, induces marginal nontargeted employees to exit the market. This relationship between E_n^* and E_n° in turn allows one to depict the effects of a targeted mandate on the wages and employment of targeted employees in simple graphical form in a figure (Figure 2) similar to Figure 1 above.

In Figure 2, the curve D_t° is the pre-mandate labor demand curve for targeted employees at the pre-mandate equilibrium level of nontargeted employment (E_n°). Thus, the curve D_t° is given by $W = M(E_t + E_n^\circ)$. Meanwhile, the curve D_t' is given by the following equation:

$$W = M(E_t + E_n^\circ) - [E_t/(E_t + E_n^*)]C.$$

The curve D_t' thus depicts the effect of the shift due to the cost of the mandated benefit after the mandate is imposed; the M term is ignored by assuming that the marginal revenue product of labor from production as a function of targeted employment (E_t) is the same pre- and post-mandate $(M(E_t + E_n^\circ)$ in both cases). Because, as noted above, $E_n^* < E_n^\circ$, it is straightforward to see that the ultimate post-mandate labor demand curve for targeted employees, with $E_n = E_n^*$, must lie above the curve D_t' in Figure 2.

[20] See Jolls (2000) for details.



Notes: Employees are assumed to be demanded and paid in accordance with their marginal revenue product of labor (rather than, for example, earning efficiency wages), and the mandated benefit is assumed to be a variable-cost rather than a fixed-cost one.

Figure 2. Effects of a targeted mandate on wages and employment levels of targeted employees.

We are now in a position to assess the effects of a targeted mandate on the wages and employment levels of targeted employees. As the discussion in Section 1.1 above suggests, it is useful to separate the analysis into distinct cases based on the relationship between the value and the cost of the mandated benefit. The straightforward case is the one in which the value of the mandated benefit equals or exceeds its cost.[21] In this case, not surprisingly, the targeted employees' wage will fall by less than the value of the mandated benefit, while their employment level will rise, as depicted in Figure 2.

The more interesting case is the one in which the value of the mandated benefit is less than its cost. In this case it is no longer certain that targeted employees will be better off after the mandate is imposed, but it remains likely; as long as the fraction of nontargeted individuals in the qualified population is not too small, and the gap between the value and the cost of the mandated benefit is not too large, a targeted mandate will always

[21] In the absence of any market failure—the assumption maintained throughout this section—a benefit whose value equals or exceeds its cost might, but conceivably would not, be provided in the absence of a mandate. Because the benefit accrues exclusively or disproportionately to targeted employees, it seems possible that the benefit would not be provided (with binding restrictions on wage and employment differentials) even if the value of the benefit exceeded its cost. This issue is not analyzed rigorously here, however.

make targeted employees better off. If, for example, nontargeted employees constitute the vast majority of the employment market in question, then the fall in the total marginal revenue product of labor for all employees in that market with the imposition of a targeted mandate will be small, as the average cost of the mandated benefit across the employment market will be small. And the smaller the downward shift in the total marginal revenue product of labor as a result of the mandated benefit cost, the smaller the gap between the curves D_t° and D_t' in Figure 2, and hence the lower the likelihood that this gap will exceed the downward shift in the labor supply curve (S_t° to S_t'). As long as the downward shift in the labor supply curve equals or exceeds the gap between D_t° and D_t' over the relevant range of employment levels, the wage of targeted employees will fall by less than the value of the mandated benefit, while their employment level will rise.

Thus, targeted mandates may be justifiable on distributive grounds even when the cost of the mandated benefit exceeds its value.[22] In the case of general mandates, by contrast, distributive considerations cannot justify legal intervention when the cost of the mandated benefit exceeds its value because, as noted in Section 1.1 above, the employees' wage will fall by more than the value of the benefit to them. This effect occurs because there is no other group to whom to shift costs. But with targeted mandates, even if the value of the mandated benefit is less than its cost, the mandate may make targeted employees better off because nontargeted employees will bear some of the associated cost.

The point about potential distributive gains wholly apart from market failure is especially important because the fact that the value of a mandated benefit is less than its cost may reflect precisely the undesirable distributive situation that employment law seeks to remedy. The reason is that "value" in this framework is measured by employees' willingness to pay for the benefit by accepting lower wages, and the distributive situation of targeted employees might preclude them from accepting lower wages (see generally Dworkin, 1980). If, for instance, a mandate requires employers to provide medical leave to employees with serious medical problems, the value (measured by willingness to pay) of this benefit to the targeted employees may be limited not by the utility of the leave to these employees but by their financial position. Mandated medical leave is discussed further in the next subsection.

6.2. *Empirical analysis of targeted mandates*

As suggested above, some targeted mandates arise under employment discrimination law and, for that reason, are not analyzed further in this chapter.[23] The primary employ-

[22] Of course, distributive goals might alternatively be achieved through a tax-and-transfer regime (Kaplow and Shavell, 1994).

[23] Disability discrimination law, for instance, requires employers to make "reasonable accommodations" for disabling conditions—a mandate targeted to employees with disabilities (Acemoglu and Angrist, 2001). Similarly, to the extent that sex discrimination law requires health insurance coverage of maternity-related medical costs, it imposes a mandate targeted to female employees (Gruber, 1994a).

ment law application of the analysis of targeted mandates is workplace leave mandates, which at least in the United States arise under employment law outside of what is generally considered employment discrimination law.[24] Workplace leave mandates in the United States disproportionately target both disabled employees (through mandated entitlement to medical leave) and female employees (through mandated entitlement to leave following the birth of a child).[25] These two aspects of mandated workplace leave are considered in turn below.[26]

As noted above, the wage and employment effects of a targeted mandate depend significantly on the degree to which restrictions on wage and employment differentials between targeted and nontargeted employees are binding. Starting with restrictions on wage differentials, such restrictions are likely to be binding in the absence of significant occupational segregation between targeted and nontargeted groups. Legal restrictions on wage differentials are generally fairly easy to enforce (e.g., Posner, 1987), and, moreover, employers may have incentives to adhere to norms of pay equity wholly apart from legal restrictions because of the potential morale problems that can result from inequity in wages between different groups performing the same work.

In the case of mandated entitlement to medical leave, the group of employees disproportionately (though of course not exclusively) targeted by such a mandate is employees with disabilities, some of whom will be significantly more likely to require leave than nondisabled employees. With respect to occupational segregation, at least in the United States employees with disabilities are not significantly segregated, so restrictions on wage differentials between employees with and without disabilities are likely to bind.

Likewise, with respect to restrictions on employment differentials, while these restrictions are unlikely to bind directly for employees with and without disabilities because of the difficulty of enforcing such restrictions (e.g., Posner, 1987), the ultimate effect may be the same in analyzing mandated entitlement to medical leave because many of the conditions for which such leave will be required are unobservable to employers at the time of hiring and, thus, cannot be the basis of differential employment decisions (Jolls, 2007). In this case, the analysis in Section 6.1 above predicts that mandated entitlement to medical leave will increase both wages and employment levels of employees with

[24] The Family and Medical Leave Act, which requires employers to provide employees with leave from work under specified conditions, is not part of any of the employment discrimination statutes in the United States and is administered by the Department of Labor rather than the agency (the Equal Employment Opportunity Commission) that administers employment discrimination statutes.

[25] Mandated leave following the birth of a child is targeted to female employees even though leave is typically available to male employees because female employees who have biological children will require at least a brief period of time off from work after a birth to recover from the temporary disability associated with giving birth and, thus, will almost certainly (for purely biological reasons) be more likely than male employees to take leave.

[26] Another targeted mandate in the United States is the Mental Health Parity Act, which requires that coverage of mental and physical conditions under health insurance plans be comparable in certain respects. Because this law is generally believed to be of modest practical impact, it seems unlikely to be the source of observable empirical effects.

disabilities, unless either the fraction of targeted individuals in the qualified population is too large or the cost of the leave significantly exceeds its value. In the United States, because medical leave under the Family and Medical Leave Act (FMLA) is unpaid, it seems unlikely that the cost of this benefit significantly exceeds its value.

A straightforward empirical measure of the wage and employment effects of mandated entitlement to medical leave under the FMLA on employees with disabilities is a comparison of wages and employment levels of such individuals before and after the FMLA went into effect—ideally with an additional comparison between states in which mandated entitlement to medical leave under the FMLA was an innovation and those in which it was not. Both approaches suggest neutral to positive effects of mandated entitlement to medical leave on disabled employment levels (relative to nondisabled employment levels), with the most credible evidence suggesting at least some positive effects (Jolls, 2007). This evidence underlines the prospects for positive effects of targeted mandates on the employment opportunities of targeted employees, consistent with the theoretical analysis offered in Section 6.1 above.

As just discussed, relatively limited degrees of occupational segregation (as in the case of employees with and without disabilities in the United States) suggest that restrictions on wage differentials will be binding. But in the case of male and female employees—the groups likely to be disparately affected by mandated entitlement to leave following the birth of a child—employment markets remain quite segregated.[27] With substantial occupational segregation, restrictions on wage differentials will tend to be of little force because the only comparisons that are drawn in the law are those between employees within the same employment market (or, more technically, those performing the same or similar work). As noted above, in the absence of binding restrictions on wage differentials, an analysis similar to that used in Sections 2 through 5 above remains applicable to the case of a targeted mandate.[28]

Mandated entitlement to leave from work following the birth of a child exists in a wide array of countries and has been extensively studied. Waldfogel (1999), for instance, examines female employees' wages and employment levels in the aftermath of the FMLA's enactment in the United States and finds no consistent pattern of statistically significant results. The lack of clear results may stem in part from the fact that, as Waldfogel notes, many female employees were entitled (by state legislation or firm policy) to FMLA-type benefits following the birth of a child even prior to the FMLA's enactment.

Another test of the effects of mandated entitlement to leave following the birth of a child comes from looking at the effects of European leave laws. Ruhm (1998), for

[27] See Chapter 25 in the *Handbook of Labor Economics* (Blau and Kahn, 1999) for further discussion of occupational segregation by sex.

[28] Of course, very few employment markets are actually perfectly segregated, so that there is absolutely no opportunity to compare targeted and nontargeted employees' wage levels. Jolls (2000) provides further discussion of the reasons that markets with significant segregation are nonetheless most naturally modeled as markets in which restrictions on wage differentials are not binding.

instance, finds that the length of the leave period under these laws generally is negatively related to the wages of female employees and positively related to their employment levels. Note that because employees' leave is paid, in contrast to the situation under the FMLA, positive employment effects are more likely under the European laws, which are presumably valued more by employees than the FMLA because of the paid nature of the leave. (Negative labor demand effects of paid leave are blunted by government rather than employer financing of the leave.) An exception to the general pattern of Ruhm's findings is that the effect of having some mandated leave versus none—as distinguished from the effect of having a short mandated leave versus a long mandated leave—has no statistically significant effect (rather than a statistically significant negative effect) on wages of female employees. This may result from the fact that a short leave, as opposed to a long leave, imposes relatively few costs on employers, and thus provides little occasion for a wage adjustment.[29]

Overall, the empirical evidence on mandated entitlement to leave following the birth of a child suggests neutral to positive employment effects—the same broad pattern as with mandated entitlements to medical leave. Wage effects, by contrast, appear more negative with mandated leave after the birth of a child—a likely consequence of the degree of occupational segregation by sex. Note, however, that because sex (in contrast to many medical conditions) is easily observable to employers, decreasing occupational segregation by sex over time is more likely to generate job losses from mandated entitlement to leave following the birth of a child (as wage adjustments are precluded by decreased segregation) than to produce the benefits for targeted employees that may come from mandated entitlement to medical leave.

7. Wrongful discharge laws

One of the most fundamental issues in employment law involves the general conditions under which employers may engage in "the industrial equivalent of capital punishment" by discharging an employee.[30] In many countries, a pervasive set of laws limits the conditions under which employees may be discharged without attendant legal obligations (e.g., Blau and Kahn, 1999). In the United States, by contrast, a sharply different approach prevails; an employee may generally be fired at any time, without identification of any reason whatsoever for the discharge and without any attendant employer obligations, unless either the employee's contract specifically provides otherwise or the discharge reflects unlawful discrimination or some other unusually arbitrary or abusive

[29] An important subtlety in discerning the effects of mandated entitlement to leave following the birth of a child is that, as Waldfogel (1998, 1999) emphasizes, such mandates may have a sort of composition effect, moving female employees into, or keeping them in, better, higher-level jobs. If this is so, then the aggregate effects of the mandates may be more mixed and more complex than the effects suggested by the theoretical framework described above.

[30] The quoted phrase comes from *Complete Auto Transit, Inc. v. Reis, 451 U.S. 401 (1981)*.

basis for decision (Kim, 1997). Only one of the fifty states in the United States (Montana) departs from this regime of "at will" employment. The United States does have an unemployment insurance system, which is a limited form of employment protection and is discussed in the next section, but (outside of Montana) there is no general requirement that employers offer any sort of justification when discharging an employee and no general set of employer financial obligations attending such discharges.[31]

The issue of the scope of wrongful discharge laws—defined for purposes of this chapter as laws requiring a threshold level of justification before employers may discharge employees—does not fit well within the framework described in Section 1.1 above because, in contrast with the legal rules discussed in the preceding sections, the legal regulation here is directly linked to the employment level. (In the case of the legal rules governing workplace safety, workplace privacy, and various types of employee benefits such as health insurance and workplace leave, the rules affect the wage and employment levels only indirectly through the shifts in labor supply and marginal revenue product of labor they will produce.) The analysis below first provides theoretical discussion of the labor market effects and the efficiency of wrongful discharge laws; it then notes evidence from empirical studies of these laws.

7.1. Theoretical analysis of wrongful discharge laws

7.1.1. Labor market effects of wrongful discharge laws

In general, wrongful discharge laws have competing potential effects on the employment level. On the one hand, they potentially increase employment because they impose barriers to firing. On the other hand, these very barriers may discourage the hiring of employees in the first place. Models of these types of competing effects are discussed in Chapter 25 in the *Handbook of Labor Economics* (Blau and Kahn, 1999), and the reader is referred to that volume for further discussion of the theoretical analysis of the labor market effects of various forms of employment protection law.

7.1.2. Efficiency analysis of wrongful discharge laws

At least in theory, several market failures may justify legal rules requiring a threshold level of justification before employees are discharged. Two leading arguments are noted below.[32]

[31] For further discussion of current law governing discharge from employment in the United States, see Kim (1997).

[32] Levine (1991) notes additional potential market failures in this context, including cognitive bias and externalities.

7.1.2.1. Employee-side information failures The most straightforward justification for wrongful discharge laws is that in the absence of such laws, inadequate employee-side information about the ability of an employer to discharge employees without offering an adequate justification for the discharge leads employees to oversupply labor at a given wage rate. Empirical evidence on this form of market failure is discussed in Section 7.2 below.

If, in the absence of wrongful discharge laws, many employees incorrectly believe that they can only be discharged after a threshold level of justification has been met, then wrongful discharge laws can improve both employee welfare and efficiency by eliminating the gap between employees' observed and "true" willingness to supply labor. With a wrongful discharge law in place, the governing regime is consistent with employees' belief—although of course in the real world the mandated level of justification for discharge may not match employees' beliefs, or its imposition may impose very large costs. As described below, the available empirical evidence on employee-side information failure in the wrongful discharge context—while strongly supportive of the existence of such information failure—does not allow any inference about whether existing wrongful discharge laws are ultimately a welfare-enhancing response to employee-side information failure in this context.

7.1.2.2. Employer-side information failures Alongside employee-side information failure, Levine (1991) develops the argument that employer-side information failures may also justify wrongful discharge laws. Intuitively, an individual employer may not offer wrongful discharge protection on its own, even if such protection, adopted universally, would increase both efficiency and employee welfare, because the individual employer who offers the protection may become a magnet for less desirable employees.

In Levine's model, employees are of two different types, and discharge occurs either on an at will basis—where employers need not provide any reason for discharge—or on a "just cause" basis—where employers must prove employee malfeasance in order to justify a discharge. Note that, by contrast to the focus in Sections 2 through 6 above on models in which employees are paid in accordance with the marginal revenue product of labor, Levine's model is a species of efficiency-wage model, in which employees may earn rents in order to discourage shirking. Levine shows that a single firm may not profit from adopting a just cause approach when its competitors do not adopt it, even if such an approach, adopted universally, would be efficiency-increasing. Because of the phenomenon of adverse selection, a mandated just cause requirement may enhance both efficiency and employee welfare.

7.2. Empirical analysis of wrongful discharge laws

7.2.1. Labor market effects of wrongful discharge laws

The evidence on the labor market effects of various forms of employment protection law is discussed in Chapter 25 in the *Handbook of Labor Economics* (Blau and Kahn,

1999). In addition, in the time since this volume of the *Handbook of Labor Economics* was published, studies by Miles (2000), Autor (2003), and Autor, Donohue, and Schwab (2006) have further examined the wage and employment effects of wrongful discharge laws in the United States. Both Miles (2000) and Autor (2003) find a significant positive relationship between more restrictive wrongful discharge laws and the substitution of temporary for regular employees. Meanwhile, Autor, Donohue, and Schwab (2006) find negative employment effects of some, though not all, types of wrongful discharge laws.

7.2.2. Efficiency analysis of wrongful discharge laws

The available empirical evidence on the efficiency of wrongful discharge laws bears most strongly on the first potential market failure discussed above—employee-side information failure. Kim (1997, 1999) presents evidence that most individuals in the United States are not aware of the fact that employers need not meet some threshold level of justification in order to discharge their employees. Kim surveyed a total of 921 recently discharged employees at unemployment benefits offices in California, Missouri, and New York in the late 1990s. To file a claim for unemployment benefits, claimants in these states must appear in person at a benefits office. In Kim's study, claimants were approached while waiting for assistance at benefits offices and asked if they would complete a written survey. Presumably because the claimants were already waiting and did not have the option to be elsewhere, response rates were high—85% in California and Missouri and 69% in New York.

The surveys in Kim's study asked respondents to consider specific scenarios in which individuals were discharged without a good reason and then to indicate whether they believed such discharges would be found to be lawful by a court of law. For instance, one question asked about the lawfulness of a termination based upon personal dislike of the employee; another asked whether it was lawful to terminate an employee based on a mistaken belief that the employee had stolen money (a belief that the employee could prove was incorrect). While such discharges are unquestionably permitted under the law in the United States (outside of Montana), the vast majority of respondents believed that the discharges were unlawful. Table 1 below summarizes Kim's main findings.

One issue with Kim's findings is whether unemployment insurance claimants are representative of the overall employee population in the United States. On the one hand, as Kim notes, individuals who recently lost their jobs may be more likely than the average employee to have some familiarity with the legal regime governing discharge, precisely because of their recent experience with being discharged. In addition, because eligibility for unemployment benefits requires both that the claimant have some prior attachment to the labor market (in the form of a minimum number of weeks employed) and that the claimant undertake an active search for new employment, claimants' beliefs should provide a good measure of what a job seeker with labor market experience knows at the moment at which a decision is made about accepting a job offer from a particular employer. On the other hand, unemployment benefits claimants may be *less* informed than the average employee about the legal rules governing termination because those

Table 1
Employees' knowledge of legal rules governing discharge

Reason for discharge	State	Legal rule: discharge is	% of total responses asserting discharge is unlawful
Employer plans to hire another person to do the same job at a lower wage	California	Legal	81.3%
	Missouri	Legal	82.2%
	New York	Legal	86.1%
Retaliation for reporting theft by another employee to supervisor	California	Legal	80.9%
	Missouri	Legal	79.2%
	New York	Legal	81.8%
Mistaken belief that employee stole money (employee can prove mistake)	California	Legal	82.7%
	Missouri	Legal	89.4%
	New York	Legal	91.6%
Personal dislike of employee	California	Legal	88.1%
	Missouri	Legal	91.7%
	New York	Legal	90.6%

Source: Kim (1997, 1999).

who are better informed—and thus realize that they have no legal protection against discharge without an adequate justification—may better protect themselves from involuntary termination. Balancing these various factors, there does not appear to be a strong a priori reason to believe that unemployment benefits claimants will be systematically less aware than other employees of the United States rule of at will employment, though only additional empirical study can definitively resolve the question.

Note that because unemployment benefits claimants are disproportionately individuals who earned low to moderate, rather than high, wages prior to their discharge from employment, Kim's evidence bears most directly on the existence of employee-side information failure in employment markets other than those involving highly compensated employees. However, Kim finds that erroneous beliefs about the legal rules governing discharge are common regardless of education level. In Missouri, for instance, respondents with college degrees still exhibited strongly mistaken beliefs about the legal rules governing discharge (Kim, 1997).

Kim's findings suggest that employees often do not have good information about the United States rule of at will employment. While in some search models a limited proportion of informed actors may eliminate the effects of information failure in markets (e.g., Wilde and Schwartz, 1979), Kim's evidence suggests that the proportion of United States employees suffering from information failure is extremely large. While the empirical support for employee-side information failure is thus large, it is, as noted above,

not clear exactly what legal reform would bring the treatment of discharge in the United States in line with employees' expectations or what the costs of such reform would be.

8. Unemployment insurance systems

Unemployment insurance, briefly noted in the preceding section, is an extremely common, although limited, form of employment protection. Unemployment insurance requires the payment of benefits in lieu of wages upon discharge from employment, typically (at least in the United States) without regard for the reason for the discharge except in extreme circumstances. Under the United States system, benefits are financed by a payroll tax that employers must pay on a per-employee basis.

Both theory and evidence relating to unemployment insurance are developed at length in Chapters 13 and 33 of the *Handbook of Public Economics* (Atkinson, 1987; Krueger and Meyer, 2002), and, thus, the reader is referred to that volume for further discussion of the economics of unemployment insurance. In brief, Krueger and Meyer, in the more recent of the two chapters, conclude that the empirical evidence suggests longer periods out of work with more generous unemployment insurance benefits—a finding, like its counterpart in the workers' compensation context from Section 3 above, with ambiguous welfare consequences.

9. Minimum wage rules

Among the most extensively discussed rules within the economic analysis of employment law are minimum wage rules. Under these rules, employers must pay a legally-specified minimum wage to employees. An important opening observation about these rules is that, akin to the analysis in Section 6 above, the usual role of market failure in justifying employment law rules is altered. Even in the absence of market failure, a minimum wage rule may increase total employee income and, thus, at least by this measure, may enhance employee welfare.

For minimum wage rules to increase not only total employee income but also efficiency and the income of each individual worker, market failure is necessary. It is often observed in discussions of minimum wage rules that market failure in the form of monopsony may justify such rules. This is so because, as discussed in Section 1.2 above, under monopsony the wage is set below the marginal revenue product of labor. Anything that moves the wage closer to the marginal revenue product of labor will both enhance efficiency and make all employees better off.

Both theory and evidence relating to minimum wage rules are developed in Chapter 32 of the *Handbook of Labor Economics* (Brown, 1999), and the reader is referred to that volume for further discussion of these rules and their effects on efficiency and employee welfare.

10. Overtime pay requirements

The Fair Labor Standards Act (FLSA) in the United States imposes both minimum wage rules—discussed in the preceding section—and overtime pay requirements, which are the focus of the present section. Under overtime pay requirements, employers are legally required to pay a wage premium for hours above a specified weekly or other threshold. In the United States, employers generally must pay one and a half times the normal wage for hours worked above 40 hours per week. This section describes both theoretical and empirical analysis of overtime pay requirements.

10.1. Theoretical analysis of overtime pay requirements

Overtime pay requirements, like the employment law rules discussed in the preceding sections, represent a legally prescribed term in the employment relationship. However, there is an important analytic difference between overtime pay requirements and the rules considered in all of the preceding sections. In the case of overtime pay requirements, the legally prescribed term simply concerns the *form* in which dollars are paid to an employee, rather than some other aspect (such as the safety, privacy, or wage level) of the job. In principle, then, it may be possible for employers to undo completely the effects of overtime pay requirements (Ehrenberg and Schumann, 1982, pp. 36–37). To borrow Trejo's (1991) example, suppose that in the absence of overtime pay regulation, an employee works 50 hours per week and earns $11 per hour. The employer in this circumstance could satisfy the FLSA by reducing the employee's straight-time wage to $10 per hour and then paying time and a half for the 10 hours in excess of the 40-hour-per-week cutoff. The employee would be in an identical situation, working 50 hours per week for $550 in pay. By contrast, when, for instance, a particular workplace safety mandate is put into effect, employers in the new equilibrium must provide the new safety feature (possibly compensated by lower wages) and cannot directly replicate the prior equilibrium.

Of course, there are limits on the foregoing account of the undoing of overtime pay requirements. To the extent that straight-time wages are at least somewhat sticky, employers may not be able to adjust such wages continuously over time in response to the number of desired overtime hours. In addition, adjustment will be either impossible or limited for employees whose straight-time wages are at or near the legal minimum wage level prior to any requirement of overtime pay. Both of these points are emphasized by Trejo (1991). Ultimately, then, empirical evidence is necessary to determine whether the theoretical account of the irrelevance of overtime pay requirements is true in practice.

To the extent that overtime pay requirements are not completely undone—as a result of various forms of wage stickiness or other factors—is there a market failure to which overtime pay requirements might be thought to be responsive? The conventional view is that intensive use of a smaller set of employees, instead of reliance on a broader group of employees, has a negative externality effect on those individuals who are unable to obtain employment when a smaller group of employees is used intensively. To the extent

that overtime pay requirements increase employment levels to some degree, it is possible that these requirements represent an effective response to the negative externality just noted.

10.2. Empirical analysis of overtime pay requirements

In an effort to determine whether overtime pay requirements are effectively undone by changes in base wages, Trejo (1991) studies differences in base wages across employees covered and not covered by the overtime pay requirements of the FLSA. The noncovered employees used in Trejo's analyses—who are comparable to covered employees in being paid on an hourly basis—include nonsupervisory agricultural workers, many transportation employees, and certain retail and services employees. (The largest group of employees not covered by the overtime pay requirements are employees who are paid on a salary rather than an hourly basis, but these individuals would, for obvious reasons, provide a weak control group for examining the effects of overtime pay requirements on covered, hourly wage employees.) Using a sample of repeated cross sections, Trejo finds that the hourly wages of covered employees are significantly lower than the hourly wages of noncovered employees. According to Trejo, this evidence provides some suggestion that overtime pay requirements are partially, although not completely, undone by adjustments in straight-time wages. Potential limits on Trejo's empirical analysis include the fact that the results are somewhat sensitive to the measure of wages used (with weekly earnings yielding results different from those just described) and the limited degree of identification resulting from possible unobservable differences across individuals in hourly wage jobs within covered versus noncovered sectors. In contrast to Trejo, Hunt (1999), examining evidence from Germany, finds increases rather than decreases in wages in response to strengthened overtime pay requirements negotiated between unions (which are industry-wide in Germany) and employers; wages may well be less flexible in Europe than in the United States. Hunt's identification strategy has the virtue of employing both cross sectional and time series variation.

In addition to examination of whether base wages adjust to offset overtime pay requirements, empirical work has directly studied the linked question of whether workweeks longer than 40 hours in the United States are less frequent with overtime pay requirements in place. (Only with imperfect wage adjustments should such workweeks be less frequent with overtime pay requirements in place.) Costa (2000) examines changes in the frequency of overtime hours with the passage of the FLSA and finds a significant reduction in the proportion of employees working such hours. Costa also finds much greater reduction in the South—where more employees were at the minimum wage level and, thus, could not be subjected to wage adjustments in response to overtime pay requirements—than in the North. By contrast, Trejo (2003), using as a source of identification variation over the 1970s and 1980s in the proportion of employees within an industry who are subject to the overtime pay requirements of the FLSA, finds no effect of overtime pay requirements on the frequency of workweeks longer than 40 hours in specifications that include industry time trends. However, as

Trejo notes, the coefficients on coverage changes are imprecisely estimated, making it difficult to discern whether the absence of a statistically significant effect suggests no underlying effect or simply the limited power of the empirical approach. Hamermesh and Trejo (2000) study not the FLSA but a California state law expanding the set of employees covered by a strict form of overtime pay and find significant negative effects on the level of overtime employees work. None of these studies, however, offers any evidence that, in response to a decline in long work hours, aggregate employment rises in the employment market in question—although it is obvious that a decline in work-weeks longer than 40 hours is a logical predicate for such an increase. Using German data and examining the effects of industry-specific changes in overtime pay requirements, Hunt (1999) finds that, if anything, aggregate employment levels fall rather than rise in response to strengthened overtime pay requirements.

11. Conclusion

This chapter has offered economic analysis (or, at times, referred to economic analysis in the *Handbook of Labor Economics* or the *Handbook of Public Economics*) of the major topics in employment law. Topics addressed above include workplace safety mandates (Section 2), compensation for workplace injuries (Section 3), workplace privacy mandates (Section 4), fringe benefits mandates (Section 5), targeted benefits such as mandated medical and family leave (Section 6), wrongful discharge laws (Section 7), unemployment insurance (Section 8), minimum wage rules (Section 9), and overtime pay requirements (Section 10).[33] There remains much interesting work to be done on many of these topics, including examining the effects of some of the relatively new laws mentioned above. Moreover, while the focus of most existing employment law is on mandating particular features within the employer-employee relationship, recent legal scholarship in employment law has given increased emphasis to prospects for using "default" rather than mandatory rules. Relatively unexplored in existing work is theoretical and empirical analysis of the effects of default versus mandatory rules in the regulation of the employment relationships. In this and other areas, there are important potential future synergies between legal scholarship on employment law and economic analysis of this body of law.

As noted above, this chapter has focused on employment law rather than labor law. However, it bears emphasis that the two fields of law do not operate in any way independently of one another. The contours of labor law, in affecting the prevalence and power of unions, may greatly affect what sorts of employment law rules are enacted and, perhaps most importantly, the degree to which employment law rules are effectively enforced against employers.[34] With respect to the importance of enforcement, it is difficult

[33] A separate area of inquiry concerns the causes (rather than the desirability and the effects) of employment law rules. This question is analyzed by Botero et al. (2004).

[34] For discussion of the difficulties of enforcing employment rules in a non-union context, see Jolls (2004b).

to disagree with Pound (1910, pp. 16, 34–35), who noted long ago that employment law "in action" may look quite different from employment law "on the books."

References

Acemoglu, D., Angrist, J.D. (2001). "Consequences of employment protection? The case of the Americans with Disabilities Act". Journal of Political Economy 109 (5), 915–957.

Aghion, P., Hermalin, B. (1990). "Legal restrictions on private contracts can enhance efficiency". Journal of Law, Economics, and Organization 6 (2), 381–409.

Akerlof, G.A., Dickens, W.T. (1982). "The economic consequences of cognitive dissonance". American Economic Review 72 (3), 307–319.

Arnould, R.J., Grabowski, H. (1981). "Auto safety regulation: an analysis of market failure". Bell Journal of Economics 12 (1), 27–48.

Atkinson, A.B. (1987). "Income maintenance and social insurance". In: Auerbach, A.J., Feldstein, M. (Eds.), Handbook of Public Economics, vol. II. Elsevier, Amsterdam, pp. 779–908.

Autor, D.H. (2003). "Outsourcing at will: the contribution of unjust dismissal doctrine to the growth of employment outsourcing". Journal of Labor Economics 21 (1), 1–42.

Autor, D.H., Donohue, J.J., Schwab, S.J. (2006). "The costs of wrongful-discharge laws". Review of Economics and Statistics 88 (2), 211–231.

Autor, D.H., Duggan, M.G. (2003). "The rise in the disability rolls and the decline in unemployment". Quarterly Journal of Economics 118 (1), 157–205.

Bartel, A.P., Thomas, L.G. (1985). "Direct and indirect effects of regulation: a new look at OSHA's impact". Journal of Law and Economics 28 (1), 1–25.

Blau, F.D., Kahn, L.M. (1999). "Institutions and laws in the labor market". In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 3A. Elsevier, Amsterdam, pp. 1399–1461.

Botero, J.C., Djankov, S., La Porta, R., Lopez-de-Silanes, F., Shleifer, A. (2004). "The regulation of labor". Quarterly Journal of Economics 119 (4), 1339–1382.

Brown, C. (1980). "Equalizing differences in the labor market". Quarterly Journal of Economics 94 (1), 113–134.

Brown, C. (1999). "Minimum wages, employment, and the distribution of income". In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 3B. Elsevier, Amsterdam, pp. 2101–2163.

Camerer, C.F., Kunreuther, H. (1989). "Decision processes for low probability events: policy implications". Journal of Policy Analysis and Management 8 (4), 565–592.

Costa, D.L. (2000). "Hours of work and the Fair Labor Standards Act: a study of retail and wholesale trade, 1938–1950". Industrial and Labor Relations Review 53 (4), 648–664.

Currie, J., Madrian, B.C. (1999). "Health, health insurance and the labor market". In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 3C. Elsevier, Amsterdam, pp. 3309–3416.

Cutler, D.M. (2002). "Health care and the public sector". In: Auerbach, A.J., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 4. Elsevier, Amsterdam, pp. 2143–2243.

Cutler, D.M., Madrian, B.C. (1998). "Labor market responses to rising health insurance costs: evidence on hours worked". RAND Journal of Economics 29 (3), 509–530.

DeJoy, D.M. (1989). "The optimism bias and traffic accident risk perception". Accident Analysis and Prevention 21 (4), 333–340.

Donohue, J.J. (2007). "Antidiscrimination law". In: Polinsky, A.M., Shavell, S. (Eds.), Handbook of Law and Economics. Elsevier, Amsterdam.

Dworkin, R.M. (1980). "Is wealth a value?" Journal of Legal Studies 9 (2), 191–226.

Ehrenberg, R.G., Schumann, P.L. (1982). Longer Hours or More Jobs? An Investigation of Amending Hours Legislation to Create Employment. Cornell University, Ithaca.

Gibbons, R., Katz, L.F. (1991). "Layoffs and lemons". Journal of Labor Economics 9 (4), 351–380.

Gray, W.B., Mendeloff, J.M. (2005). "The declining effects of OSHA inspections on manufacturing injuries, 1979–1998". Industrial and Labor Relations Review 58 (4), 571–587.
Greenwald, B.C. (1986). "Adverse selection in the labour market". Review of Economic Studies 53 (3), 325–347.
Gruber, J. (1994a). "The incidence of mandated maternity benefits". American Economic Review 84 (3), 622–641.
Gruber, J. (1994b). "State-mandated benefits and employer-provided health insurance". Journal of Public Economics 55 (3), 433–464.
Gruber, J., Krueger, A.B. (1991). "The incidence of mandated employer-provided insurance: lessons from workers' compensation insurance". In: Bradford, D. (Ed.), Tax Policy and the Economy, vol. 5. MIT Press, Cambridge, MA, pp. 111–143.
Gruber, J., Madrian, B.C. (1994). "Health insurance and job mobility: the effects of public policy on job-lock". Industrial and Labor Relations Review 48 (1), 86–102.
Gruber, J., Madrian, B.C. (1995). "Health-insurance availability and the retirement decision". American Economic Review 85 (4), 938–948.
Gruber, J., Madrian, B.C. (1996). "Health insurance and early retirement: evidence from the availability of continuation coverage". In: Wise, D.A. (Ed.), Advances in the Economics of Aging. University of Chicago Press, Chicago, pp. 115–143.
Gruber, J., Madrian, B.C. (1997). "Employment separation and health insurance coverage". Journal of Public Economics 66 (3), 349–382.
Hamermesh, D.S., Trejo, S.J. (2000). "The demand for hours of labor: direct evidence from California". Review of Economics and Statistics 82 (1), 38–47.
Hunt, J. (1999). "Has work-sharing worked in Germany?" Quarterly Journal of Economics 114 (1), 117–148.
Ippolito, R.A. (1988). "A study of the regulatory effect of the Employee Retirement Income Security Act". Journal of Law and Economics 31 (1), 85–125.
Jacobson, M. (2003). "Drug testing in the trucking industry: the effect on highway safety". Journal of Law and Economics 46 (1), 131–156.
Jolls, C. (2000). "Accommodation mandates". Stanford Law Review 53 (2), 223–306.
Jolls, C. (2004a). "Identifying the effects of the Americans with Disabilities Act using state-law variation: preliminary evidence on educational participation effects". American Economic Review 94 (2), 447–453.
Jolls, C. (2004b). "The role and functioning of public-interest legal organizations in the enforcement of the employment laws". In: Freeman, R.B., Hersch, J., Mishel, L. (Eds.), Emerging Labor Market Institutions for the Twenty-First Century. University of Chicago Press, Chicago, pp. 141–176.
Jolls, C. (2007). "Workplace leave mandates and the employment of individuals with disabilities". Manuscript.
Jolls, C., Prescott, J.J. (2007). "Disaggregating employment protection: the case of disability discrimination". Manuscript.
Kaestner, R., Simon, K.I. (2002). "Labor market consequences of state health insurance regulation". Industrial and Labor Relations Review 56 (1), 136–159.
Kaplow, L., Shavell, S. (1994). "Why the legal system is less efficient than the income tax in redistributing income". Journal of Legal Studies 23 (2), 667–681.
Kim, P.T. (1997). "Bargaining with imperfect information: a study of worker perceptions of legal protection in an at-will world". Cornell Law Review 83 (1), 105–160.
Kim, P.T. (1999). "Norms, learning, and law: exploring the influences on workers' legal knowledge". University of Illinois Law Review 1999 (2), 447–515.
Krueger, A.B., Meyer, B.D. (2002). "Labor supply effects of social insurance". In: Auerbach, A.J., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 4. Elsevier, Amsterdam, pp. 2327–2392.
Levine, D.I. (1991). "Just-cause employment policies in the presence of worker adverse selection". Journal of Labor Economics 9 (3), 294–305.
Mendeloff, J. (1979). Regulating Safety: An Economic and Political Analysis of Occupational Safety and Health Policy. The MIT Press, Cambridge, MA.

Miles, T.J. (2000). "Common law exceptions to employment at will and U.S. labor markets". Journal of Law, Economics, and Organization 16 (1), 74–101.
Morantz, A.D. (2005). "Has regulatory devolution injured American workers? A comparison of state and federal enforcement of construction safety regulations". Manuscript.
Neumark, D., Stock, W.A. (1999). "Age discrimination laws and labor market efficiency". Journal of Political Economy 107 (5), 1081–1125.
Posner, R.A. (1987). "The efficiency and the efficacy of Title VII". University of Pennsylvania Law Review 136 (2), 513–521.
Pound, R. (1910). "Law in books and law in action". American Law Review 44 (1), 12–36.
Rea, S.A. (1981). "Workmen's compensation and occupational safety under imperfect information". American Economic Review 71 (1), 80–93.
Ruhm, C.J. (1998). "The economic consequences of parental leave mandates: lessons from Europe". Quarterly Journal of Economics 113 (1), 285–317.
Ruser, J.W., Smith, R.S. (1988). "The effect of OSHA records-check inspections on reported occupational injuries in manufacturing establishments". Journal of Risk and Uncertainty 1 (4), 415–435.
Sanz-de-Galdeano, A. (2006). "Job-lock and public policy: Clinton's second mandate". Industrial and Labor Relations Review 59 (3), 430–437.
Scholz, J.T., Gray, W.B. (1990). "OSHA enforcement and workplace injuries: a behavioral approach to risk assessment". Journal of Risk and Uncertainty 3 (3), 283–305.
Shavell, S. (2007). "Liability for accidents". In: Polinsky, A.M., Shavell, S. (Eds.), Handbook of Law and Economics. Elsevier, Amsterdam.
Simon, K.I. (2005). "Adverse selection in health insurance markets? Evidence from state small-group health insurance reforms". Journal of Public Economics 89 (9-10), 1865–1877.
Smith, R.S. (1976). The Occupational Safety and Health Act: Its Goals and Its Achievements. American Enterprise Institute for Public Policy Research, Washington, D.C.
Smith, R.S. (1992). "Have OSHA and workers' compensation made the workplace safer?" In: Lewin, D., Mitchell, O.S., Sherer, P.D. (Eds.), Research Frontiers in Industrial Relations and Human Resources. Industrial Relations Research Association, Madison, pp. 557–586.
Summers, L.H. (1989). "Some simple economics of mandated benefits". American Economic Review 79 (2), 177–183.
Trejo, S.J. (1991). "The effects of overtime pay regulation on worker compensation". American Economic Review 81 (4), 719–740.
Trejo, S.J. (2003). "Does the statutory overtime premium discourage long workweeks?" Industrial and Labor Relations Review 56 (3), 530–551.
Viscusi, W.K. (1978). "Wealth effects and earnings premiums for job hazards". Review of Economics and Statistics 60 (3), 408–416.
Viscusi, W.K. (1979). "The impact of occupational safety and health regulation". Bell Journal of Economics 10 (1), 117–140.
Viscusi, W.K. (1983). Risk by Choice: Regulating Health and Safety in the Workplace. Harvard University Press, Cambridge, MA.
Viscusi, W.K. (1986). "The impact of occupational safety and health regulation, 1973–1983". RAND Journal of Economics 17 (4), 567–580.
Viscusi, W.K., O'Connor, C.J. (1984). "Adaptive responses to chemical labeling: Are workers Bayesian decision makers?" American Economic Review 74 (5), 942–956.
Waldfogel, J. (1998). "Understanding the 'family gap' in pay for women with children". Journal of Economic Perspectives 12 (1), 137–156.
Waldfogel, J. (1999). "The impact of the Family and Medical Leave Act". Journal of Policy Analysis and Management 18 (2), 281–302.
Weil, D. (1996). "If OSHA is so bad, why is compliance so good?" RAND Journal of Economics 27 (3), 618–640.
Weil, D. (2001). "Assessing OSHA performance: new evidence from the construction industry". Journal of Policy Analysis and Management 20 (4), 651–674.

Weinstein, N.D. (1980). "Unrealistic optimism about future life events". Journal of Personality and Social Psychology 39 (5), 806–820.
Wilde, L.L., Schwartz, A. (1979). "Equilibrium comparison shopping". Review of Economic Studies 46 (3), 543–553.

Chapter 18

ANTIDISCRIMINATION LAW

JOHN J. DONOHUE*

School of Law, Yale University, and National Bureau of Economic Research

Contents

* The author wishes to thank Mitch Polinsky, Steve Shavell, Peter Siegelman, Paul Oyer, and Devah Pager for their extremely valuable comments, and Stanford and Yale Law Schools for research support. Michael Gottfried, Christopher Griffin, Seth Stephens-Davidowitz, and Maile Tavepholjalern provided outstanding research assistance.

Handbook of Law and Economics, Volume 2
Edited by A. Mitchell Polinsky and Steven Shavell

DOI: 10.1016/S1574-0730(07)02018-X

Abstract

This essay provides an overview of the central theoretical law and economics insights and empirical findings concerning antidiscrimination law across a variety of contexts including discrimination in labor markets, housing markets, consumer purchases, and policing. The different models of discrimination based on animus, statistical discrimination, and cartel exploitation are analyzed for both race and sex discrimination. I explore the theoretical arguments for prohibiting private discriminatory conduct in light of the tensions that exist between concerns for liberty and equality. I also discuss the complexities in empirically establishing the existence of discrimination and highlight the critical point that one cannot automatically attribute observed disparities in various economic or social outcomes to discrimination. The major empirical findings showing the effectiveness of federal law in the first decade after passage of the 1964 Civil Rights Act are contrasted with the generally less optimistic findings from more recent antidiscrimination interventions.

Keywords

Discrimination, disparate impact, disparate treatment, profiling, statistical discrimination, antidiscrimination law and economic welfare

JEL classification: D72, J70, J71, J78, K31, K42

1. Introduction

The last century of world history has been marred by episodes of appalling mistreatment of various racial, ethnic, and religious groups.[1] While the acts of discrimination against and mistreatment of women around the world may have been less visible, their consequences have been even more arithmetically compelling if one credits Amartya Sen's conclusion that 100 million women are missing.[2] In this country, the horrors of slavery and the Civil War ultimately led to the first great body of American antidiscrimination law, which was followed by a century of rigid racial discrimination in the Jim Crow South. The evils of the Holocaust gave considerable impetus to the emergence of the second great body of antidiscrimination law after World War II, when New York and New Jersey became the first states to prohibit discrimination in employment. In the ensuing decades, most states followed their lead.

The final phase of expanding federal antidiscrimination law began with the adoption of the Civil Rights Act of 1964. This federal ban on discrimination, which played a critical role in dismantling the rigid discriminatory social order of the South, is now used to regulate an enormous array of social institutions from the workplace and schools to public accommodations and policing. Other important developments in the federal effort to dismantle southern racial segregation were the Supreme Court's rejection of the separate but equal doctrine in *Brown v. Bd. of Education* (1954) and the Voting Rights Act of 1965. Antidiscrimination law has been growing dramatically in scope for at least the last half century and has revolutionized the American conception of the proper role of government.[3]

Today, most Americans—and virtually all public officials—have embraced prohibitions on discrimination as an important constraint on both private contracting and public action. Moreover, even as issues such as a ban on discrimination against gays and certain types of affirmative action have generated opposition, the reach of antidiscrimination law has never been greater.

The growth in the scope of antidiscrimination law can be seen in the language of Section 12920 of California's Fair Employment and Housing Act (FEHA), which states:

> It is hereby declared as the public policy of this state that it is necessary to protect and safeguard the right and opportunity of all persons to seek, obtain, and hold employment without discrimination or abridgment on account of race, religious

[1] See Power (2002). Turkey's killing of nearly one million Armenians, the Nazi's killing of six million Jews, Iraq's slaughter of more than one hundred thousand Kurds, Bosnian Serbs' murder of some two hundred thousand Muslims and Croats, and the Rwandan Hutu's slaughter of 800,000 Tutsi—among the most horrific events of the last century—were all motivated or aggravated by racial/ethnic/religious antagonism.

[2] Sen (1990) claims this is primarily because women have not received comparable care to men in health, medicine, and nutrition. Oster (2005) argues that perhaps half of this shortfall is due not to discrimination but to an unusual effect of prevalent Hepatitis B virus that results in a greater percentage of male births (perhaps because of higher miscarriage rates of female fetuses).

[3] For further materials on the issues raised in this paper, see Donohue (2003).

> creed, color, national origin, ancestry, physical disability, mental disability, medical condition, marital status, sex, age, or sexual orientation.

When employers are not subject to such legal restrictions, advertisements seeking workers described along these precise characteristics—such as "young, white females" or "married, white male"—are abundant. In 1960, American newspapers were full of such now-prohibited advertisements, and in areas of the world that don't have antidiscrimination laws such ads are common today.[4] The growing power of antidiscrimination law in the United States represents a dramatic rejection of classical liberal notions of freedom of contract. Again, Section 12920 of California's FEHA offers the following, sweeping rationale for the legal prohibition:

> It is recognized that the practice of denying employment opportunity and discriminating in the terms of employment for these reasons foments domestic strife and unrest, deprives the state of the fullest utilization of its capacities for development and advancement, and substantially and adversely affects the interest of employees, employers, and the public in general.

The California statute goes on to set forth a similarly long list of prohibited bases for actions concerning housing accommodations (excluding the prohibition on age discrimination while adding a prohibition based on familial status). Many states have responded to tobacco industry lobbying and now prohibit "discrimination" against cigarette smokers,[5] and a few jurisdictions even ban discrimination on the basis of all physical characteristics.[6] With crime dropping sharply throughout the United States in the 1990s, the

[4] See Darity and Mason (1998). For example, want ads seeking candidates specified by race/ethnicity, gender and age are published online in the classifieds of the *New Straits Times* in Malaysia, which has no law prohibiting private employment discrimination in a country that is roughly 1/3 Bumiputra (ethnic Malay and religiously Muslim), 1/3 Indian, and 1/3 Chinese.

[5] Finkin (1996) describes the recent trend as follows:

> Employers commonly forbid drinking on the job and, more recently, prohibit smoking on the premises. But some employers have gone much further, and refuse to hire or retain employees who drink or smoke at all, in an effort to reduce medical insurance costs attendant to such behavior. Unlike many of the other invasions of individual autonomy, these policies have drawn the attention of politically influential groups, i.e., the tobacco and alcohol interests. Consequently, eighteen states now expressly forbid discrimination on the basis of off-premises use of tobacco; one forbids discrimination on the basis of off-premises use of tobacco or alcohol; and six forbid discrimination on the basis of off-premises use of any "lawful product." Absent such legislation, however, nothing prohibits such employer commands .... Colorado forbids discrimination by employers for engagement in "any lawful activity" off the employer's premises and New York forbids discrimination for engagement in "legal recreation activities."

Moreover, a new argument designed to protect smokers is that smoking is an addiction protected by the Americans with Disabilities Act. While it is doubtful that this argument will prevail, it does underscore the continued pressure for extending the reach of American antidiscrimination law.

[6] For example, the city of Santa Cruz has an antidiscrimination ordinance [Chapter 9.83.02(13)] that goes beyond the law of California to ban discrimination on the basis of "height, weight or physical characteristic."

American Civil Liberties Union (ACLU) launched a major campaign against discrimination in policing—so-called "racial profiling"—which was gaining wide support, at least until the events of 9/11 rekindled the argument that some types of profiling might serve useful law enforcement functions.[7] Meanwhile, lawsuits challenging discriminatory practices in mortgage lending, housing practices, insurance sales, and the financing of automobiles have been prominent elements of the attack on allegedly discriminatory business practices.

Judged by any measure of legal significance, American antidiscrimination law has continued to grow in importance. It has assumed a large place in current domestic legal consciousness and has been a major legal export as other countries have emulated the U.S. legal prohibitions. Another noteworthy reflection of the importance of this topic is the caliber of the contributors to the scholarship in this area—the Nobel economists alone include Gary Becker, Kenneth Arrow, Milton Friedman, George Akerlof, Amartya Sen and James Heckman. Within the legal academy, major contributions have come from many top law and economics scholars, including Richard Posner, Robert Cooter, Richard Epstein, Ian Ayres, and Christine Jolls. While a simple vision of animus-based discrimination, which Becker fashioned into the first economic theory of employment discrimination, has motivated elected officials and the public to support an elaborate body of law committed to the eradication of "discrimination" against an ever-widening group of claimants, the issue is far too multi-faceted and the nature of the protected classes far too diverse to be adequately encompassed by a single theoretical framework. In addition, the goals of antidiscrimination law have evolved: initially, legal and economic notions of "discrimination" were broadly compatible, but, over time, an increasingly expansive legal conception of discrimination has come into greater conflict with the economic notion of discrimination. As the ambitions of antidiscrimination law have advanced beyond the narrow task of eliminating economic inequities to promote broader goals of distributive justice, the costs imposed by the regulatory framework and the tensions between the demands of law and the forces of the market have grown. This chapter will address these issues from both a theoretical and empirical basis, while shedding light on what antidiscrimination law has accomplished in the past and what it is achieving today.

Section 2 begins with a brief overview of American antidiscrimination law that defines the key legal concepts of disparate treatment and disparate impact discrimination and notes the breadth of American social life that is governed by the far-reaching regulatory apparatus. Section 3 discusses the basic economic theories of discrimination and highlights their virtues and shortcomings. Section 4 then explores the theoretical arguments for prohibiting private discriminatory conduct and illustrates the tensions that exist between concerns for liberty and equality. Section 5 makes the critical point that one cannot automatically attribute observed disparities in various economic or social

[7] As Schauer (2003) notes: "As a generalization, the principle of treating all equally is a principle that ignores real differences—and consequently comes at a price."

outcomes to discrimination, and Section 6 illustrates the complexities in establishing the existence of discrimination. Section 7 discusses some practical problems with antidiscrimination law, revolving around the difficulties of motive-based litigation and the dangers of Type I and Type II error as well as the costs of preventing the use of efficient statistical discrimination. Section 8 then discusses some of the major empirical studies evaluating the impact of antidiscrimination law. One important message from this literature is that the initial adoption of Title VII of the 1964 Civil Rights Act aided black economic welfare but that further efforts to strengthen federal antidiscrimination law have been subject to the law of diminishing returns. Section 9 discusses the evidence on whether antidiscrimination law has aided female workers, examines the data on premarket factors that may influence female labor outcomes, and describes the development of one particular strand of the ban on sex discrimination—harassment on the basis of sex. The literature on discrimination in mortgage lending and major consumer markets is outlined in Section 10. Section 11 discusses "racial profiling" in policing, and Section 12 concludes.

2. The contours of antidiscrimination law

Since the various legislative mandates against "discrimination" generally offer little further guidance on what those prohibitions mean, the development of the precise contours of antidiscrimination law over the last half century has largely been the product of judicial decision-making. For example, the prohibition embodied in Title VII of the 1964 Civil Rights Act was initially thought to extend only to *intentional* employment discrimination, or so-called "disparate treatment" discrimination. The courts would ask whether the plaintiff would have been treated differently if he or she did not have the particular trait in question.[8] In the 1971 case of *Griggs v. Duke Power*, the Supreme Court fashioned an additional and potentially more sweeping theory of discrimination. This so-called "disparate impact" doctrine prohibited facially neutral acts that had an adverse impact on certain protected classes unless the employer could offer a sufficiently compelling justification for the practice. In the workplace, typical practices that might be challenged include the use of screening tests for employment or for promotion. In policing, a disparate impact charge might be used to challenge drug enforcement efforts that involved targeting certain cars or driving conduct. Determining what legally "justifies" conduct generating a disparate racial or ethnic impact involves some balancing of the benefits generated by the practice in question versus the costs to the group that is differentially impacted.

The Supreme Court has held, though, that this disparate impact doctrine is not available to litigants who base their claim of discrimination on the Constitution (typically

[8] One major issue of contention raised by this definition was whether it would prevent employers from engaging in voluntary affirmative action that provided some advantage to workers having one of the protected characteristics. This issue will be discussed below.

under the Fifth or Fourteenth Amendments) or to those suing under Section 1981 of the federal code (which provides a remedy only for intentional racial discrimination).[9] One or both elements of this disparate treatment/disparate impact structure have been applied to legal challenges to discrimination in a large array of different domains:

a. **labor markets** (Title VII of the 1964 Civil Rights Act is the primary federal law and most states have similar—or in states such as California, more stringent—prohibitions. The major judicial expansions were the creation of the disparate impact standard in 1971 and bringing sexual harassment within the ambit of the prohibition of sex discrimination in 1986.[10] The Age Discrimination in Employment Act and the Americans with Disabilities Act represent the two major non-Title VII legislative expansions of federal antidiscrimination law, with the latter explicitly going beyond the prior notion of prohibiting discrimination to mandating that employers provide "reasonable" accommodations to "qualified" disabled workers.[11])
b. **education** (Title VI prohibits discrimination in any program that receives federal funds)
c. **criminal justice and racial profiling** (the Fourteenth Amendment and Title VI have been used to challenge racially disparate outcomes in death penalty cases, drug and traffic enforcement, and street policing)
d. **the provision of health care services** (Fourteenth Amendment and Title VI)
e. **housing and lending** (The Fair Housing Act and Equal Credit Opportunity Act)
f. **purchase of goods and services** (Section 1981 prohibits intentional racial discrimination).

The threat of employment discrimination litigation, along with pressure on federal government contractors to comply with the antidiscrimination and affirmative action requirements of Executive Order 11422, has led many firms to develop affirmative action plans to reduce perceived shortfalls in the employment share of minority and female workers.[12] There is an obvious tension between the establishment of a race or gender-based affirmative action plan and the statutory language of Section 703(m) of the Civil Rights Act of 1991, which states that "an unlawful employment practice is established

[9] The enduring disagreement within the federal judiciary as to whether the disparate impact doctrine is applicable to cases brought under the federal Age Discrimination in Employment Act has recently been answered in the affirmative *Smith v. City of Jackson*, 125 S. Ct. 1536 (2005). The same decision had been imposed legislatively in California in 1999.

[10] In 1980, the Equal Employment Opportunity Commission issued "Guidelines on Discrimination Because of Sex," which declared sexual harassment a violation of Title VII of the 1964 Civil Rights Act. These guidelines defined two distinct categories of sexual harassment: "quid pro quo" and "hostile environment" harassment. In the 1986 case of *Vinson v. Meritor Bank*, the U.S. Supreme Court affirmed that "hostile environment" sex harassment violated Title VII.

[11] According to the U.S. Census Bureau, in 1997 about 52.6 million people (19.7 percent of the population) had some type of disability, and among those with a disability, 33 million people (12.3 percent of the population) had a severe disability.

[12] See Leonard (1984a, 1984b), Ashenfelter and Heckman (1976), and Heckman and Wolpin (1976).

when the complaining party demonstrates that race, color, religion, sex, or national origin was a motivating factor for any employment practice." The courts have resolved this tension by allowing *private* employers to grant preferences on race or gender grounds but only in the limited circumstances where there is a manifest imbalance in the employer's workforce and any preferential treatment does not unduly burden members of the non-preferred group.[13] *Governmental* affirmative action is subject to strict scrutiny under the equal protection clause of the Fourteenth Amendment (or under the Fifth Amendment in the case of federal governmental action), and the permissible scope of such affirmative action has been defined through a series of Supreme Court decisions. California has adopted a constitutional amendment—Proposition 209—that prohibits any form of preferential treatment based on race or gender in *state* employment or in the operation of other *state* functions, such as education and contracting for construction and other services. Private employers and universities in California are still free to pursue such race and gender-based preferences as long as they do not overstep the bounds of federal (or state) antidiscrimination law.

3. Theories of discrimination

Although the contexts in which discrimination is found vary widely, ranging from labor markets and health care to housing, education and the purchase of automobiles, the reasons for such discrimination are relatively few. In fact, the implicit internal justification of discriminators in any area generally falls into one of four categories: (1) "we don't like you" (aversion) or, what is often functionally equivalent, "we prefer someone other than you" (nepotism); (2) "it is in our self-interest to cater to the aversion or nepotism of others even though we don't share those feeling ourselves"; (3) "we can further our independent interests by acting to subordinate another group, either because such actions enhance our self-esteem or undermine economic competition from your group"; and (4) "taking your particular trait into account can help us achieve a legitimate goal more effectively." As we will see, the first two of these are predicated on certain individuals having a "taste" for discrimination, which actually tends to harm the one having such a taste, at least if it is being expressed in a competitive market setting. The third and fourth categories involve the strategic use of discrimination to further one's own interests, in some cases intentionally imposing burdens on the disadvantaged group (Category 3—the cartel model) and in others simply burdening certain members of the group without necessarily harming the group as a whole unless the group's incentives to invest in their human capital is impaired (Category 4—statistical discrimination).

To illustrate, the employer who dislikes blacks and therefore refuses to hire black applicants or the police officer who stops a black driver out of animus and gives a ticket when he would not have done so had the driver been white is engaging in animus-based discrimination (Category 1). The airline that exclusively hires young female flight

[13] *United Steelworkers v. Weber*, 443 U.S. 193 (1979).

attendants to cater to the preferences of its passengers or the restaurant that hires only white servers in response to customer preferences is acceding to the bias or nepotism of others (Category 2). The prototype of Category 3 discrimination was the informal Jim Crow restrictions that kept blacks out of the southern textile industry throughout the first two-thirds of the 20th century. The drug enforcement agent who finds that using race or ethnicity can increase the likelihood of making drug seizures, or the employer who feels that such traits are useful proxies for certain productivity-related traits is engaging in statistical discrimination (Category 4). Such behavior, though likely widespread, is illegal, whether the stereotypes are accurate or inaccurate.

Importantly, an actor who doesn't consider any protected characteristic but who acts directly on factors that are universally accepted as legitimate is not engaging in discrimination, even if a racial or ethnic disparity emerges.[14] Thus, the FBI reports that 51 percent of homicide offenders in 2002 were black, while 50 percent of those arrested for homicide were black. Even though only 12.7 percent of the population is black, the close correlation between the race of homicide offenders and race of homicide arrestees undermines the view that the police are invidiously discriminating in making homicide arrests. Assuming all the arrests are accurate, this is not discrimination by the police.[15] The next four subsections will discuss the four categories of discrimination in the context of employment discrimination, but, as we have just seen, they apply to other contexts as well.[16]

[14] Determining what factors constitute "legitimate" bases for action has been complicated since the Supreme Court created the disparate impact theory of discrimination. Disparate impact discrimination, which is prohibited in the employment realm by Title VII of the 1964 Civil Rights Act, involves the use of a neutral proxy *other than race or some other protected trait*. Because the decision-maker in such disparate impact cases is not relying on a directly relevant factor, this conduct differs from my illustration of a nondiscriminatory judgment, and because it does not rely directly on race (or other protected trait), it differs from my definition of statistical discrimination. (This distinction can break down if a court allows a disparate impact challenge to a subjective employment process that uses only factors "universally accepted as legitimate.")

[15] Of course, the pattern may result from discrimination elsewhere in society, but in a system where "discrimination" is often penalized heavily (both in terms of social opprobrium and through monetary damages), it is important to be clear about who is—and is not—engaging in discriminatory conduct. Similarly, the results of the recent New Haven taxi-cab study in Ayres, Vars, and Zakariya (2005) suggest that black drivers receive considerably smaller tips than white drivers. If this differential accurately reflected differences in driver service (for example, less help with bags or knowledge about locations), then there would be no discrimination. Instead, the authors conclude the disparity results from the animus-based discrimination of taxi passengers. Note, though, that a system designed to reduce discrimination by including tips in the cab fare would help black drivers, while hurting black customers since the latter tended to provide lower tips than white customers (which may not be surprising given the links between race and income).

[16] The theories and empirical studies discussed in subsequent sections have been chosen for their applicability to the functions and forms of antidiscrimination law. For a more general discussion of the theoretical and empirical literature on discrimination, see Altonji and Blank (1999) and Cain (1986).

3.1. Employer discrimination

The core understanding of the electorate about the nature of discrimination and the motivating dynamic propelling the expanding prohibitions against discrimination has consistently been that animus by prejudiced employers is pervasive and seriously harms the employment prospects of women and minorities. According to this view, elderly and disabled workers are disadvantaged, not only because on average their age and incapacities lower their productivity but also because of irrational bias. As will be discussed below, one of the key features of this type of discrimination is that it burdens not only the victim, but also imposes a cost on the *discriminator*.

3.1.1. The Becker employer-animus model

Gary Becker's theory of employer animus-based discrimination attempts to model this conception of discrimination by positing that a discriminating employer must bear not only the wage, w, when he or she hires a disfavored worker but also pay a discriminatory psychic penalty, δ.[17] The following condition should hold for the last disfavored worker hired:

$$mp = w + \delta, \tag{1}$$

where *mp* is marginal productivity. Becker's model of the discriminating employer is mathematically equivalent to a case in which the government imposes a variable tax on employers who hire workers of a certain group (defined by race, gender, or other immutable trait), where the tax ranges from zero (for the non-discriminators) to the maximum value M (the value of δ for the most highly discriminating employer).[18] Just as an employment tax would be expected to lower the quantity demanded and earnings of workers, the psychic "tax" lowers the quantity demanded and earnings of disfavored workers. Of course, if there is enough heterogeneity in the population of employers, the actually observed psychic tax should be far less than M, since disfavored workers will tend to gravitate to employers who bear a smaller (or no) psychic penalty.

Conversely, the greater the number of disfavored workers in any labor market, the greater the observed value of δ will be in that labor market as the marginal disfavored workers will have to deal with employers characterized by higher values of δ in order to be absorbed into the market. Indeed, this was one of the primary motivations of Becker's model: it provided an explanation for the greater apparent discriminatory penalty in the South without resorting to differences in tastes for discrimination between northern and southern employers. Even if employers in both regions were identical in

[17] The variable δ is also known as the discrimination coefficient.

[18] Conceivably, the tax could be negative, suggesting that the relevant group of workers is preferred rather than disfavored. This, then, could be thought of as a model of nepotism (or attraction to workers possessing certain non-productivity-related desirable traits).

the extent of their racial prejudice, the much higher percentage of black residents in the South meant that to find jobs, blacks had to associate with increasingly higher δ employers, leading to widening earnings disparities between black and white workers. The Becker model of employer animus thus provided an explanation for an important observed phenomenon—the greater black-white earnings disparity in the South than the North—without resorting to "difference in tastes" as the cause.

The model also generates a number of other predictions, which can be illustrated with the simple supply and demand model of Figure 1. The intersection of demand curve D_1 and the supply curve S illustrates the short-term equilibrium for the market for black workers in a world without discrimination. Black workers would earn a wage of W_1, and Q_1 black workers will be hired. Becker models the introduction of employer discrimination by positing a downward shift in the demand curve to D_2 (which, as a way of reflecting the higher "cost" of black workers, could alternatively be modeled as a pseudo-upward shift in the supply curve, as indicated by Equation (1) above—see Donohue (1986) and Donohue (1989)). In this simple case, it is assumed that the vertical distance AC reflects the uniform psychic penalty associated with hiring black workers. The consequence of this discriminatory animus is that fewer blacks would be hired (only Q_2 instead of Q_1) and black wages would fall from W_1 to W_2. Thus, at least in the short-run, the Becker model predicts that the disfavored group will experience job losses and receive lower pay relative to the non-discriminatory equilibrium given by point A in Figure 1.

As Donohue (1987) underscored, the Becker model predicts that discriminating employers are hurt by their discriminatory preference in that their net profit (monetary

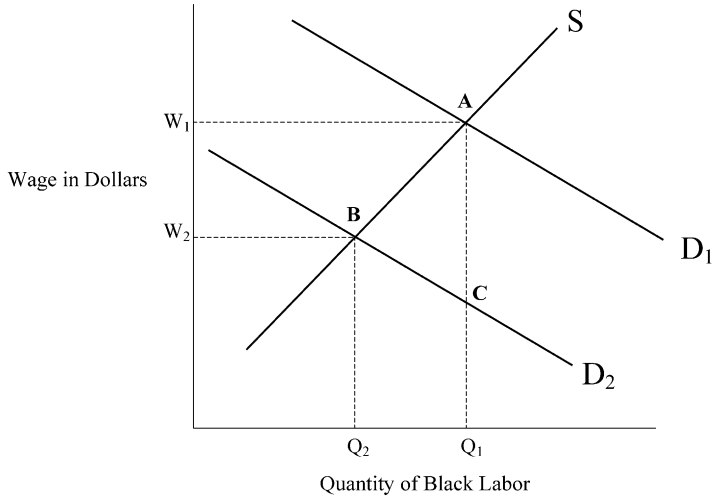


Figure 1. The effect of discrimination on the wages and quantity of black labor under the Becker employer-animus model.

minus psychic penalty) is actually larger in the non-discriminatory world than in the discriminatory world.[19] Note, too, that in the example illustrated in Figure 1, while the discriminatory employer's net benefit is lower, this employer's total monetary reward is greater (because the entire psychic cost denoted by *AC* represents money that ends up in the pocket of the discriminator). In summary, the presence of discrimination can cost jobs and money to the disfavored group and increase income inequality as black earnings fall and the monetary profits of discriminators rise (even as their total psychic welfare declines). One might use these conclusions as the basis for constructing an equity argument for prohibiting employer animus-based discrimination. In this case, equity would be defined by reference to the non-discriminatory equilibrium, and thus legal efforts to move closer to that equilibrium would be defined as promoting equity.[20] Conversely, if one were willing to value all preferences expressed in the market and no others (as many economists are inclined to do), then any attempt of law to move away from the discriminatory equilibrium would undermine efficiency (since the efficient solution would be defined by intersection *B* in Figure 1). If preferences for redistribution or racial justice (or against racial discrimination) are honored, however, then antidiscrimination law can have a role.[21] Note that the Becker model assumes away any possible harm to victims other than lost wages. Since juries now routinely award compensatory damages in addition to back pay to victims of discrimination, this suggests that the dishonor or psychological costs of discrimination are deemed to be significant.

An important conclusion from the Becker model is that, if there are enough non-discriminatory employers around, disfavored workers may not suffer any monetary penalty. In this event, blacks would work for the non-discriminators and would be paid

[19] Note that the "employer surplus" is the area under the demand curve above the relevant wage. Clearly, this is lower for demand curve D_2 than for demand curve D_1.

[20] Note that this is not the customary equity argument based on the perceived social benefit of enhancing the wealth of the least well off members of society, although coincidentally this value may be served by antidiscrimination policy because of the lower wealth of women and minorities. For this reason, taxes and transfers, which are generally preferred to the use of the legal system as a way to redistribute wealth may not be appropriately targeted to achieve the non-discriminatory equilibrium. Still, it is possible that tax incentives to hire women and minorities might be a more efficient mechanism to reach the non-discriminatory equilibrium at least cost (in light of the not inconsiderable costs of a litigation-based antidiscrimination regime in which difficult issues of motives must be determined by judges and juries).

[21] One can craft a theoretical argument based on the Coase Theorem that in a zero transactions cost world, one would expect discrimination to be banned if the costs of discrimination exceeded their benefits. In this calculus, the preferences of those who are offended by racism and other forms of prejudice would be weighed fully. The outcome would then depend on whether the victims of discrimination and those who decried their plight would be willing and able to compensate those who gained from discrimination. The operation of free labor markets in a discriminatory environment will not reveal the answer to this empirical question since transactions costs will prevent the opponents of discrimination from contracting with the discriminators. (Query whether in practice an effort to bargain for respect will ipso facto be futile, since a Coasean payment to induce someone to respect the payer in itself undermines the payer's sense of self-worth. Of course, this problem can always be elided by considering it to be a violation of the zero transactions costs assumption.)

equally to white workers working for the discriminatory employers. Thus, even in the presence of discriminatory attitudes by some employers, effective discrimination (in terms of lower black wages and employment) may be eliminated by the operation of the market, albeit by encouraging the segregation of workers. Correspondingly, mandated segregation will increase the Beckerian psychic costs of discrimination by increasing the interaction between discriminatory employers and disfavored workers.

3.1.2. An empirical challenge to the Becker model

Note that we have been discussing the short-run predictions of the Becker model, which at first would appear to explain lower black wages and employment—at least for the case of the Jim Crow South and the discrimination confronted by southern black workers prior to the adoption of the Civil Rights Act of 1964. But what would happen in the long run under the Becker model? The market should discipline such discriminators, and, at least under constant returns to scale, should ultimately drive out the discriminators. For this reason, Milton Friedman argued that legal prohibitions on discrimination in employment would be unnecessary since the market would solve the problem. But this is where the Becker employer-animus model ran into problems. It failed to explain the enduring exclusion of blacks from entire industries in which they were fully capable of performing the work.[22] In the competitive market that Becker premised, the cost to discriminators of forsaking talented black workers should have engendered a painful market response, ultimately leading to the elimination of the discriminators if the employers operated in a world of constant returns to scale. Indeed, Kenneth Arrow observed that Becker had developed a theory of employment discrimination that "predicts the absence of the phenomenon it was designed to explain."[23] Becker anticipated Arrow's point by asserting that the shortage of entrepreneurial skill prevented the elimination of the discriminating employers.[24] This claim is unpersuasive, though, since very little talent was needed to see that blacks could be hired at lower cost into low-skilled industries such as textiles. Yet this never happened until Title VII took effect. Clearly, Title VII and not some newfound entrepreneurial talent explains the large gains of blacks in the decade from 1965 to 1975 (as discussed in Section 8.1 below).

Thus, while Becker's theory of employer animus generated some useful predictions, it failed to explain perhaps the key feature of racial discrimination in the South—its relentless persistence in excluding black workers from entire industries over more than half a century. Nor can the employer model explain another common characteristic of the pre-Title VII world—occupational segregation (as opposed to segregation across firms).

[22] Heckman and Payner (1989) note that blacks were largely excluded from the low-skill southern textile industry for the entire 20th century—until the effective date of Title VII of the 1964 Civil Rights Act. I analyzed the strengths of this article in Donohue (2002).

[23] Arrow (1972).

[24] Becker (1968).

3.1.3. Is the Becker model undermined by positive search costs?

Black (1995) attempts to address the empirical inadequacies of the Becker employer model of discrimination by introducing search costs into the analysis of discrimination. Black relaxes the assumptions of frictionless hiring and perfect competition to conclude that victims of discrimination may not be protected from earnings discrimination even when numerous nondiscriminating firms are present. Two assumptions about market behavior drive Black's model: workers participate in a costly search process and employers have some degree of monopsonistic power. In the population of available workers, some portion λ is male (m) and the other $1-\lambda$ is female (f).[25] Members of each group have a reservation utility U_j ($j = m, f$) that they compare to wage offers from firms, which are labeled prejudiced (p) or unprejudiced (u). Prejudiced firms, with market share θ, are known to employ only men at the wage w_p^m while unprejudiced firms with market share $1-\theta$ employ both men and women at wages w_u^m and w_u^f, respectively. Total utility for workers is the sum of wages received plus some non-pecuniary job satisfaction parameter α_k^j ($k = u, p$). This value, which for instance measures the success in matching individuals to occupation, is a random variable with cumulative distribution function $F(\alpha)$ and density $f(\alpha)$.[26] For ease of exposition, let α be distributed uniformly over the unit rectangle.

The essence of the Black (1995) model is the search process involved in matching workers with employers. When considering employment at a particular firm, a worker accepts a wage offer whenever $w_k^j \geq U_j - \alpha_k^j$. If a match does not take place, the worker incurs cost C. In equilibrium, the marginal worker will be indifferent between accepting a job and continuing search. For men, this condition yields the equation:

$$C = \theta \int_{\alpha_p^m}^{1} \left(w_p^m + \alpha_p^m - U_m\right) f(\alpha)\,\mathrm{d}\alpha + (1-\theta) \int_{\alpha_u^m}^{1} \left(w_u^m + \alpha_u^m - U_m\right) f(\alpha)\,\mathrm{d}\alpha \tag{2}$$

which, given our assumptions about α, becomes:

$$C = \theta\left[w_p^m + \frac{1+\alpha_p^m}{2} - U_m\right] + (1-\theta)\left[w_u^m + \frac{1+\alpha_u^m}{2} - U_m\right] \tag{3}$$

Equation (3) states that the cost of search must equal its expected benefits, or the net gain derived from employment at either a u or p firm, weighted by the probability of meeting that firm type. According to Equation (3), the derivative of U_m with respect

[25] Black's model, although formulated here in terms of sex discrimination, applies to any type of preference-based discrimination on the part of the employer.

[26] In order to ensure that second order conditions are satisfied, the hazard function $f(\alpha)/[1-F(\alpha)]$ must be strictly decreasing.

to w_p^m is θ and with respect to w_u^m is $(1-\theta)$, which means that reservation utility increases with wage offers but at a lower rate than the wage increase.

Women, however, can gain no utility from visits to prejudiced firms since such firms refuse to hire them. Therefore, their equilibrium condition is characterized by:

$$C = (1-\theta)\int_{\alpha_u^f}^{1}\left(w_u^f + \alpha_u^f - U_f\right) f(\alpha)\,\mathrm{d}\alpha \tag{4}$$

or

$$C = (1-\theta)\left[w_u^f + \frac{1+\alpha_u^f}{2} - U_f\right] \tag{5}$$

The comparative statics of Equation (5) again imply that reservation utility increases with the wage offer, but now on a one-for-one basis. However, an increase in the number of prejudiced firms (given by θ) reduces reservation utility. The intuition for this result is that a higher share of prejudiced firms increases search costs for women, which allows unprejudiced firms to offer a lower wage and still attract female employees. Unprejudiced firms therefore enjoy some degree of monopsonistic power due to the costly search process.

Turning to the employer's decision, let V denote the marginal product of labor, which is assumed to be the same for men and women. Then, an unprejudiced firm's expected profit per applicant is given by:

$$\pi_u^j = \left[1 - F\left(U^j - w_u^j\right)\right]\left(V - w_u^j\right) \tag{6}$$

or

$$\pi_u^j = \left(1 - U^j + w_u^j\right)\left(V - w_u^j\right) \tag{7}$$

and profit maximization entails a wage offer of:

$$w_u^j = 0.5(V + U^j - 1) \tag{8}$$

On the other hand, prejudiced firms that only hire men can expect a per applicant profit of:

$$\pi_p^m = \left(1 - U^m + w_p^m\right)\left(V - w_p^m\right) \tag{9}$$

and will offer the wage:

$$w_p^m = 0.5(V + U^m - 1) \tag{10}$$

Constant returns to scale ensure that wage offers are equalized for men across firm types, which in turn equates the "male" profits of prejudiced and unprejudiced employers. From Equations (8) and (10) we see that wages increase by one half for a unit increase in either male or female reservation utility. However, so long as there is at least one prejudiced firm, the negative effect of their presence on

women's reservation utility guarantees that $w^f < w^m$. Thus, even when workers are equally productive, the existence of any employers harboring taste-based discrimination against a minority group (or women) will result in lower earnings for the members of that group—whether or not they work for non-prejudiced employers.[27]

3.1.4. Will increased competition reduce labor market discrimination?

For both the Becker employer discrimination model and the Black search cost model, one would expect greater competition would dampen the degree of labor market discrimination. This claim has recently been invoked as part of an argument in support of globalization. Specifically, Jagdish Bhagwati's latest book *In Defense of Globalization* considers the impact of greater world economic integration on the fortunes of women by posing the question: "has globalization accentuated, or has it been corrosive of, the discriminations against women that many of us deplore and wish to destroy?"[28] Bhagwati answers this question by arguing that increased trade flows tend to narrow the male–female wage gap.

According to Becker's theory of the taste for discrimination, the decision to hire a male rather than a female of equal or greater potential productivity places the firm at a competitive disadvantage relative to its counterparts. In an autarkic world, uniform taste discrimination will not affect the home country, since all firms are making hiring decisions in the same way. However, once foreign trade is allowed, the forces of external competition reduce the viability of such prejudice since relative productivity now matters. Thus, "[t]he gender wage gap will then narrow in the industries that must compete with imports produced by unprejudiced firms elsewhere." In general, Bhagwati notes, competition, regardless of its source, will elevate the price to the firm of indulging in prejudiced behavior. As a result, "all fat [must] be removed from the firm" and the wage gap will contract.

Bhagwati cites Black and Brainerd (2002) as empirical validation of this theory. Specifically, they report that American firms that experienced an openness/competitiveness shock displayed a faster decline in their gender wage gap. Using Current Population Survey (CPS) data from 1977 through 1994, the authors try to proxy the degree of discrimination against women by computing a "gender wage gap" for 63 industry group-

[27] Duleep and Zalokar (1991) had previously observed that even if we do see equalization of gross wages, if minorities have to search harder, then the cost of discrimination shows up as lower *net* wages including search costs.

[28] Bhagwati (2004).

ings.[29] The authors then estimate the following equation:

$$\Delta(\ln(wage)_{xm} - \ln(wage)_{x\omega}) = \alpha + \beta\Delta trade_x + \gamma concen_x + \psi(trade^* concen)_x \quad (11)$$

where the dependent variable is the change in the residual gender wage gap, *trade* is the import share and *concen* is an indicator of whether in 1977 the industry was concentrated (i.e., had a four-firm concentration ratio of at least 0.4). Using the measure of the residual gender wage gap reduces confounding from other sources of variation in earnings such as education and labor market experience. Estimation of Equation (11) focuses on the coefficient on the trade-concentration interaction term, which is found, as hypothesized, to be negative and statistically significant. The authors conclude that "a 10 percentage point increase in import share in a concentrated industry would lead to a 6.6 percent decline in the residual gender wage gap."

Black and Brainerd's paper constitutes evidence that increased competition, in this case engineered by increased trade, can narrow discriminatory wage gaps.[30] I am skeptical, though, that the U.S. gender wage gap narrowed because of "imports produced by unprejudiced firms elsewhere," since many of our trading partners over this period were much less oriented towards womens' rights than we were. But note that even if employers in a country such as China are biased against women, the competitive pressure that their lower wage structure puts on U.S. firms gives American employers an incentive not to pass up lower cost but equally productive female workers. (The pressure would presumably be greater still if the Chinese employers were not discriminating against women since their costs would be even lower, but competition from low-cost producers should discipline American discriminators in any event.)

If the Black and Brainerd study is correct, then two conclusions follow: (1) there was considerable discrimination against women more than a decade after Title VII was enacted and (2) increased international trade following 1977 eliminated some portion of the discriminatory male-female wage gap differentially across the 63 industry groupings.[31] I suspect, however, that the black-white earnings gaps did *not* narrow in the

[29] Black and Brainerd (2002) describe the derivation of the gender wage gap, the within-year earnings disparity between men and women at the industry level, as follows:

> The log wage is first regressed on four categorical education variables, age, age squared, and a nonwhite dummy variable; this regression is estimated for the pooled sample of men and women in each year of interest. The residual gender wage gap is then generated as the difference in the average residual wage for men and women, calculated at the industry- or MSA-level.

[30] Query whether the Black and Brainerd results suggest that men earned rents in concentrated industries and that increased competition from trade dissipated these rents, leading to an observed shrinking of the male-female earnings disparity.

[31] Mulligan and Rubinstein (2005) conclude that the primary factors leading to the narrowing of the measured gender wage gap are that women are investing more in their market productivity and there is a positive

same fashion. If this surmise is correct, then either there was less discrimination against blacks by 1977 or Bhagwati's claim needs to be refined.

Neither the Becker model of animus-based employer discrimination nor the Black search cost model can explain the enduring exclusion of blacks in the southern labor market from entire industries, such as textiles. The following sections will discuss three rival theories to the notion of employer animus to explore whether these other types of discrimination better explain the empirical evidence of the past or present.

3.2. Customer and fellow-worker discrimination

The first alternative theory of discrimination, also originally crafted by Becker, posits that discrimination emanates not from the employer but from customers and fellow workers. This model has very different implications from the employer animus model in that the market tends to punish employer animus but clearly rewards efforts to accommodate the discriminatory preferences of customers and fellow workers. This model would seem to have at least one major advantage over the employer animus model since it can explain an enduring pattern of discrimination. Several recent papers have tested empirically for customer discrimination and found evidence of unequal treatment for minorities.[32] The most basic theoretical formulation of customer discrimination posits that buyers in the prejudiced group base their decisions on an adjusted price $p(1+\delta)$ for minority sellers, where δ again represents the discrimination coefficient. Realizing this, firm managers will attempt to assign minorities to jobs with the least amount of customer contact. In the polar case in which labor is perfectly mobile and positions within the firm are easily segregated by group characteristic, then any wage disparity between majority and minority workers of equal productivity will vanish as firms compete for the cheaper supply. When the stringent assumptions of the polar case are not met, however, the Becker model of customer discrimination can explain enduring wage shortfalls for the dispreferred group.

3.2.1. Borjas and Bronars' customer discrimination model

In contrast, the Borjas and Bronars (1989) model of self-employment and customer discrimination introduces imperfect information and search costs, which generate not only workforce segregation but also income inequality. The model assumes that individuals

selection effect operating in that, unlike in the 1970's, "women working in the 1980's and 1990's typically had better backgrounds (in terms of own schooling, cognitive test scores, and parental schooling) than did non-working women."

[32] Nardinelli and Simon (1990) discover that baseball trading cards featuring nonwhite players sell at a discount compares to that of white athletes and List (2004) uses an experimental design to uncover discriminatory bargaining offers between white and nonwhite card traders. Holzer and Ihlanfeldt (1998) find that the racial composition of a firm's clientele affects hiring decisions with respect to race (especially for positions with customer contact) as well as the wages that workers receive.

can be divided on two dimensions: (1) between white (w) and black (b), and (2) between buyers and sellers of each race. The percentage of black sellers in the population is given by γ, while that of white sellers is $1 - \gamma$. Assume also that these fractions hold for the buyer population. The mechanism through which discrimination operates is the price markup that the representative white buyer perceives to pay for the product of a black seller, denoted by δ.[33] Therefore, the maximum price that a white buyer will pay for a good produced by a black seller is $R(1 - \delta)$, where R is the consumer's valuation of the product.

The value of a price offer from a seller of race i to a buyer of race j, $V(P, i, j)$, can take one of three values. In the event that a transaction occurs, the price paid may equal the buyer's reservation price (which yields a net payoff of zero), or there may be some positive net gain $R - D(i, j)P$, where P is the seller's price offer. If the buyer rejects a proposal, then she incurs cost C and has expected valuation $EV(P, i, j)$ in the next round of search.

Sellers, on the other hand, seek to maximize the utility function $U = I - (H^{\lambda}/\lambda)$, where I is income, H represents hours worked and $\lambda > 1$. They produce goods at the rate β, conduct α transactions per unit of time and a fraction τ of those transactions result in sales. The variable τ is determined by the segregation strategy of the seller, which equals one for sales to all consumers or γ and $1 - \gamma$, respectively, for exclusive sales to blacks or whites. If production and sales cannot be performed simultaneously, then the efficient portion of the workday devoted to production is $s = \alpha\tau/(\alpha\tau + \beta)$.[34] Substituting these values into the function U and maximizing over H yields an optimal number of hours worked, which in turn generates the indirect utility function:

$$U^* = \left(\frac{1}{\sigma}\right)\left[\frac{\alpha\tau\beta}{\alpha\tau + \beta}P(\tau)\right]^{\sigma} = \frac{y}{\sigma} \tag{12}$$

where $\sigma = \lambda/(\lambda - 1)$ and y represents income consistent with utility maximizing behavior. The segregation strategy is then chosen to maximize indirect utility according to Equation (12). Finally, the model allows for differences in seller productivity with high-ability (h) and low-ability (l) individuals in each race group.

Borjas and Bronars close the model by characterizing the equilibrium distributions of prices and income for the various types of market actors. Based on their assumptions about preferences and production technologies, the equilibrium set of prices is described by the following observations:

(1) The price that sellers charge is the minimum of the reservation prices of the consumers they opt to serve.
(2) The order of reservation prices is: $P^*(w, w) \geq P^*(w, b) = P^*(b, b) \geq P^*(b, w)$.

[33] It is assumed that white consumers do not discriminate against white sellers and that black buyers are indifferent between sellers of either race.

[34] Efficiency dictates that the part of the work day spent in production, $s\beta H$, equals the portion spent conducting transactions, $\alpha\tau(1 - s)H$. Solving for s gives the equation in the text.

(3) If the market segregates by race, then sellers will be of the same race as buyers.
(4) Since high-ability sellers segregate only if their low-ability counterparts do, then the offer price distribution is ordered as: $P_{wl} \geq P_{wh} \geq P_{bl} \geq P_{bh}$.

Thus, the price schedule of white sellers constitutes a ceiling above which black sellers will never charge. Perhaps more striking is the result that high-ability sellers of both races never price above their low-ability equivalents.

These results determine the first two moments of the income distributions for white and black sellers. From the fourth condition above for prices, it is clear that even if both black and white sellers are retaining all contacts ($\tau = 1$), the latter can always charge a higher price and, on average, generate more revenue. Therefore, the mean of the white income distribution will be greater than that of the black sellers. As for the variance of the distributions, Borjas and Bronars note that higher returns to ability follow from greater variance in the distribution of income.[35] They define the variable Δ to be the ratio of the relative incomes of high-ability sellers:

$$\Delta = \frac{(y_{wh}/y_{wl})}{(y_{bh}/y_{bl})} \tag{13}$$

Whenever $\Delta > 1$, the returns to ability for whites exceed those for blacks, and Borjas and Bronars derive a set of outcomes for the variance based on different market segregation patterns.

These features of the price and income distributions indicate that incomplete information (and attendant search costs) for consumers coupled with discriminatory tastes will not only affect the size of the minority class in the market but also its quality composition. Since high ability members of the minority class face lower returns to their skill, they also have fewer incentives to engage in the prejudiced market.

3.2.2. Chiswick's employee discrimination model

In addition to customer discrimination, Becker identified fellow employees' tastes as an alternative source of biased outcomes. Chiswick (1973) developed a model of employee discrimination (which was then used to conduct an empirical analysis) as follows. Consider the case in which white workers in some location prefer not to work with non-whites. For a given amount of human capital, the wage for white employees will be:

$$Y_i = Y_i^*(1 + \delta_i X_i) \tag{14}$$

where Y_i^* is the wage of a white worker who works only with other whites, X_i equals one if he works with nonwhites and zero otherwise, and δ_i is the discrimination coefficient, which may vary with i. Furthermore, let μ_x be the mean of the X_i's, or the

[35] Borjas and Bronars specifically examine the variance of log income, a standard measure of income inequality, given in this model by $\pi(1-\pi)[\ln(y_{ih}/y_{il})]^2$. Since this variance measure increases with the ratio of high-ability to low-ability incomes, greater variance is suggestive of a higher return to ability.

proportion of white workers in integrated firms, p represent the share of nonwhites in the population, μ_δ the mean of the δ_i's and $k = \mu/p$ an index of integration.

As in the customer discrimination model, Chiswick's analysis focuses on the degree of income inequality. Taking first the natural log and then the variance of Equation (14) yields the following two equations (for small δ_i):

$$\ln(Y_i) = \ln(Y_i^*) + \delta_i X_i \tag{15}$$

$$\sigma^2(\ln Y_i) = \sigma^2(\ln Y_i^*) + \sigma^2(\delta_i X_i) + 2\mathrm{cov}(\ln Y_i^*, \delta_i X_i) \tag{16}$$

After invoking the formula for the variance of the product of random variables and applying some additional algebra, Chiswick presents the final version of the model as:

$$\sigma^2(\ln Y) = \sigma^2(\ln Y^*) + \left\{k\left(\mu_\delta^2 + \sigma^2(\delta)\right)\right\}p + \left(-\mu^2 k^2\right)p^2 + U \tag{17}$$

where U is the residual capturing the unmeasurable covariance term in Equation (16).

Inspection of Equation (17) generates several hypotheses governing the relationship between income inequality and the level of integration in the workforce. For a fixed variance of Y^*, white wage disparity increases with the nonwhite population (p).[36] Holding constant the discriminatory preferences of white workers, the more labor market integration, the greater the inequity of white earnings. Finally, inequality increases with the prevalence of discriminatory tastes (holding integration constant) and with the variance in those tastes (holding integration and mean tastes fixed). In fact, Chiswick's empirical analysis suggests that white employee discrimination against nonwhites raises average white income inequality by 2.3 percent in the entire U.S. and 3.1 percent in the South.[37] However, discrimination against whites by nonwhites was not found to be a significant factor in explaining their income inequality.

3.2.3. Did ideology influence the acceptance of the employer animus model?

Given the relative value of the various models in explaining persistent labor market discrimination, it is important to consider why the employer-animus model (henceforth "the employer model") became the dominant economic model of discrimination from the time Becker advanced it in 1959 and Milton Friedman championed it in the early 1960s. One possible answer is that ideology often trumps truth when high political stakes are involved, and both the left and right had reasons for preferring the employer model. The left found it more palatable to blame employers instead of customers and fellow workers for the ravages of discrimination, and with the enormous political battle brewing over adoption of a federal antidiscrimination law, it was better strategy to say that the law was needed to deal with bad southern employers than with the bad citizens of the South (the rest of the country had already adopted state antidiscrimination laws, so the heart of the debate was whether Congress should impose an antidiscrimination

[36] For this result to hold, Chiswick observes that μ_x must be less than ½.

[37] Chiswick (1973).

law on the South). At the same time, the right embraced the Becker employer-animus model out of dislike for antidiscrimination law and the desire to follow the lead blocking of Milton Friedman (a major opponent of the state antidiscrimination laws) in arguing that legal intervention was unnecessary since the market would effectively discipline discriminatory employers. Writing in 1963 and hoping to derail the federal efforts to adopt an antidiscrimination law, Milton Friedman clearly did not want to draw attention to the fact that the market rewards and hence encourages obeisance to the discriminatory preferences of fellow employees and customers: rather than driving out these discriminators, the market will serve to entrench those who cater to the discriminatory tastes of fellow workers and customers.

Friedman engendered much antagonism for the discipline of economics by his strident opposition to the 1964 Civil Rights Act. While his equation of this federal law to the Nazi Nuremberg laws (see the quote below in the text at footnote 67) was puzzling, and his claim in the early 1960s that blacks would be better off without the law has now been widely rejected, Friedman's opposition to an antidiscrimination law that injected government into a large realm of hitherto unregulated private contracting is consistent with his larger goal of promoting freer markets and less government. It is not implausible that he believed that in the long-run even southern blacks would be better off with this policy mixture. One sees this sentiment expressed in the work of Bhagwati (discussed above), in which he invokes the Becker employer model in stressing the benefits of competition and free trade in reducing earnings disparities by sex. The bottom line then is that market incentives can encourage discrimination in some settings and discourage it in others, and therefore, one needs to know a considerable amount about the nature of the discrimination and the institutional context before one can predict which of these outcomes will occur. In the Jim Crow South, federal law was needed to protect blacks since both the market and the local judicial/political system had failed, but that doesn't necessarily imply that federal antidiscrimination law is serving a similarly important function today. Nor can one conclude that because the market catered to the discriminatory preferences of workers and customers at an earlier time, increased market pressures today from free trade and globalization will cause more discrimination (although these economic forces might harm workers at the bottom end of the socioeconomic scale, who are disproportionately black, not because of discrimination but because of the downward pressure on the wages of low-skill workers).

In certain settings, customers have demonstrated strong discriminatory preferences. In the early days of Title VII, the courts addressed the issue of whether an employer could lawfully accommodate the discriminatory preferences of customers in a series of cases challenging airline rules that favored the hiring of young, unmarried, and attractive women to be flight attendants. The law is now well-established that such conduct constitutes unlawful employment discrimination. Clearly, the market and the law have clashed in this arena.[38]

[38] Indeed, the airline run by the restaurant chain Hooters has found a way around the prohibitions of Title VII by structuring its service of Hooters Girls on its planes as a way of selling sex appeal (since federal law allows

Fellow-worker discrimination is also a problem, but it need not undermine employment prospects (in terms of jobs and wages) if firms simply move to more segregated workforces. Of course, because current law prohibits segregation, the market and the law are in direct conflict in this area as well. But while customer and fellow-worker discrimination can explain the enduring character of racial discrimination in the South (in contrast to the employer model), neither customer nor fellow-worker discrimination can explain the pattern of exclusion from the southern textile industry. Customers would have no way of knowing, nor presumably would they care about, the race of the workers in textile plants. Although fellow-worker discrimination might explain exclusion from a particular plant, if this problem were rampant, it would, again, only lead to segregated plants—not the complete exclusion of blacks from the industry that Heckman and Payner (1989) have so thoroughly documented in the pre-Title VII world.[39]

3.3. The cartel model of discrimination

The second rival to employer animus models is the cartel model of discrimination, which was designed to explain the enduring discriminatory patterns of the Jim Crow South. While Becker premised his models of discrimination on the operation of a perfectly competitive market operating without constraint, the cartel model posits that white employers and workers managed to thwart the operation of the market by exploitatively relegating blacks to low-paying positions.[40] While Becker argued that discriminators were hurt by their discriminatory attitudes, the cartel model posits that discrimination generates supra-competitive profits for the discriminators.[41] The cartel model has a clear advantage over the Becker employer-discrimination model by better conforming with the reality that "the market" never seemed to thwart Jim Crow restrictions, which endured for decades until federal intervention finally brought them crashing down. Similarly, the cartel model conforms to the historical evidence better than the customer and fellow-worker discrimination models. Specifically, the customer discrimination model founders because discrimination was present even when customers could not identify workers, and the fellow-worker discrimination model predicts segregation

sex discrimination if it is based on a bona fide occupational qualification) rather than hiring these women for customary flight attendant jobs, for which the airline would have no right to discriminate on the basis of sex.

[39] One Chicago scholar responded to hearing of the findings of Heckman and Payner (1989) by pointing out that a powerful industry-wide union could have explained the exclusion of blacks. While the point is correct as a matter of theory, it fails as a matter of fact—such powerful unions were largely absent in the southern textile industry throughout the first sixty-five years of the 20th century.

[40] See Altonji and Blank (1999) for an excellent review of the theoretical literature on occupational exclusion.

[41] Recall from the discussion of the Becker employer model depicted in Figure 1 that the monetary profits of the discriminators *rose* even though their utility fell because of the psychic cost of having non-preferred workers. In the cartel model, the employers are able to restrict the hiring of black labor exactly as depicted in Figure 1, but in the cartel model they do so in order to increase their monetary profits without paying the psychic cost. The absence of competition enables the discriminatory employers to maintain the supra-competitive price.

(not wholesale exclusion as was observed in the southern textile industry). Libertarians, such as Richard Epstein, have also endorsed the cartel model since it has an appealing built-in remedy—simply destroy the power of the cartel, and the market will be able to operate in the protective manner that it should. Government might be needed to break up the cartel, but once competition is restored, government should revert to its more limited role of preventing the use of force and fraud.[42] In short, the libertarian argument is that no antidiscrimination law is needed once the stranglehold of the racist cartel is broken.

What is still unclear, though, is whether the "cartel" was the product of pernicious governmental restrictions and private violence, as Epstein insists, or the product of status-enhancing norms that generated utility gains for the white community, as Richard McAdams has argued.[43] While both Epstein and McAdams agree that whites clearly collaborated to help themselves at the expense of blacks, and while some of this action was enforced by law—for example, many aspects of segregation were legally mandated—there were many arenas in the South in which the law did not speak, yet the cartel was maintained. For example, there was no legal requirement that firms refuse to hire blacks, yet entire industries in the South did this for decades without any sign that the market was eroding this pattern—until the passage of the 1964 Civil Rights Act. McAdams and Epstein then stand in contrast to Becker and Friedman in arguing that federal governmental action was necessary to break up the power of the white cartel. But McAdams rejects the Epsteinian view that the cartel could only have been sustained by virtue of discriminatory governmental action, which itself was the problem. According to McAdams, legal intervention was needed to overcome discriminatory patterns that were not backed up by governmental restrictions. Whites had a vested interest in policing the informal cartel because they benefited from the increase in prestige and status, as well as the monetary benefits that were derived from subordinating blacks. McAdams marshals the social science literature showing that groups can enforce norms that enable a cartel to persist and thwart the power of those whose efforts to cheat on the cartel—for example, firms hiring cheap black labor—might eliminate the discriminatory conduct.

Interestingly, even if McAdams' hypothesis provides a better theoretical explanation for discrimination during the Jim Crow era than that offered by Becker or Friedman, there remains a question about its relevance to current American conditions. Are self-enforcing discriminatory norms still powerful enough to undermine the protective forces of competitive labor markets? Current stories of American companies establishing customer service call centers in foreign countries suggest a type of aggressive pursuit of profits that is hard to square with any pure taste for discrimination or self-enforcing discriminatory norms.[44]

[42] Epstein (1992).

[43] McAdams (1995).

[44] Query whether these global shifts only suggest that enormous profit opportunities can overcome discrimination. According to "Financial Firms Hasten Their Move to Outsourcing" (2004): in 2003 the average

3.4. Statistical discrimination

The models of statistical discrimination are the third set of rivals to the employer animus models. A central feature in these models is that unobservable attributes of workers that differ by sex, race or ethnicity prevent employers from ascertaining their true individual capabilities. Consequently, the existence of imperfect information induces employers to form hiring and wage decisions based on whatever observable information they can gather (including the worker's race or sex) as well as their prior beliefs about the expected ability of potential workers. This concept was first introduced by the models of Phelps (1972) and Arrow (1973). Since then, extensions of these theories have been highly prominent in the economic analysis of discrimination. Section 1 begins with a discussion of "statistical discrimination" in which the underlying productivities of the two groups are unequal, perhaps because of poorer schooling and lower SES (blacks) or because of expected differences in tenure (women). Section 2 then discusses two models of statistical discrimination in which the underlying productivities of the two groups are equal, yet because of the imperfection of the signal available to the employer which is more variable for the dispreferred group, earnings disparities emerge between equally productive groups of workers. The two models are the standard, static model of statistical discrimination as depicted in Aigner and Cain (1977) and the dynamic extension in Oettinger (1996), which allows for learning over time.

3.4.1. "Statistical discrimination" where group productivities are unequal

Statistical discrimination models would appear to have greater explanatory power than the Becker employer model in that they are consistent with the persistence of discrimination over long periods of time. Note, though, that while statistical discrimination as here defined is clearly prohibited under federal law, it does not necessarily constitute discrimination in the economist's sense of the term. Specifically, if on average Group A is less productive than Group B (perhaps because of poorer schooling options), then it is not discriminatory (in the economist's definition) for members of Group A to be paid less than those of Group B. Indeed, to the economist, a situation in which the earnings shortfall accurately reflects the productivity shortfall is the definition of a nondiscriminatory outcome. Federal law is clear, however, that ascribing the qualities of a racial or gender group to an individual member of that group (even if correct on average)

M.B.A. working in the financial services industry in India, where the cost of living is about 30 percent less than in the United States, earned 14 percent of his American counterpart's wages. Information technology professionals earned 13 percent, while call center workers who provide customer support and telemarketing services earned 7 percent of their American counterparts' salaries. Experts say that with China, India, the former Soviet Union and other nations embracing free trade and capitalism, there is a population 10 times that of the United States with average wage advantages of 85 percent to 95 percent.

constitutes unlawful discrimination against that individual. The greater individualized treatment in the hiring process that the law mandates will tend to help the more elite members of the group, who will not be tainted with the lower-average quality predictions that would otherwise be ascribed to them.[45] Conversely, eliminating statistical discrimination tends to harm those with the least human capital, as more individualized consideration will confirm their likely lower-than-average productivity. Thus, unless reliance on statistical discrimination impairs human capital accumulation or on the job performance—a subject addressed below—the legal attack on statistical discrimination should not be expected to improve overall welfare of any disfavored group but only to redistribute wealth from the poorest members of that group to its more affluent members.

Three main points should be made about this simple model of statistical discrimination. First, while the practice will be persistent to the extent that it is profitable to employers (presumably in helping them select good workers at low cost), there are ways for the above-average members of the group to protect themselves by signaling their high productivity to employers. Of course, without the benefit of Title VII, these workers would have to pay the costs of such signaling, which may be an argument for the legal prohibition if these signaling costs are high relative to the costs of legal enforcement. Second, there is a potential inefficiency associated with statistical discrimination in that it undermines *ex ante* incentives for investment in human capital if workers perceive that they will be treated ex post as "average" members of their group when seeking employment.[46] Unless a signal can overcome this treatment—note the distinction between a directly productive investment in human capital and a pure signal that is merely revelatory but unproductive—a worker who invests in greater human capital incurs a cost that is not fully rewarded, which will therefore result in inefficient underinvestment. Again, this inefficiency can provide a second basis for the legal prohibition of statistical discrimination. But while the risk of underinvestment in human capital as a response to statistical discrimination may have been more problematic in the past, the current empirical evidence raises doubt that this is a major concern today since returns to education are as high for blacks and other groups as they are for

[45] The divergence between the economist's conception of discrimination and the legal or popular conception of discrimination is illustrated by the manner in which the law tries to encourage employers not to consider the fact that young women are highly likely at some point to have children, which may impose costs on the employer. Indeed, if the only salient difference between male and female workers was that women needed to take time off for childbirth and subsequently spent more time in child-rearing than their male counterparts, and these traits were costly or less desirable to employers, than an economist would say that paying women identically for what is on average a less valuable contribution to an employer would be discrimination. Any attempt to pay women less on these grounds would clearly violate both the legal and popular conception of discrimination.

[46] Lundberg and Startz (1998).

whites.[47] Whether underinvestment induced by statistical discrimination has been a major problem for other protected workers, such as women or the disabled, is unclear.

The third point may be the most telling criticism of statistical discrimination when practiced against previously subordinated groups. In this situation, it may be unrealistic to assume that employer judgments will be correct on average since cognitive biases may tend to confirm the negative stereotypes that are retained from the time of subordination. Indeed, numerous studies find that individuals focus more on confirming evidence while tending to discount disconfirming evidence, which suggests that a legacy of past subordination may tend to be self-perpetuating if stereotypes tend to be reinforced.[48] Of course, competitive markets will tend to discipline firms that are subject to such cognitive biases and reward those who more accurately set the price of labor. But as Bhagwati suggested in his discussion of the gains from introducing external competition, a social consensus can develop about the attributes of previously subordinated groups that is highly resistant to change. Psychologists have developed an implicit attitudes test that reveals that most Americans strongly associate—albeit at an unconscious level—positive attributes with being white and negative attributes with being black.[49] As Kenneth Arrow has observed:

> "Suppose Blacks and Whites do in fact differ in productivity, at least on the average. This is in turn due to some cause, perhaps quality of education, perhaps cultural differences; but the cause is not itself observable. Then the experience of employers over time will cause them to use the observable characteristic, race, as a surrogate for the unobservable characteristics which in fact cause the productivity differences."[50]

At some point, the repeated association of race and lower productivity may become an embedded truth for many, as well as a factor that systematically undermines the productivity and performance of the victims of this stereotype, through what psychologists have called the "stereotype threat."[51] Indeed, once this "self-confirming" stereotype becomes entrenched, it might well lead to the type of Beckerian animus-based discrimination that was discussed above.[52] The growing literature that supports this view strengthens the case for legal prohibition of such discrimination.[53]

[47] One might still argue that, perhaps owing to lower parental investment and lower quality education, the costs of greater personal investment in human capital are higher for blacks, which might suggest that blacks would need more than equal percentage returns to elicit the optimal level of human capital investment.

[48] Loury (2002). It appears that certain beliefs are important to many as a basis of providing a sense of order or security, and these beliefs can be highly resistant to change even in the face of compelling evidence to the contrary.

[49] Greenwald, Banaji et al. (2002).

[50] Arrow (1998).

[51] Steele and Aronson (1995).

[52] Loury (2002) and Ramirez (2004).

[53] Arrow finds the model of statistical discrimination can also explain discrimination in the mortgage market because blacks default more on loans than whites. Discrimination in this market is statistical rather than

Arrow also believes that none of the economic models of discrimination can fully explain the observable patterns of behavior. For example, he finds that there is no market-based explanation for discrimination against black *consumers*, yet he considers the evidence in support of such discrimination to be compelling. Arrow also conjectures that beliefs and individual preferences may themselves be the product of social interactions unmediated by prices and markets. He focuses on the non-market network of social acquaintances and friends in the labor market that are often stratified by sex and race. Arrow emphasizes that this "network model" may be the most appropriate for the labor market, because each transaction within the employment sphere is a social event.[54] Since employment may occur by means of referral from current employees, labor segregation and discrimination can easily arise, particularly if profit maximization takes a subordinate role to maintaining a social network.

3.4.2. True discrimination—equally productive groups yet unequal pay

If employers simply generalize about individuals based on group difference in average productivity, we have seen that true discrimination is difficult to generate. Seeing every women as the average woman does hurt some women, but helps others (those below the average for women) and hence can not really explain group differences in average pay that are not based on group differences in average productivity. To explain this kind of discriminatory outcome, we need to focus on differences in the reliability of productivity-related information across groups. This approach is taken by Aigner and Cain (1977), whose model develops as follows. Consider again a labor market comprised of two groups, X and Y, and let their underlying productive ability, α, be a random variable distributed such that $\alpha \sim \mathrm{N}(\mu, \sigma_\alpha^2)$. Aigner and Cain assume that the employer does not know the value of α for a job applicant but possesses information about the distribution of ability in the overall population. Although the employer *ex ante* cannot learn α, he observes a noisy signal s (such as the score from an aptitude test), which equals the sum of true ability and a normally distributed error term with mean zero and variance σ_g^2 for $g = \mathrm{X}, \mathrm{Y}$. The crucial assumption of the model is that $\sigma_\mathrm{X}^2 > \sigma_\mathrm{Y}^2$, which means that the signal more accurately reflects latent ability for Group Y. Presuming that workers and firms are risk neutral, competition for workers bids wages up to their expected productivity conditional on the signal. Therefore, the wage is set according to:

$$w = E(\alpha|s) = \mu\left(\frac{\sigma_g^2}{\sigma_g^2 + \sigma_\alpha^2}\right) + s\left(\frac{\sigma_\alpha^2}{\sigma_g^2 + \sigma_\alpha^2}\right) \tag{18}$$

In words, the wage is a weighted average of mean ability (group characteristic) and the signal (individual characteristic) where the weights are derived from the projection

taste-based because the mortgage-lender is simply attempting to minimize risk when providing more loans to whites.

[54] See Bayer, Ross, and Topa (2005) for empirical evidence revealing a significant effect of social networks on a variety of labor market outcomes.

of α onto s. Differences in the information content of s alter the two weights but do not generate differences in the average wage of the two groups.[55]

The Oettinger (1996) model explores the simplest form of learning using a two-period horizon. Let those periods be indexed by $t = 1, 2$ and allow α and s to vary with time. As before, $\alpha_t \sim \mathrm{N}(\mu, \sigma_\alpha^2)$ and $s_t = \alpha_t + \varepsilon_t$ where $\varepsilon_t \sim \mathrm{N}(0, \sigma_g^2)$ and the variance of the error term for Group X exceeds that for Group Y. With only two periods, workers may either remain in their period 1 job during period 2 or switch to another position. Oettinger continues to assume a wage structure based on expected productivity but also allows for piece rate wages; therefore the wage in period t is $w_t = \theta\hat{\alpha}_t + (1-\theta)\alpha_t$ where $\hat{\alpha}_t$ is the conditional expectation of ability and $0 \leq \theta \leq 1$. This setup departs from the static model of statistical discrimination most notably in its assertion that productivity is match-specific. In other words, neither employee nor employer knows the former's ability in a given position before a match takes place. Thus, Oettinger remarks that for this model of learning to matter, "the arrival of this new information . . . must vary across job matches, and job mobility must be feasible."[56]

Since the firm will discover α_1 for a worker that stays in period 2, the wage schedule over time will be:

$$\begin{aligned} w_1 &= \theta\hat{\alpha}_1 + (1-\theta)\alpha_1 \\ w_2 &= \begin{cases} \alpha_1 & \text{if } \alpha_1 \geq \hat{\alpha}_2 \text{ (true for stayers)} \\ \theta\hat{\alpha}_2 + (1-\theta)\alpha_2 & \text{if } \alpha_1 < \hat{\alpha}_2 \text{ (true for movers)} \end{cases} \end{aligned} \tag{19}$$

Recall from above that the conditional expectation of α_1 under risk neutrality does not differ between Group X and Group Y; it is simply μ. Since period 1 wages are a weighted average of that conditional expectation and its true value, initial wages for X and Y are then expected to be equal.

The model then analyzes expectations of between-period wage changes and second period wages to derive hypotheses about the effects of signal extraction on wage disparities. Workers will choose to remain in their period 1 jobs if they expect a wage decrease from moving. Hence their expectation, conditional on staying, is:

$$E(\theta(\alpha_1 - \hat{\alpha}_1) | \alpha_1 - \hat{\alpha}_2 \geq 0) = \left(\frac{(1-\rho^2)\theta}{\sqrt{1+\rho^2}}\right)\left(\frac{2\sigma_\alpha^2}{\pi}\right)^{1/2} \tag{20}$$

[55] Taking the expectation of wages over s, it is evident that the average wage will be the mean level of ability μ since that is the mean of s. For this reason, Aigner and Cain take issue with the original Phelps (1972) model because differences in mean ability or signal variances do not engender average wage inequality. Thus, in order to get true discrimination, they demonstrate that employers must also be risk averse. This means that employer utility depends on signal variances, which causes the group with the less informative signal to receive a lower average wage. The model is subject to criticism, not only on the grounds that the signals are not less reliable for different groups, but also because it would seem to be unlikely that employers are sufficiently risk averse to penalize blacks to a large degree.

[56] Oettinger (1996).

where ρ is the signal to noise ratio. The implication of Equation (20) is that the expected value of staying decreases as the signal becomes more precise. Therefore, under the model's assumptions, X stayers can expect larger average wage gains than Y stayers. On the other hand, a mover faces an expected wage of:

$$E(\theta(\hat{\alpha}_2 - \hat{\alpha}_1) + (1-\theta)(\alpha_2 - \alpha_1)|\hat{\alpha}_2 - \alpha_1 > 0) = \left(\frac{(1+\rho^2) - (1-\rho^2)\theta}{\sqrt{1+\rho^2}}\right)\left(\frac{2\sigma_\alpha^2}{\pi}\right)^{1/2} \tag{21}$$

Now, enhanced signal precision (higher ρ) raises expected wages and thus predicts greater wage gains for Y movers.

In the second period, only the level of w_2 matters since the decision to move or stay has already been made. The expectations of a worker that remains at his first period firm and one that relocates are, respectively:

$$E(\alpha_1|\alpha_1 - \hat{\alpha}_2 \geq 0) = \mu + \left(\frac{1}{\sqrt{1+\rho^2}}\right)\left(\frac{2\sigma_\alpha^2}{\pi}\right)^{1/2} \tag{22}$$

$$E(\theta\hat{\alpha}_2 + (1-\theta)\alpha_2|\hat{\alpha}_2 - \alpha_1 > 0) = \mu + \left(\frac{\rho^2}{\sqrt{1+\rho^2}}\right)\left(\frac{2\sigma_\alpha^2}{\pi}\right)^{1/2} \tag{23}$$

These two equations differ solely by the factor ρ^2 in the second term on the right hand side. Since that coefficient is necessarily positive but less than one, the model suggests positive returns to job tenure. However, as ρ rises, so too does the return to job switching. Therefore, Group X should benefit more from staying in period 2, whereas Group Y profits from changing positions. Using Equations (22) and (23), one can compute the unconditional second period wage to be:

$$E(w_2) = \mu + \left(\frac{(1+\rho^2)\sigma_\alpha^2}{2\pi}\right)^{1/2} \tag{24}$$

This equation unambiguously predicts that Group Y, with a higher value of ρ, will have higher wages in the second period. Thus, "the model predicts that while no wage gap should exist at the time of labor force entry, one should develop as time in the labor force accumulates." Oettinger's intuition for this outcome is that the random draw that characterizes period 1 matching precludes any wage gap. However, more successful matches in the future lead to higher wages, and minorities (in our case, Group X) are disadvantaged because of their noisy productivity signal.[57]

The theoretical models discussed in this section are interesting in that they purport to explain how groups of equal productivity receive unequal compensation without resort to animus or bias. Nonetheless, Cain (1986, p. 729) does "not find the empirical counterparts to [these] models of statistical discrimination and signaling to be convincing."

[57] One should keep in mind that these results are predicated on the assumption that θ does not vary between the two groups. If it did, the results would be ambiguous.

Altonji and Blank (1999, p. 3190) concur: "we are unaware of any empirical work that systematically investigates the proposition that the 'signal to noise' in employer assessment of workers is lower for women than men or for blacks than whites, despite the prominence of the idea in the discrimination literature. For this reason, we are not clear how much weight should be placed on the statistical discrimination/information quality explanations for differences in group outcomes . . ."

4. Should private discrimination be prohibited?

Although the American public overwhelmingly endorses the view that labor market discrimination on the basis of race and gender should be unlawful, standard economic theory can be invoked to argue that such discrimination should not be prohibited. This argument proceeds from a basic assumption of neoclassical economics that utility maximization is an attractive principle of social welfare. As a first approximation, permitting individuals to make choices that may reflect discriminatory preferences maximizes utility. In essence, a partial equilibrium analysis invests the intersection of supply and demand curves with normative significance, and neoclassical economics tends to view discrimination as simply one more preference that shapes those curves.

Put differently, standard neoclassical economics usually begins with the assumption that, *in the absence of market failure*, there is no economic argument for government intervention into a competitive labor market. In the various taste-based models, discrimination merely reflects a personal preference of an employer, fellow employee, or customer not to associate with a certain category of individuals.[58] This model of personal preference implies that discrimination does not constitute a "market failure" in that competitive markets will still generate the most efficient allocation of resources. In terms of the partial equilibrium analysis of a labor market depicted in Figure 1, discrimination is a factor that influences the contours of the relevant supply and demand curves, but the intersection of those curves still represents the efficient solution in that any deviation from that outcome would lower welfare (as long as one honors discriminatory preferences). Similarly, under the statistical discrimination models, firms are assumed to be profit-maximizing, so again there is no market failure, which ordinarily implies that there is no efficiency argument for governmental intervention.[59]

In a competitive market, the price of a good or service should equal its value to the marginal buyer. (If not, competitive pressures will cause the price to rise or fall to restore the equality.) Thus, from the perspective of consumer sovereignty, a free market economist might even say that, in a competitive market, there can be no "discrimination" in the sense that the price paid for any good or service should equal the value

[58] Becker (1957).

[59] For an argument that antidiscrimination law can promote efficiency in a world of Becker employer discrimination by driving the discriminators from the market more rapidly, and thereby reducing the costs of discrimination, see Donohue (1986), the reply in Posner (1987) and a final rejoinder in Donohue (1987).

that the marginal purchaser places on it. Figure 1 reveals that the marginal purchaser of black labor values the marginal worker by BQ_2, which is clearly lower than AQ_1, the value of the marginal worker in the absence of discrimination. This is somewhat of a semantic point, but one would ordinarily not say that a customer who prefers the voice of Singer A to that of Singer B and thus is willing to pay more for recordings of A's music is "discriminating" against Singer B. The willingness to spend more to enjoy A's music reflects the customer's preference for that artist, and the market will ordinarily cater to that preference. In essence, the modern legal prohibition of "discrimination" posits that any preference based on race, gender, and a host of other factors is illegitimate and therefore, the goal of law is to ascertain the equilibrium intersection of the supply and demand curves—point A in Figure 1—that would exist if no economic agent had any awareness of these traits (or at least no differential valuation of them). Of course, the pure preference distinction between Singer A and Singer B does not capture the historical context in which blacks have endured a long history of oppression and subordination. This example does raise the issue of whether discrimination in the absence of such history of oppression should be treated differently from other market preferences. Would white males over 40—the primary litigants in age discrimination cases—have a strong claim for legal protection on these grounds?

The moral judgment that discriminatory preferences should not enter the social welfare calculus might be analogized to the standard philosophical argument that malicious preferences—those benefits that derive from the suffering of others—must be outside the welfare calculus. Clearly, some discrimination has been of this type in that it has been used to subordinate certain groups as a means to elevating the well-being of members of the dominant group.[60] But not all discrimination has this malicious, other-regarding character, and philosophers have had trouble justifying why non-malicious discriminatory preferences should be disregarded. Consider in this regard the words of Ronald Dworkin arguing against Catherine MacKinnon's view that pornography should be prohibited because it represents impermissible discrimination against women.[61] Dworkin argues that the "principle that considerations of equality require that some people not be free to express their tastes or preferences anywhere" is "frightening" and that if liberty and equality really conflict "we should have to choose liberty because the alternative would be the despotism of the thought-police."[62]

But in the realm of discrimination in employment, housing, and education, we do have to choose between liberty and equality (at least in the negative conception of liberty implying freedom from external restraint). As Dworkin states:

> Exactly because the moral environment in which we all live is in good part created by others ... the question of who shall have the power to help shape that environment, and how, is of fundamental importance, though it is often neglected in

[60] See the discussion of McAdams (1995) in Section 3.3, above.
[61] Dworkin (1993).
[62] Dworkin (1993).

> political theory. *Only one answer is consistent with the ideals of political equality: that no one may be prevented from influencing the shared moral environment, through his own private choices, tastes, opinions, and example, just because these tastes or opinions disgust those who have the power to shut him up or lock him up*. Of course, the ways in which anyone may exercise that influence must be limited in order to protect the security and interests of others. People may not try to mold the moral climate by intimidating women with sexual demands or by burning a cross on a black family's lawn, **or by refusing to hire women or blacks at all**, or by making their working conditions so humiliating as to be intolerable (emphasis supplied).
>
> But we cannot count, among the kinds of interests that may be protected in this way, a right not to be insulted or damaged just by the fact that others have hostile or uncongenial tastes, or that they are free to express or indulge them in private. Recognizing that right would mean denying that some people—whose tastes these are—have any right to participate in forming the moral environment at all.
>
> . . . This is an old liberal warning—as old as Voltaire—and many people have grown impatient with it. They are willing to take the chance, they say, to advance a program that seems overwhelmingly important now. Their impatience may prove fatal for that program rather than essential to it, however. If we abandon our traditional understanding of equality for a different one that allows a majority to define some people as too corrupt or offensive or radical to join in the informal moral life of the nation, we will have begun a process that ends, as it has in so many other parts of the world, in making equality something to be feared rather than celebrated, a mocking, "correct" euphemism for tyranny.[63] (Emphasis supplied.)

There is an interesting contradiction here. At first, Dworkin seems to be making a strong philosophical argument for why individuals should be allowed to exercise their private tastes and choices by discriminating against other groups even if society finds that offensive (note specifically the language in italics). But then, in the language highlighted in bold, he specifically exempts employment discrimination from his general view that liberty interests must trump equality. Note that in detailing the types of conduct that society can legitimately curtail, Dworkin includes acts of intimidation and coercion—which were traditionally prohibited as common law torts—as well as simple refusals to deal with certain groups, which have traditionally been permitted. Richard Epstein notes that "the parallel between force and discrimination has an apparent verbal seductiveness, but the differences between the two types of behavior are so profound that it is unwise to move from a condemnation of force to an equal condemnation of discrimination . . ."[64]

As Epstein notes, acts of coercion and force undermine security because even if 99 percent of the populace is not a threat, we must be concerned about the one percent

[63] Dworkin (1993).
[64] Epstein (1992).

that "bears us the most ill will." But with competitive labor markets, we need not worry about the person who most dislikes us, but can seek out and contract with those who bear us the least ill will. As long as a few employers are willing to hire members of any protected class, these workers will have options that are far more attractive than if they had to deal with the most discriminatory employer. Although violence and discrimination have often gone together, Epstein argues that, if government had to choose which it would focus on stopping first, it should choose violence.[65] Compared to discrimination, violence is easier to identify and hence sanction (with lower Type I and Type II error), and the benefits of stopping violence will be more far-reaching. This distinction is also important in evaluating the results of audit experiments that seek to uncover the proportion of employers who harbor bias against certain groups. If, for example, 20 percent of potential employers would discriminate against a particular group, this does not necessarily tell us whether the members of that group will suffer significant harm in the labor market—as long as a considerable portion of the labor market is open to them. Thus, competition dampens the "effective discrimination" experienced by those who are victims of bias—at least in terms of wage impairment, even if not in terms of the psychological harm of being rejected.[66]

While some economists stress the pragmatic point that competitive markets can reduce the need for antidiscrimination law, Milton Friedman, writing in Capitalism and Freedom, argues that such laws are not only unnecessary from a consequentialist perspective, but also defective as a matter of deontology:

> [Antidiscrimination] legislation involves the acceptance of a principle that proponents would find abhorrent in almost every other application. If it is appropriate for the state to say that individuals may not discriminate in employment because of color or race or religion, then it is equally appropriate for the state, provided a majority can be found to vote that way, to say that individuals must discriminate in employment on the basis of color, race or religion. The Hitler Nuremberg laws and the law in the Southern states imposing special disabilities upon Negroes are both examples of laws similar in principle to [antidiscrimination legislation].[67]

Note that Friedman's dubious equation of governmental prohibition and mandates of discrimination would be correct if liberty interests always trump equality interests as Dworkin argues because curtailing liberty is to be avoided regardless of whether the law

[65] Epstein urges that the notion of "rounding up" or enslaving certain groups is completely different from merely refusing to deal with the group or choosing only to deal on more favorable turns. The first is imposing harm and the second is simply failing to confer a benefit, which explains why at common law the first was unlawful and the second was not.

[66] This applies for both victims of labor market discrimination as well as victims of product market discrimination. Still, the transaction costs of having to seek out the non-discriminators are real (and in one respect tend to be exacerbated by an antidiscrimination law since employers can't advertise only for their desired candidates on racial or ethnic grounds).

[67] Friedman (1962).

promotes equality (Title VII) or is designed to stifle it (Nuremberg laws and Jim Crow). In the latter case, both liberty and equality are infringed, so the argument against the Nuremberg laws is particularly strong. One can easily imagine arguments—contrary to the views of Dworkin and Friedman—that limited curtailments of liberty could be justified by important enhancements of equality (and indeed Dworkin seems to have embraced exactly this principle with his endorsement of employment discrimination law). Still, one should be mindful of the admonitions of Dworkin, Friedman, Isaiah Berlin and others, that governments have inflicted much harm not only when deliberately curtailing freedom in order to inflict greater inequality, but also when sacrificing liberty in the name of greater equality.[68]

Ironically, we have the great liberal Dworkin arguing for what seems to be a libertarian position vis-à-vis employment discrimination (although he denies this conclusion), while the conservative Frank Easterbrook has rather forcefully articulated an argument for societal concern about the harm caused by discriminatory attitudes and behavior. In a judicial decision striking down an Indianapolis anti-pornography ordinance, Judge Easterbrook offers the following argument for why equality should trump liberty in the domain of employment discrimination (although on First Amendment grounds he then rejects the force of this argument):

> Indianapolis enacted an ordinance defining "pornography" as a practice that discriminates against women. "Pornography" is to be redressed through the administrative and judicial methods used for other discrimination . . . Indianapolis justifies the ordinance on the ground that pornography affects thoughts. Men who see women depicted as subordinate are more likely to treat them so. Pornography is an aspect of dominance. It does not persuade people so much as change them. It works by socializing, by establishing the expected and the permissible. In this view pornography is not an idea; pornography is the injury.
>
> There is much to this perspective. Beliefs are also facts. People often act in accordance with the images and patterns they find around them. People raised in a religion tend to accept the tenets of that religion, often without independent examination. People taught from birth that black people are fit only for slavery rarely rebelled against that creed; beliefs coupled with the self-interest of the masters established a social structure that inflicted great harm while enduring for centuries. Words and images act at the level of the subconscious before they persuade at the level of the conscious. Even the truth has little chance unless a statement fits within the framework of beliefs that may never have been subjected to rational study.
>
> Therefore we accept the premises of this legislation. Depictions of subordination tend to perpetuate subordination. The subordinate status of women in turn leads to affront and lower pay at work, insult and injury at home, battery and rape

[68] See Berlin (1969).

> on the streets.[69] In the language of the legislature, "[p]ornography is central in creating and maintaining sex as a basis of discrimination. Pornography is a systematic practice of exploitation and subordination based on sex which differentially harms women. The bigotry and contempt it produces, with the acts of aggression it fosters, harm women's opportunities for equality and rights [of all kinds]." Indianapolis Code Section 16-1(a)(2). *American Booksellers Ass'n, Inc. v. Hudnut*, 771 F.2d 323, C.A.7 (Ind.), 1985, at 323, 328-29.

Note that while the specific Indianapolis statute in question was aimed at pornography, the statute was deemed to be a form of antidiscrimination protection for against women and the arguments advanced in support of the legislation are frequently advanced in support of all antidiscrimination law. If one accepts the view that acts of discrimination serve to construct and buttress a powerful and subconscious framework of discriminatory beliefs, then the case for governmental intervention is greatly strengthened. In contrast, if discrimination is simply one of the infinite tastes or preferences that individuals express through their private choices and labor markets are highly competitive and governed by Beckerian notions of discrimination, then it is difficult to construct either an economic or philosophic argument for why government intervention would be needed. The two moves that those arguing for employment discrimination law have advanced are that 1) the Becker model is incorrect in that it fails adequately to capture the causes and consequences of discrimination, and 2) there is some other impediment that keeps the labor markets from operating competitively (thereby undermining confidence that competition will protect workers from exploitation). One can translate these two arguments into economic terms. The Becker model is inadequate if discrimination (1) creates important negative externalities, (2) is perpetuated by other types of market failure, or (3) leads to socially undesirable distributional outcomes, as in the various search models described above. In that event, the Beckerian vision of the competitive laissez-faire equilibrium maximizing social welfare is no longer theoretically assured even if the discriminatory preferences are fully honored.[70] The second argument posits either that discriminators are able to act as an exploitive and anticompetitive cartel in which blacks are essentially denied access to jobs, or there is some other friction that prevents the discriminators from being driven from the market at the optimal rate, as discussed in Donohue (1986). Of course, the enduring apparatus of Jim Crow in the South prior to the adoption of the 1964 Civil Rights Act would seem to support this characterization, as discussed in Section 3.3, above.

In the end, it is important to realize that both opponents and supporters of bans on discrimination have at times relied on rhetorical excess to advance their positions. Milton

[69] "... In saying that we accept the finding that pornography as the ordinance defines it leads to unhappy consequences, we mean only that there is evidence to this effect, that this evidence is consistent with much human experience, and that as judges we must accept the legislative resolution of such disputed empirical questions ...."

[70] See generally, Sen (1970).

Friedman's attempt to equate a law like Title VII to something that is now universally reviled (the Nazi Nuremberg laws) is one obvious example. A law that is designed to be an integral part of a mandated system of subordination and ultimate extermination simply cannot be equated with one that prohibits discrimination. Similarly, supporters of laws such as Title VII link the prohibition of discrimination to a battle against slavery and violence against subordinated groups, but again this link need not exist if the only practice that is being banned is the individual decision not to deal with a certain group (or not to deal in as favorable terms as those given to some other group). Presumably, acts of slavery and violence can (and of course should) be prohibited directly, regardless of whether discrimination is tolerated or banned.

Nor is every act of discrimination so obviously malicious or mean-spirited. There is considerable evidence that employers pay more for "attractive" workers.[71] According to one study: "The 9 percent of working men who are viewed as being below average or homely are penalized about 10 percent in hourly earnings. The 32 percent who are viewed as having above-average looks or even as handsome receive an earnings premium of 5 percent." The study goes on to note that the findings for women are similar although somewhat smaller: the best looking women earned 4 percent more, while the least attractive earned 5 percent less than average looking workers.[72] Another study found that obese women suffer in the labor market, but that 95 percent of their lower economic status comes from their poorer prospects in the marriage market (in terms of lower probability of marriage and a lower earning spouse if married).[73] Consequently, if race or ethnicity influences one's notion of attractiveness, it would not be surprising if some employers gravitated more to certain racial or ethnic groups in making their employment decisions. Employers also may gravitate to certain personalities, which could again be influenced by the culture of certain racial or ethnic groups. While society seems to accept preferences for attractive physical or personality traits, the distinction between such permissible preferences and impermissible discrimination when the preferences correlate with race or ethnicity is not easy to discern either conceptually or in practice.

[71] Hamermesh and Parker (2005) examine the effect of physical attractiveness on student evaluations of professors. Relying on student evaluations of 94 professors in 463 courses at the University of Texas at Austin, they found that teachers' attractiveness directly impacts the students' evaluations of their teachers: increasing attractiveness by 1 standard deviation increased the evaluation of the professor by roughly one-half a standard deviation. The authors consider, but do not resolve, the question of whether the good looks correlate with better teaching skills or effectiveness, which might provide a productivity-based explanation for the disparity. But Hamermesh (2006) shows that, in elections for officers of the American Economic Association, more attractive pictures in the election brochure increased the votes for the election candidates. Indeed, based on his examination of all 312 candidacies over the period from 1966–2004, Hamermesh finds that "a particular real-world outcome becomes more favorable for the same person when perceptions of his/her looks improve exogenously." This finding underscores that perceived attractiveness, not some underlying productivity-enhancing characteristic, influences the votes of economists in officer elections for a person that they will likely never see in person. Presumably this effect would be stronger still if the decisionmaker were choosing a person with whom he or she would be working.

[72] Hamermesh and Biddle (1994).

[73] Averett and Korenman (1996).

Nonetheless, federal antidiscrimination law embodies the judgment that society is willing to allow discrimination against "unattractive" individuals as long as the reason for the judgment of unattractiveness is not one of the precluded traits of race, sex, religion, ethnicity, disability, or age.

5. Discrimination versus disparities

Continued concern about the existence and consequences of discrimination is primarily driven by large and enduring racial/ethnic disparities in poverty and unemployment rates as well as earnings and wealth. For example, 22.7 percent of all blacks earned wages below the poverty line in 2001 as opposed to only 9.9 percent of whites.[74] For decades, black unemployment rates have typically been twice those of whites: recent data in December 2002 shows the unemployment rate was 11.5 percent for blacks and only 5.1 percent for whites.[75] For full-time workers, the median white male full-time worker had an income of $40,790 in 2001 while the median black male full-time worker only made $31,921 (see Table 1). Moreover, racial disparities in wealth are vastly greater.

Wide gaps in various employment measures also exist between male and female workers. Perhaps not surprisingly, the employment-to-population ratio for females was 56.3 percent in December 2002 but 68.8 percent for males.[76] Even for full-time workers who worked the entire year, women earned less than men: the median male full-time white worker had an income of $40,790 in 2001, while white female workers earned only $30,849.[77] The comparable numbers for black full-time, full-year workers were $31,921 for men and $27,297 for women, as shown in Table 1. Note that black men are earning more on average than white women.

Table 1 indicates that blacks (both male and female) made considerable progress in narrowing the earnings gap for full-time, full-year workers (FTFY) in the decade following the implementation of Title VII (from 1965 to 1975). During that decade, this black-white earnings gap narrowed to roughly 75 percent for men and 95 percent for women. Since 1975, the earnings growth of black women has not kept pace with that of white women, although the earnings grewth of FTFY black men have kept pace with that of white men. (Indeed, black men even narrowed the gap to 78 percent by 2001, although the weakening of the economy thereafter may have undercut this progress to some degree.)[78]

[74] See Table B-33 of The Economic Report of the President (2003).

[75] See Table B-42 of The Economic Report of the President (2003).

[76] See Table B-39 of The Economic Report of the President (2003).

[77] Part of this disparity is explained by the fact that, on average, male full-time, full-year workers work longer hours than female full-time, full-year workers.

[78] Greater rates of departure from the labor market by lower wage black men may have artificially improved these black-white earnings ratios to some degree. See Carneiro, Heckman, and Masterov (2005). Couch and Daly (2002) conclude that the black-white wage gap had indeed narrowed in the 1990s and by 1998 was the narrowest gap ever.

Two other notable points can be seen in Table 1, although significant selection effects are likely operating in both cases. First, while their declining earnings ratio suggests that Hispanics have lost out relative to whites in the 1990s, the huge influx of low-skilled, low-education Hispanics obscures this comparison. Note that the number of Hispanics in FTFY employment more than doubled between 1988 and 2001. Second,

Table 1
Median income and number of full-time, year round workers for selected groups, 1962–2001

	Males							
	1962		**1975**		**1988**		**2001**	
White	29,774		38,374		40,635		40,790	
	–		(34.0)		(42.7)		(50.0)	
Black	17,768	59.7 percent	28,553	74.4 percent	29,786	73.3 percent	31,921	78.3 percent
	–		(2.8)		(4.1)		(5.5)	
Asian	–	–	–	–	40,369	99.3 percent	42,695	104.7 percent
	–		–		(1.2)		(2.7)	
Hispanic	–	–	27,804	–	26,154	61.7 percent	25,271	58.5 percent
	–		(1.5)		(3.6)		(7.6)	
White, non-Hispanic	–	–	–	–	42,364	–	43,194	–
	–		–		(39.3)		(42.8)	
	Females							
	1962		**1975**		**1988**		**2001**	
White	17,793		22,436		27,064		30,849	
	–		(15.1)		(26.3)		(33.4)	
Black	10,858	61.0 percent	21,436	95.5 percent	24,252	89.6 percent	27,297	88.5 percent
	–		(2.0)		(4.0)		(5.9)	
Asian	–	–	–	–	28,536	105.4 percent	31,284	101.4 percent
	–		–		(0.9)		(2.0)	
Hispanic	–	–	19,073	–	21,856	79.4 percent	21,973	69.1 percent
	–		(0.6)		(2.0)		(4.4)	
White, non-Hispanic	–	–	-	–	27,521	–	31,794	–
	–		–		(24.4)		(29.3)	

Notes:

(a) Data source is the U.S. Census Bureau.
(b) Full-time workers are defined as persons on full-time schedules and include persons working 35 hours or more, persons who worked 1–34 hours for non-economic reasons (e.g., illness) and usually work full-time, and persons "with a job but not at work" who usually work full-time.
(c) Median income numbers are in 2001 dollars based on the CPI-U-RS price index of inflation.
(d) The figures in parentheses represent the number in millions of full-time, year round workers making up each group.
(e) For blacks and Asians, the percentage is the median income compared to the white median income.
(f) For Hispanics, the percentage is the median income compared to the white, non-Hispanic median income.

the table reveals that Asians have done extremely well in the labor market, but immigration may have generated the opposite selection effect from the Hispanic in-migration. Asian FTFY employment also more than doubled over the 1988–2001 period. The selection of Asian immigrants from the right tail of the skill distribution at least raises the possibility that if one controlled for education and hours worked, one might also observe an unexplained earnings shortfall for Asians. Nonetheless, it is striking that Table 1 reveals that the raw median earnings ratios for FTFY workers show Asians at or above the white levels (despite the difficulties imposed by the need to learn English for some recent immigrants).

Clearly, there are substantial disparities in earnings and other economic and social outcomes among groups, but, of course, labor market disparities can exist even in the absence of employment discrimination. Indeed, one of the greatest challenges in ascertaining the presence of employment discrimination is that *most employment discrimination cases are filed by groups that we would expect to have lower earnings even in the absence of discrimination*. For example, lower socioeconomic status correlates with lower levels of education obtained in lower quality schools and, consequently, lower levels of human capital attainment, which is an unfortunate fact of life in America for many blacks and Hispanics. While one would not expect to see women disadvantaged in schooling (at least in this country), the fact that women become pregnant and tend to assume primary care for young children imposes burdens on them that male employees less frequently shoulder. The result is that hiring female workers of a certain age predictably imposes certain higher costs on employers that are not borne if male workers are hired.[79] Obviously, individuals with physical or mental disabilities, medical conditions, or advanced age will have attributes that for many jobs will be less attractive to employers, again even in the absence of "discrimination." To the extent that the baseline for determining disparate impact is equality among groups that are not equally productive, employment discrimination law becomes a mechanism for providing preferences in the guise of enforcing an antidiscrimination mandate.[80] To the extent that the protected groups are encumbered by taste-based discrimination in addition to the productivity-based reasons for lower pay or other differential treatment, then

[79] Males of course have their own disadvantages in that they tend to be more violent and more prone to criminal conduct than women are. Age and marriage, as discussed below, seem to dampen these antisocial tendencies. Moreover, the percentage of males that engage in serious antisocial conduct is substantially lower than the percentage of women who have children, so employers may be able more effectively to identify less desirable male employees.

[80] In a disparate treatment case, a plaintiff would ordinarily buttress the claim of intentional discrimination with a simple showing that the protected group was being treated less favorably than the comparison group and that this difference was statistically significant. The same calculation would ordinarily be made in a disparate impact case, although the employer would argue that there was no disparate impact as long as the relevant rate for the protected group was at least 80 percent of the rate of the comparison group. Thus, if an employer hired 20 percent of white applicants and 17 percent of black applicants, the employer would invoke the Equal Employment Opportunity Commission's 80 percent rule to argue there was no disparate impact. See Meier, Sacks, and Zabell (1984).

the current approach of unacknowledged preferences may be helping to achieve the non-discriminatory equilibrium. If, however, the degree of preference exceeds any true economically discriminatory disadvantage, then the law is providing welfare benefits through the antidiscrimination framework.[81]

This brief discussion underscores that disparity is not a sufficient condition for the existence of discrimination. While discrimination can and has contributed to racial and ethnic inequity in earnings, the mere presence of such a disparity does not establish the presence of discrimination. For example, differences between racial or ethnic groups in basic levels of education or even earlier differences in pre-natal exposure to drugs and alcohol can lead to disparities in ultimate outcomes, even in the absence of any invidious discrimination.[82] Conversely, the absence of discrimination in one stage does not eliminate the possibility that discrimination was present and caused harm at some earlier stage. Thus, disparities observed in the labor market may not reflect discrimination in that domain, but could be generated by discrimination that occurred in a prior sphere (health, housing and education, for example). A growing literature has tried to ascertain what portion of the large disparities in economic welfare between men and women, whites and non-whites, and among other classes of protected individuals stems from discrimination and what is the product of differences in human capital, personal preferences, and ambitions.[83]

Some articles focusing on earnings disparities between gay and heterosexual workers illustrate these issues. We know that many individuals and employers discriminate against gays (the U.S. military for one[84]), although one consequence of the widespread bias is that many gays do not advertise their sexual orientation to employers. How does this bias influence the labor market outcomes of gay workers? Using simple earnings regressions, one study found that "lesbian women earn more than comparable single and married women, while gay men earn less than their married male counterparts and also perhaps somewhat less than comparable single heterosexual men."[85] Assuming that such findings are correct, how are they to be interpreted? In general, it appears that married men earn more than both straight unmarried men and gay men (the latter two earning about the same).[86] Does this show that being gay doesn't matter at all and the

[81] Providing welfare benefits through the antidiscrimination apparatus tends to target the benefits to the elite members of the protected classes rather than the neediest members of such classes. This has some obvious undesirable distributional consequences but may have offsetting external benefits if the visible advancement of the elite members both undermines stereotypic attitudes about members of the protected classes and serves a useful mentoring or role-modeling effect. For example, readily visible black economic advancement may generate positive externalities.

[82] Carneiro, Heckman, and Masterov (2005).

[83] See Blank, Dabady, and Citro (2004).

[84] "Military Loses Able Recruits with Gay Rule; ousted linguists' skills badly needed" (2003).

[85] Black et al. (2003).

[86] Allegretto and Arthur (2001) use 1990 Census data and find that gay men in unmarried partnered relationships earned 15.6 percent less than similarly qualified married heterosexual men and 2.4 percent less than similarly qualified unmarried, partnered heterosexual men.

only important stimulant to earnings for a man is being married (to a woman)? Perhaps married men are simply more attached to the labor force, so that they work a greater portion of the year and more hours while working. One might suspect that married men work harder or more reliably because they need to support a family. Alternatively, the married men may not work harder but may simply bargain harder or more effectively given the ability to reference family needs as a justification for higher earnings. Since lesbians earn more than married women, the various earnings disparities are clearly not explained by a simple story of discrimination against gays—again one suspects that marriage (at least when accompanied by child rearing) facilitates a sexual division of labor that probably leads to higher earnings for men and lower earnings for women as husbands focus more on work and wives concentrate on family matters.[87]

Thus, it may be unsurprising that lesbians earn more than married women, but why they earn more than unmarried women is potentially more puzzling (one assumes discrimination is not the explanation). Perhaps, the anticipation of the marriage effect by single women who plan to marry (and therefore feel less of a need to invest in their careers) explains why they earn less than gay women. It is also worth speculating whether being gay influences preferences in ways that might impact earnings. For example, one could examine the collegiate educational choices of gay and straight individuals to see if lesbians choose majors more often selected by straight men than straight women (e.g., more business and economics and less art and sociology), and, conversely, whether gay men choose undergraduate majors more like straight women than straight men. The bottom line is that an analysis will need many steps of elaboration before a finding of disparity can be taken as proof of discrimination.

6. Measuring the extent of discrimination

Psychologists have argued that the aforementioned implicit attitudes test reveals that most Americans harbor unconscious bias against blacks. Experimental studies in which individuals were put in situations in which they needed help reveal that both whites and blacks frequently are more likely to give aid to their own race than to the opposite race.[88] Researchers have tested for the presence of discrimination in an enormous array of settings, and typically conclude that some discrimination against blacks (and perhaps against Hispanics) is present. For example, discrimination has been uncovered in car buying, access to kidney transplants, and tipping in taxicabs.[89] In general, as we

[87] Posner (1989) argues that because women and men are economically linked (through marriage or relationships), the shortfall in earnings of women is less problematic than, say, the shortfall in earnings experienced by blacks, who do not have the same strong economic interdependency with whites. While Posner's point may lessen the sting of lower earnings for women it does not eliminate the impact on women if bargaining within the family is influenced by each partner's individual contribution to family wealth or if, in the (common) event of marital dissolution, divorce law inadequately protects women's contributions to family well-being.

[88] Crosby, Bromley, and Saxe (1980).

[89] Ayres and Siegelman (1995), Ayres (2001) and Ayres, Vars, and Zakariya (2005).

saw earlier, raw (or unadjusted) disparities across racial groups are often considerable. Nonetheless, because of the difficulties in trying to control for legitimate nondiscriminatory factors in regression analyses, most crude documentations of racial (or other) disparities probably overstate the presence of discrimination.

6.1. Regression studies

Regression analysis is frequently used to ascertain whether earnings disparities can be fully explained by various explanatory variables. These efforts to control for some of the non-discriminatory reasons for such disparities (such as lower levels of human capital attainment) tend to shrink the disparities considerably. This leaves the researcher with the nagging concern that any remaining disparities after adjustment may not reflect discrimination, but only the imprecision of the controls. For example, many studies use earnings regressions of the following form to test for discrimination:

$$\ln(earnings) = \alpha + \beta X + \delta(black = 1 \text{ or } sex = 1)$$

where X is a vector of explanatory variables including observable traits such as age and years of education. If the observable controls can capture all the factors that both influence earnings and are correlated with race or sex, then δ provides a consistent estimate of the percentage shortfall in earnings resulting from discrimination. But the observable variables are crudely measured in a way that overestimates the likely human capital of women and minorities so that the estimate of δ is likely more negative than is in fact the case. For example, age may not be a good proxy for work experience for those (such as women) who may have spent many years out of the labor market. Indeed, the enormous increase in the incarceration of blacks means that age may considerably overstate years of human capital accumulation for blacks vis-à-vis whites. Similarly, years of education may not be an ideal proxy for human capital in estimating racial disparities if whites are attending substantially higher quality schools. In both cases, the coefficient δ on the race or gender dummy may be statistically significant, but its value might drop to insignificance (or conceivably even reverse sign) if better controls for years of job experience and for quality of education could be found. Indeed, statistical tests for discrimination confront researchers with the vexing and opposing problems of omitted variable bias (where the inability to capture all of the factors that affect productivity can exaggerate the unexplained residuals in earnings functions) versus multicollinearity (where many of the included variables that proxy for productivity are highly correlated with race or sex). Many factors that could lead to greater productivity and that are correlated with race or gender may be left out of standard earnings equations. As a result, most regression studies only succeed in generating an unexplained residual in the earnings equations rather than identifying with precision the shortfall in wages caused by discrimination.[90]

[90] For example, these earnings equations rarely control for choices that workers make that may not maximize their own earnings, such as the decisions of many wives to focus more on the family than their husbands do. In

The raw comparisons of the relative earnings of black and white full-time workers are considerable (roughly 22 percent in 2001), as are the raw disparities for the other groups listed in Table 1. Not surprisingly, a portion of these disparities can be explained as regression studies control for various human capital differences. If everything can be explained by legitimate human capital factors, the case that discrimination is still harming blacks and other groups is weakened. For example, Altonji and Blank (1999) estimate racial, ethnic, and gender disparities (using 1995 CPS data) while controlling for education, potential experience, region, occupation, and industry. Their finding is that blacks suffer a 9 percentage point shortfall; Hispanics, 10 percentage points; and women, 22 percent. The big questions about these regression results for blacks and Hispanics are whether years of education can capture the difference in schooling quality for the different populations, and whether some better measure of human capital attainment is needed. For women, the big questions are whether potential experience can be an adequate control since time off for childrearing can dampen years of actual experience considerably, and whether differing female job preferences undermine the validity of the male-female regression results.

6.2. The debate over the current degree of discrimination

The competing positions in the vigorous debate concerning the significance of race and sex discrimination in the contemporary labor market were well-captured in an excellent symposium in the *Journal of Economic Perspectives* (JEP). In that exchange, William Darity and Patrick Mason aligned with Kenneth Arrow in arguing that discrimination is still widespread and significantly diminishes the economic opportunities of women and minorities, while James Heckman disagreed with this assessment. Many non-economists are highly resistant to Heckman's position because they heard this claim made by earlier University of Chicago economists at a time when it clearly was not true. The view was that the market would eliminate discriminators so discrimination can't be a substantial problem. This was analogous to arguing based on principles of aerodynamics that bumble bees can't fly, and the economics profession was probably justly given a black eye for elevating selected theory over unassailable empirical evidence. It must be remembered, though, that Heckman did not accept the Chicago orthodoxy when the empirical evidence refuted it, and he has written some of the major papers establishing the burdens of discrimination on blacks during the Jim Crow period and beyond, as discussed in Section 3.1 above and Section 8.1 below. Therefore, simply because discrimination was once a major impediment to economic advances by blacks and women does not necessarily mean that it remains so in the very different legal and economic environment that exists today. (Conversely, even if the current impact of existing labor market discrimination is small, one cannot simply assume that this would be the case if antidiscrimination laws were eliminated.) Heckman no longer believes that market

some instances, this pattern may reflect intra-family discrimination against women, rather than labor market discrimination.

discrimination substantially contributes to the black-white wage gap (as it once clearly did), and therefore he doubts that at present racial discrimination in the labor market is a first-order problem in the United States. Rather, Heckman looks to other factors (i.e., those that promote skill formation) to explain the black-white earnings gap—a theme that he builds on in Carneiro, Heckman, and Masterov (2005). The following discussion will summarize the competing evidence amassed in the JEP symposium.

Darity and Mason (1998) cite articles that estimate earnings functions using Census data that show unexplained disparities for women and minorities, which they interpret as a measure of discrimination. For the reasons discussed above, Heckman is skeptical that regression analysis of Census data is able to discern the extent of labor market discrimination against black workers.[91] Heckman also notes the disparity between the list of human capital characteristics used to measure a difference in earnings that are available in standard data sets, such as those provided by the Census, and the more complete list of characteristics available to employers when they make their employment decisions. Of course, even if current labor market discrimination is not a major impediment to blacks, discrimination in public education or housing could still be a factor, as could a set of choices by blacks that would not have been made in a non-discriminatory environment.

In arguing for the relative unimportance of labor market discrimination in affecting black economic outcomes today, Heckman (1998) instead relies on an important set of articles that correct for the problem of omitted productivity variables by adding to the earnings functions the Armed Forces Qualifying Test (AFQT) to measure an important dimension of worker quality. These articles have found that the previously unexplained earnings disparities are eliminated by the inclusion of this human capital measure. Heckman contends that the studies purporting to show the existence of racial discrimination are flawed by their failure to adequately control for underlying racial differences in human capital attainment.

Darity and Mason remain unconvinced by these articles, arguing that "the results obtained by O'Neill (1990), Maxwell (1994), Ferguson (1995), and Neal and Johnson (1996) after using the AFQT as an explanatory variable are, upon closer examination, not robust to alternative specifications and are quite difficult to interpret."[92] Specifically, Darity and Mason contend that there is a conceptual flaw in Neal and Johnson's earnings equation in that it controls for age and the AFQT but does not control for education. Both Darity and Mason as well as Lang and Manove (2004) have found that when the control for education is added to the earnings equation a black-white wage gap reemerges. The contrasting findings, then, are not in dispute but there is debate over their proper interpretation. If the AFQT measures aptitude and years of schooling measures additional productivity attributes such as acquired skill or knowledge (as well as motivation or perseverance), then both the AFQT and years of schooling should be included in the regression. In this case, the racial gap in earnings is significant. But,

[91] Heckman (1998).
[92] Darity and Mason (1998).

Heckman supports Neal and Johnson in the view that the AFQT score captures the contribution to productivity of intelligence and education and that therefore it is inappropriate to also include years of education in these earnings functions. The Neal and Johnson specification that Heckman endorses eliminates the racial gap in earnings. Ross (2003) addresses the issue as follows: "Johnson and Neal control for the age of the individual [at the time of the AFQT test], but they did not control for the education of the individual when he or she took the AFQT. While [arguably] education should not be included in the wage specification, it is certainly important to remove the influence of educational differences on AFQT performance if one is to obtain a measure of premarket ability. When this correction is made the influence of prejudice on earnings is 11 percentage points."

Similarly, Darity and Mason argue that measures of psychological well-being should be included in wage equations. They claim that their inclusion again causes the black-white wage gap to resurface. They also find that the results of the above-cited "AFQT studies" are not robust since using the math and verbal subcomponents of the AFQT leads to conflicting implications for discriminatory differentials. Given these flaws, Darity and Mason do not trust the results of studies based on the AFQT data. Even though the results may suggest that there has been a decrease in the black-white wage gap, the authors assert that blacks still suffer from discrimination in the employment market.

In addition to their claim that the aggregated regression data document the existence of race and sex discrimination, Darity and Mason argue that the evidence from selected discrimination lawsuits and audit pair studies further buttress this conclusion. They highlight the 1996 *Texaco* case as the most notorious in recent years in which top corporate officials were caught on tape making highly demeaning remarks about blacks, which then translated into discriminatory employment practices. Similar evidence was uncovered about the racist language and behavior of Ray Danner, who was the CEO of the restaurant chain Shoney's.[93]

Darity and Mason summarize the findings of five separate audit-pair studies assessing race and sex discrimination, noting:

- The Urban Institute audits from the early 1990s found that both black and Hispanic males were three times as likely to be turned down from a job as white males.
- Bendick, Jackson, and Reinoso (1994) found that whites were 10 percent more likely to receive job interviews than blacks, half of the white interviewees received job offers versus 11 percent of the black interviewees, and blacks who did receive jobs were offered 15 cents per hour less than whites.
- The Fair Employment Council found that both Hispanic and black women were three times as likely to encounter discrimination when compared to Hispanic or black males, respectively.
- To address the methodological complaints of Heckman and Siegelman (1993) that audit pairing fails to adequately hold constant all relevant traits, Neumark, Bank,

[93] See Steve Watkins, "Racism du jour at Shoney's," excerpted in Donohue (2003).

and Van Nort (1995) designed a study to eliminate personality and appearance variables by relying on manipulated resumes sent to selected employers (restaurants). The results show that a man always had a higher probability of receiving a job offer, and Darity and Mason interpret this to mean that within a particular occupation, gender discrimination is still prevalent.

- Goldin and Rouse (2000) found that hiding the identity of orchestra applicants (behind a screen) raised the probability that a female musician was selected by 50 percent.

Heckman emphasizes that the evidence from the audit studies must be evaluated in light of the distinction between market discrimination and individual discrimination. He stresses that the "impact of market discrimination is not determined by the most discriminatory participants in the market nor by the average level of discrimination among firms, but rather by the level of discrimination at the firms where ethnic minorities or women actually end up buying, working, and borrowing." That is, market discrimination occurs at the margin. While the audit studies can establish that a certain percentage of employers are discriminatory, this does not imply that there will be any effective market discrimination in an active labor market. If lots of employers refuse to hire Jews, but there are others who don't share this view, Jews may suffer no shortfall in earnings. Therefore, since Heckman concludes from the AFQT studies that blacks are receiving wages consistent with their productivity, he is skeptical of the importance of the audit study findings that some percentage of employers harbors discriminatory attitudes towards blacks.

In addition, Heckman argues that the audit pair studies may not correctly achieve even the more limited goal of identifying individual examples of discriminatory conduct. Heckman notes the following weaknesses in the audit studies:

- Audit pair studies have primarily been conducted for hiring in entry-level jobs in certain low-skill occupations using overqualified college students during summer vacations.
- Audit pair studies do not sample subsequent promotion decisions. Since only jobs found through newspapers clippings are audited, other avenues of securing work are underrepresented.

Heckman is also uncomfortable with some of the methodological assumptions that underlie the audit pair methodology. It is quite unlikely that all characteristics affecting productivity can be perfectly matched between two job candidates. If, because of the effort required to match the candidates, the researcher assumes that they have equal strength on *all* characteristics, she can mistakenly assume discrimination where there is none. For example, if the black auditors are better at Skill X but the white auditors are better at Skill Y, an audit researcher who equalizes blacks and whites only on Skill X will find discriminatory practices in firms (that are in fact looking for Skill Y workers) even when there is no discrimination.

In the end, Heckman believes that more strenuous enforcement of civil rights laws will henceforth be a costly and ineffective way to narrow the black-white wage gap. Rather, efforts should focus on enriching family and preschool environments so that

skills are strengthened before job candidates enter the market. The need for early intervention is highlighted by the recent findings of Fryer and Levitt (2005, p. 5): "By the end of third grade, even after controlling for observables, the black-white test score gap is evident in every skill tested in reading and math .... The largest racial gaps in third grade are in the skills most crucial to future academic and labor market success: multiplication and division in math, and inference, extrapolation, and evaluation in reading. Any initial optimism is drowned out by the growing gap."

6.3. Some new audit pair studies

A recent study by Devah Pager concludes that the degree of discrimination in employment is so great that blacks without criminal records are treated as badly as whites with criminal records.[94] The Pager study has been widely cited as establishing the existence of a high level of discrimination, but there are some reasons for caution in interpreting this work. This study employs an experimental audit approach, varying only criminal record, to chronicle the success of candidates' interviews in Milwaukee. Using matched pairs of individuals, the author is able to control for other characteristics and isolate the effect of the criminal record alone. Pager finds that a criminal record has a substantial effect on employment opportunities, particularly for black applicants.

Pager's audit experiment involved four male participants, two blacks and two whites, applying for entry-level job openings. The auditors formed two teams such that the members of each team were of the same race.[95] The teams applied to 15 jobs per week and the final data included 150 applications by the white pair and 200 by the black pair.[96] The auditors applied to the jobs and advanced as far as they could during the first visit. The application was considered a success only if the auditors were called back for a second interview or hired.

The results showed that 34 percent of whites with no criminal record were called back while only 17 percent of those with a criminal record were; 14 percent of blacks without a criminal record were called back while only 5 percent with a criminal record were. Notably, the black auditor without a criminal record received a smaller percentage of callbacks than the white auditor with a criminal record, suggesting the presence of substantial discrimination against blacks in general. Note that the extent of the disparity that Pager found was quite a bit higher than that found in other audit pair studies

[94] Pager (2003).

[95] The auditors were chosen based on similarity of characteristics, and all background information was made similar for the job applications. The only difference in the application was that one of the testers in each team was assigned a criminal record, a felony drug conviction, and 18 months of prison. The member of each team with the criminal record was rotated on a weekly basis to control for any unobserved differences. Both members of a team would apply for the same job, one day apart with the order determined randomly.

[96] The job openings were all within 25 miles of downtown Milwaukee and were selected from the classified section of a Milwaukee newspaper and a state-sponsored internet job service. The project occurred between June and December 2001 and focused on a range of entry-level jobs, such as restaurant workers and production workers.

in the employment realm. One issue to consider is that same-race pairs would visit the same employers but the cross-race pairs visit different employers. This is an efficient protocol for testing the impact of a criminal record on labor market success but a less efficient method for testing for race discrimination. Nonetheless, Pager notes that black and white testers were carefully matched to each other as if they were participating on the same team, so that these estimates should be unbiased even if less efficient. Moreover, while there may be some heterogeneity among the employer samples tested by each pair, even after random assignment, Pager's approach yields an offsetting advantage: the black pair and the white pair were able to use *identical* sets of resumes, which would not have been possible had they been visiting the same employers.[97] Another concern, albeit one about which Pager is well-aware, is the possibility of experimenter effects in-person audit studies. When the variables of interest, i.e. race and criminal record, are known to the auditors, there is potential for bias if the person conducting the study signals even subtly what the study hopes to accomplish.[98]

A closely related technique is used in Bertrand and Mullainathan (2004) to measure the extent of race-based labor market discrimination. Employing a so-called correspondence test methodology, they submitted about 5,000 fictitious resumes in response to nearly 1,300 employment advertisements posted in *The Boston Globe* and *The Chicago Tribune*. Their experiment was designed to estimate the racial gap in response rates, measured by phone calls or emails requesting an interview. The authors deliberately chose a correspondence test in order to circumvent some of the weaknesses associated with audit studies, such as the confounding effects of human interaction in a face-to-face interview and the difficulty of "matching" two different individuals. Randomly assigning traditionally black or white names to resumes, on the other hand, ensures (1) race remains the only component that varies for a given resume and (2) heterogeneous responses to behavior or appearance do not affect outcomes (as often occurs with human auditors).

The Bertrand and Mullainathan paper also differs from specific features of Pager's audit study. First, they analyze hiring practices for two large cities in different regions of the country. In addition, they submitted four applications to each employer, one for each race/quality cell.[99] Bertrand and Mullainathan submitted applications for three occupational categories—sales, clerical services, and administrative support—while Pager's study includes entry-level sales and clerical positions, restaurant and warehouse jobs, customer service positions, and cashiers. Finally, and perhaps most important, differences in race can only be *inferred* by the employer in the Bertrand and Mullainathan study. Since no personal contact with the potential employer ever takes place, Bertrand

[97] The resumes of test partners were similar but not identical.

[98] This is the "experimenter" effect that Heckman and Siegelman (1993) discuss in the context of the Urban Institute audit studies and that social psychologists have long recognized and stressed.

[99] Quality, which can either be "high" or "low," refers to a subjective classification of attributes across a range of standard resume components. For example, a high-quality applicant might possess (among others) an email address, computer skills, honors and volunteer or military experience.

and Mullainathan randomly assign names that are typically or exclusively associated with blacks or whites.[100] As the authors note, the correspondence test—like most audit studies—captures only the initial stage of the hiring process and excludes other important sources of employment news such as social networks. One drawback to this approach is that it can only address jobs in which mailed resumes is an appropriate application method, which may miss lower level jobs where discrimination at the point of hire may be most acute.

Bertrand and Mullainathan find significant differences in callback rates for whites and blacks: "applicants with White names need to send about 10 resumes to get one callback whereas applicants with African-American names need to send about 15 resumes."[101] Put differently, the advantage of having a distinctly white name translates into roughly eight additional years of experience in the eyes of a potential employer. Whites also appear to benefit much more than blacks from possessing the skills and attributes of a high-quality applicant and from living in a wealthier or whiter neighborhood.[102]

Although these results represent compelling evidence of labor market discrimination, it is important to bear in mind the study's underlying assumptions, particularly the likelihood that distinctive names map as expected to racial identity in the minds of potential employers. The results of Fryer and Levitt (2004) also indicate that distinctive names do not disadvantage blacks for a variety of adult outcomes. They offer some potential arguments for reconciling their findings with those of Bertrand and Mullainathan (2004). First, if names are considered a noisy initial indicator of race, then they should have no effect once a candidate arrives for the interview.[103] Second, if distinctively black names damaged labor market prospects, one might observe more name changes than appear to occur. Finally, with only about 10 percent of jobs being secured through formal resume-submission processes, the disadvantage of being screened out by certain employers may not be high when other employers and other job search paths remain open.

The combination of the audit studies and the better regression studies seems to tell us that (1) there are enough discriminators around that blacks do have to search harder to find employment, (2) the resulting unexplained earnings shortfall is not terribly high, and (3) the unexplained earnings shortfall will overstate discrimination if other legitimate factors are omitted, but will understate the cost of discrimination to blacks because

[100] Bertrand and Mullainathan express concern that employers might not recognize racial identities based on distinctive names and that such labeling may not reflect the identity of the average African-American. However, their informal survey of Chicago residents confirmed that people associate their list of distinctive names with the expected race.

[101] Bertrand and Mullainathan (2004).

[102] The difference in callback rates between high and low quality whites is 2.3 percentage points, while for blacks the difference is a meager one half of one percentage point.

[103] Fryer and Levitt also hint at the possibility that discrimination at the resume submission stage against individuals with distinctively black names will reduce the search costs of those applicants and perhaps direct them more rapidly toward employers that prefer to hire blacks. Still, this saving in search costs may come at a price if it eliminates the opportunity for high-quality black applicants to present themselves in a manner that will dampen the employer's discriminatory response.

they bear the added search costs and any attendant psychological burden that it imposes. Eliminating discrimination could bridge that earnings gap and remove the added search costs, but this would still leave a substantial unadjusted disparity in black and white earnings. Heckman is trying to emphasize that current black earnings shortfalls should be thought of as emanating more importantly from lower levels of human and cultural capital, and that efforts to address those deficits will yield greater rewards than further heightened antidiscrimination measures in the labor market. Heckman fears that efforts to aid groups that have been languishing in socio-economic attainment will be more impeded rather than advanced by a predominant focus on discrimination.

7. Antidiscrimination law in practice

Rather than a focus on ability enhancement, which Heckman would prefer, the theoretical goal of antidiscrimination law is the attainment of the equilibrium that would exist in the counterfactual world in which every individual retained his or her same abilities but the employer (or purchaser) was somehow prevented from observing any of the prohibited traits (such as race or sex). This equilibrium is given by point A in Figure 1.[104] In effect, this implies that the legislation is premised on the view that discriminatory preferences should not be registered in the social calculus and that any benefits that occur from taste-based or even statistical discrimination should be foregone. Since the antidiscrimination regime is implemented largely through private litigation, it is encumbered by all of the costs of any litigation-based scheme in which motives are highly relevant to determining liability. Thus, post hoc decision-makers must determine whether protected workers have been fired because of their protected race/gender/age/disability or for some other legitimate reason such as their shortcomings relative to other available workers. Obviously, this type of litigation is costly and prone to error.

The effort to discern the motive of employers may be particularly difficult because (as considerable psychological evidence suggests) much discrimination is unconscious. This implies that an employer might believe that he or she has not discriminated even when discrimination has occurred. The difficulty this poses for a trier of fact is clear: if the employer doesn't know that he or she has acted in a discriminatory way, how easily can the jury discern this fact? Certainly demeanor evidence at trial would be misleading if an employer who sincerely believes there has been no discrimination did in fact discriminate. The inability to readily and accurately identify intentional discrimination provides a rationale for the disparate impact doctrine and reliance on statistical proof of discrimination. Statistical models are informative about the probability that an observed disparity would occur if workers were selected in a random process. Statistically

[104] This conclusion depends on the assumption that the law is pursuing the color-blind view of discrimination. To the extent that the law is seeking to pursue another goal—such as, providing preferences for a disadvantaged group—then the demands of the law might be to generate a more favorable level of wages and employment than would exist in a wholly nondiscriminatory environment. See Donohue (1994).

significant disparities therefore suggest that the likelihood that the observed employment patterns emerged from a random process is low. Such a finding, however, does not always provide useful evidence that the non-random process was discriminatory: underlying differences in productivity may be correlated with race yet not accounted for in the statistical model. Therefore, reliance on statistical models to prove intentional discrimination will likely generate too high a level of Type I error (where the innocent employer is wrongfully found to have discriminated). Assuming reverse discrimination lawsuits are possible, the standard level of statistical significance (5 percent) would indict 5 percent of all employers, even with purely random employment selection. Not using statistical evidence, though, increases the risk of Type II error (where the unlawfully discriminating employer avoids sanction). Presumably, markets will provide some discipline on employers who engage in unconscious discrimination, so in evaluating the costs and benefits of antidiscrimination law, the imperfect market sanction needs to be compared with the imperfect legal remedy.

Antidiscrimination law may also undermine the efficient use of statistical discrimination, thereby lowering overall wealth to the extent that statistical discrimination has real cost advantages to employers seeking to minimize the cost of selecting their workforce.[105] Moreover, the prohibition on statistical discrimination can potentially turn antidiscrimination law into a mechanism for generating preferential treatment of protected workers. As we have seen, the law clearly prevents an employer from acting on the knowledge that most women will leave the labor market when they have children. If women and men were otherwise identical, then burdens of childbearing would imply that, on average, the marginal product of men would be higher than that of women. Requiring that employers ignore this fact tends to increase the demand for female workers beyond what it would be if the outcome could be reached in which all animus against women was absent. This highlights a difference between an economic and a legal definition of discrimination, since economists would say that an employer who pays a class of workers $\$x$ less because on average the members of that class impose $\$x$ greater costs on the employer is not discriminating. Indeed, the economist would likely say that in this scenario, if the employer did *not* pay less to this class of workers, then the employer would be discriminating *in favor* of this group. Thus, the legal definition would mandate economic discrimination by requiring that male and female workers must receive equal compensation and employment despite this productivity differential. Similar issues arise for racial and ethnic minorities (their relative poverty has led to less desirable school options and hence lower human capital attainment), the elderly (on average they are slowing down), and the disabled (at the very least they require reasonable accommodation).

[105] A fascinating recent paper revealed that the introduction of a personality test into the hiring process for a large retail firm did not reduce the employment of blacks even though black workers did score lower on the test. The authors conclude that "these results imply that employers were ... statistically discriminating prior to the introduction of employment testing—that is, their hiring practices already accounted for expected productivity differences between minority and non-minority applicants." See Autor and Scarborough (2004).

The previous discussion suggests an inherent tension in employment discrimination law. If, in the economist's terms, employers are appropriately paying members of a certain group less because on average the members of that group are either less productive or more costly to employ, then the legal requirement not to discriminate will be in tension with the economic incentives faced by employers. In essence, a tradeoff emerges between the equal hiring requirement and the equal wage requirement. If, as is generally believed, the latter is more binding, then the law may actually dampen employment while raising wages of those who secure employment (the "minimum wage" scenario). There is empirical support for the view that antidiscrimination laws may help those who keep jobs while reducing the total number of jobs. In general, the minimum wage effect predicts higher wages and lower employment for protected workers, while the equal hiring component suggests that protected workers will experience some demand stimulus. The bottom line is that both factors predict higher wages for protected workers but the employment effects are ambiguous depending on whether the demand stimulus offsets the incentive to cut back on more costly workers.

8. The impact of antidiscrimination law on black economic welfare

8.1. Title VII of the Civil Rights Act of 1964 and black employment

As previously noted, the major law prohibiting employment discrimination on the basis of race, sex, religion, and national origin was Title VII of the Civil Rights Act of 1964. Congress later broadened the coverage of this statute when it enacted the Equal Employment Opportunity Act (EEOA) of 1972, and then further expanded federal antidiscrimination law (primarily in providing greater damage remedies for successful sex discrimination plaintiffs and workers discharged because of their race) in passing the Civil Rights Act of 1991. The 1964 Act has received the most scholarly attention, for it was clearly the most momentous piece of antidiscrimination law ever enacted. Initially, James Smith and Finis Welch attempted to carry the mantle of Milton Friedman by arguing that the Civil Rights Act of 1964 had not advanced black economic welfare.[106] The thrust of the argument was simply that blacks had low skill levels and little education and as they secured more human capital their wages rose appropriately. Smith and Welch argued that the economic gains of blacks were no different during the period from 1940 through 1960 than they were in the following two decades. They took this as evidence against the view that Title VII generated any benefits for black workers.

More nuanced examinations of this issue have now confirmed that Title VII did indeed generate economic gains for blacks, although these gains were largely concentrated in the first ten years after adoption and in the South. As Donohue and Heckman (1991) note:

[106] Smith and Welch (1989).

> "the evidence of sustained economic advance for blacks over the period 1965–1975 is not inconsistent with the fact that the racial wage gap declined by similar amounts in the two decades following 1940 as in the two decades following 1960. The long-term picture from at least 1920–1990 has been one of black relative stagnation with the exception of two periods—that around World War II and that following the passage of the 1964 Civil Rights Act."

It is now widely accepted that in helping to break down the extreme discriminatory patterns of the Jim Crow South, Title VII did considerably increase the demand for black labor, leading to both greater levels of employment and higher wages in the decade after its adoption.[107]

8.2. The Equal Employment Opportunity Act (EEOA) of 1972

As the literature examining the effects of Title VII illustrates, attempts to estimate the impact of a federal law that has universal application at a single date in time are difficult, since any perceived changes may at least arguably be the product not of law but of broader shifts in the economy or society that either led to the legal change or just happened to coincide with it. Differential geographic impact turned out to strongly buttress the conclusion that Title VII mattered. The area of the country that had no antidiscrimination law in 1964 and that fought desperately against the passage of the 1964 Civil Rights Act was the South, and it was this region that experienced the most profound narrowing of the black-white wage gap after the federal law took effect. A recent, interesting effort addresses these issues in attempting to determine whether the EEOA, which broadened the coverage of Title VII in 1972, provided additional independent stimulus beyond that provided by the initial Civil Rights Act of 1964. Ken Chay used the fact that the EEOA had a predictably different impact across industries and between the South and the non-South as a way to estimate the economic consequences for blacks of this strengthening in the federal antidiscrimination law.[108] Prior to 1972, Title VII's prohibition against employment discrimination only applied to firms with 25 or more employees. The Equal Employment Opportunity Act (EEOA) of 1972 lowered this threshold to include employers with 15 to 24 employees. Moreover, many states already had fair employment practice (FEP) laws that covered these employers, so if the legal prohibition in these states was as effective as the federal prohibition, then the EEOA would be redundant in those states. Of the nine states that did *not* have FEP laws before 1972, eight were in the South.

Chay analyzes CPS data for the years 1968–1980 in order to assess the relative trends in black and white earnings at the two-digit industry level. Using the fraction in each industry-region employed by establishments with fewer than 25 employees (note: this is not limited to 15–24 employee establishments), Chay is able to divide the industries into

[107] Freeman et al. (1973), Donohue and Heckman (1991), Conroy (1994), and Orfield and Ashkinaze (1991).
[108] Chay (1998).

three groups for both the South and the non-South: industries with high, medium, and low fractions of workers in establishments with fewer than 25 employees. Chay's "treatment group" consists of the high fraction group (H-Group) industries in the South, since these were assumed to be the most affected by the EEOA. The low fraction (L-Group) industries are essentially considered unaffected by the EEOA and serve as the control group.

Chay estimates the share of black employment by industry, region (South or one of five non-South regions), and year while controlling for region-specific economic measures, black-white relative demographic characteristics, and a time trend.[109] The variables of interest are the post-policy effects for each region-industry group, which were defined to equal zero before March 1973 and are captured by a trend term thereafter. Chay calculates two estimates: 1) a difference-in-differences estimator comparing the post-policy changes for the South H-Group to the changes for the South L-Group; and 2) a "triple differences" estimator that compares the difference-in-differences estimate (H-Group vs. L-Group) for the South relative to the one for the non-South. Both sets of estimates indicated that the relative employment of blacks grew more after March 1973 in industries and regions with a greater proportion of small firms.[110] Chay concludes from this that the EEOA strongly increased relative black employment shares and earnings: "black employment shares grew 0.5–1.1 log points more per year and the black-white earnings gap narrowed, on average, 0.11–0.18 log points more at newly covered than at previously covered employers after the federal mandate." The evidence on the increasing relative wages of blacks is important to help exclude the possibility of white disemployment or simple black re-shuffling of employment. As a result, Chay concludes that the EEOA increased the demand for black workers among small employers not previously covered by FEP laws.

The Chay paper is persuasive, and in fact may *understate* the boost to black employment from the 1972 law for two reasons. First, Chay's control group contains some employers who had 15 to 24 employees and were not covered by a state antidiscrimination law. Thus, Chay's control group would contain some employers who shared whatever impact the EEOA had on black employment. Second, by the mid-1970s, the Supreme Court had interpreted Section 1981 of the Civil Rights Act of 1866 as providing another federal remedy for intentional discrimination without any explicit exemption for small firms. Both of these factors would lead Chay's estimates to *understate* the true impact of the law.

[109] While conducting this analysis on states rather than regions would have provided more variation and greater precision of the estimates by enabling Chay to directly control for the establishments that were already covered by pre-existing FEP laws, the small sample size of the CPS made statewide analysis impossible.

[110] Instead of separating the industries into the three groups (H-, M-, L-Group), Chay might have experimented with interacting the post-1972 trend term with the fraction of establishments in that industry that had less than 25 employees. This technique would have allowed Chay to test whether black employment share and the fraction "treated" are directly related instead of dividing the industries into somewhat arbitrary groups.

8.3. The Civil Rights Act of 1991

8.3.1. Did the CRA alter terminations of black and female workers?

In the summer of 1989, the Supreme Court cut back on a previous holding that enabled blacks to sue under §1981 for compensatory and punitive damages when discharged because of their race. While discriminatory discharges continued to be unlawful under Title VII, the 1989 decision meant that such discharged blacks were limited to remedies of reinstatement and back pay until the Civil Rights Act of 1991 restored the pre-1989 law on this issue. This Act also gave workers dismissed (or otherwise discriminated against) because of their sex the right, for the first time, to seek compensatory and punitive damages for such dismissals (although the damages that such sex discrimination cases could generate were subject to caps depending on the size of the discriminating firm's workforce). Paul Oyer and Scott Schaefer have tried to explore different effects generated by the Civil Rights Act of 1991 (henceforth "CRA") by examining whether the elevated penalties for discriminatory discharge might have altered employer behavior in predictable ways. If it is costly to fire minority and female employees because of legal restrictions such as federal and state antidiscrimination laws, then firms will have an incentive to find ways to get rid of *ex post* low-productivity protected workers that circumvent the legal prohibitions. Oyer and Schaefer (2000) suggest that one possible mechanism is to try to push such workers out the door in the course of a larger layoff, and if this strategy lowers the cost of discharge one might expect to see more firms relying on this approach as federal and state antidiscrimination laws become more stringent. To test this proposition, Oyer and Schaefer posit that the CRA would have increased employer concern about discharging minority and female workers and might have prompted the hypothesized effort to use layoffs to avoid litigation.

The major findings of the paper are that:

(a) black male full-time workers aged 21–39 were more likely to be fired than comparable non-Hispanic white men during the period from 1987 to 1991 (before the CRA of 1991 went into effect), but that this differential disappears over the period 1992–1994 (after the legislation). The paper notes, "These estimates are strongly consistent with our model's prediction that the firing rates of protected workers should go down when the potential costs of wrongful discharge litigation go up."

(b) Among the black workers who were involuntarily separated from their jobs, the proportion fired went down by more than a third after the CRA of 1991 went into effect. The suggestion is that firms were shifting away from firing blacks to terminating them during layoffs: "While the overall rate of displacement for protected workers was unaffected by the law, the share of involuntary displacements coming in the form of firings fell significantly."

Layoffs provide a great opportunity to unload dead wood of any kind (with an added advantage of getting rid of protected workers who might sue if discharged for cause). But loading up the layoff with too high a percentage of blacks might draw the attention of plaintiffs' lawyers too readily. Oyer and Schaefer note a finding that would seem to

buttress their theory of the causal impact of the CRA of 1991: they find no effect on relative firing of blacks in California but do in the rest of the country.[111] This is supportive because California had a very generous state antidiscrimination law throughout the 1987–1994 period, so one would not have expected the effective legal regime in California to be significantly impacted by the CRA of 1991. That is, if because of the application of state law, California employers were already subject to the full penalties for terminating blacks throughout the study period, then no shift in minority termination behavior should have been observed after 1991. The fact that such a shift is *not* seen in California but is seen outside California lends credence to the claim that the 1991 change in federal antidiscrimination law has influenced termination patterns.

It should also be noted that Oyer and Schaefer's before and after comparisons of the impact of the CRA are not entirely pristine because of certain judicial decisions in the pre-CRA period that were alluded to above. Federal antidiscrimination law afforded a somewhat restricted set of remedies to victims of race discrimination (damages limited to back pay and no right to jury trial) between 1965 (the effective date of Title VII of the Civil Rights Act of 1964) and 1976, when the Supreme Court ruled that a suit alleging intentional *racial* discrimination could be brought under a law passed at the end of the Civil War (Section 1981) without these restrictions. Thus in 1976, blacks could get to a federal court jury if they alleged intentional racial discrimination and sought not only back pay, but compensatory and punitive damages without limit. The Supreme Court then cut back on the sweep of the 1976 ruling in the June 1989 case of *Patterson v. McLean Credit Union*, which "held that claims of racial harassment on the job are not actionable under sec. 1981 and indicated that many promotions do not amount to the making of a new contract. Further, its decision clearly suggested that discharge for racial reasons is also outside the statute's purview."[112] The Civil Rights Act of 1991 then restored the pre-*Patterson* interpretation of Section 1981.

If the *Patterson* case had been decided before the pre-CRA data period used by Oyer and Schaefer, then their conceptual approach of defining a before/after comparison of the legal regime relevant to blacks would provide a clean test of their hypothesized effects. Instead, for the period from 1987 to mid-1989, the predominant view of the Section 1981 law concerning discriminatory discharge on grounds of race was exactly the same as the legal regime after 1991.[113] Perhaps then, a more precise test of the Oyer-Schaefer hypothesis would only compare mid-1989 to 1991 as the "before" period to post-1991 as the "after" period. The bottom line is that the before and after comparisons are likely muddied because of the way in which the law concerning race

[111] Oyer and Schaefer (2000, p. 356).

[112] Zimmer et al. (1994).

[113] Oyer and Schaefer recognize that their pre-CRA of 1991 period essentially divides into a pre-*Patterson* (pre-June 1989) period and a post-*Patterson* period. During the pre-*Patterson* period, the majority of federal courts permitted Section 1981 wrongful discharge claims, which were then extinguished when *Patterson* was decided before being restored by the CRA of 1991. If by 1987 employers fully anticipated the Supreme Court's decision in *Patterson*, then the Oyer-Schaefer pre-post comparison would be pristine.

discrimination in employment was weakened by the Supreme Court in 1989 and then restored by Congress in late 1991.

8.3.2. Did the CRA affect black and female employment levels?

In another paper, Oyer and Schaefer compare CPS data for 196 three-digit SIC code industries for two four-year periods prior to the passage of the CRA of 1991 (1983–1986 and 1988–1991) and for the period from 1993–1996 to determine if the CRA affected the employment of blacks and women.[114] The basic conclusion is that in the years leading up to the CRA of 1991, industries with relatively few women and blacks had been increasing their share of such workers (if one compares data from 1983–1986 with that from 1988–1991) but that this trend fades if one looks at data from 1993–1996. It is not all that surprising that the CRA did not enhance black employment since the only real changes it effectuated for blacks was the restoration of the law that had existed in June of 1989 with respect to discriminatory discharge and the standards for employer justification of practices with disparate racial impacts. The disparate impact standard (used to attack neutral acts that have an adverse impact on protected workers) was stringent until mid-1989, then virtually eviscerated by the *Ward's Cove* decision, and eventually restored by the CRA of 1991. Once again, though, the major difference in the law concerning racial discrimination was between mid-1989–1991 versus the end of 1991 on (when the CRA went into effect), so the Oyer-Schaefer comparison is somewhat muddied.

Moreover, to the extent that the boom of the 1990s was disproportionately driven by white and Asian males harnessing the opportunities of the internet, Oyer and Schaefer's finding that relative black employment growth slowed in the post-1991 period may be more the product of overall economic trends than the consequence of law. Note that, in any event, Oyer and Schaefer show, in their Table 2, that the percentage of blacks in overall employment was 7.8 percent for 1988–1991 as well as for the period 1993–1996, so there was no "reversal" in black employment, even if there was a slowing of gains observed across the time periods in the 1980s. For women, the small percentage decline from 39.8 to 38.9 again may be more a product of internet-driven growth in male employment than a law-driven reversal in the hiring of protected workers.[115] Thus, while I am skeptical that the CRA hurt black and female employment, I agree with one of the main themes of the Oyer and Schaefer papers that there is little support for the view that the strengthening of federal antidiscrimination law in 1991 stimulated black or female employment, as occurred with the federal laws passed in 1964 and 1972.

[114] Oyer and Schaefer (2002b).

[115] One would have expected the CRA of 1991 to have had far more impact on gender discrimination cases than on race cases (since before and after the CRA of 1991 blacks could sue for failure to hire under Section 1981 and get compensatory and punitive damages with a right to a jury trial, while women could only do this afterwards). Thus, the pattern of no decline in black employment coupled with a modest decline in female employment is at least consistent with my reading of the extent of the legal change for race and sex discrimination generated by the CRA.

Table 2
Counts of men and women in the 2002 and 2004 national scrabble tournaments

		Men	Women
2004			
	Division 1	145	25
	Division 2	78	56
	Division 3	80	88
	Division 4	61	89
	Division 5	30	59
	Division 6	26	49
	Division 7	19	21
	Total	439	387
2002			
	Division 1	113	17
	Division 2	73	29
	Division 3	54	76
	Division 4	42	82
	Division 5	40	70
	Division 6	25	61
	Total	347	335

Sources: http://www.scrabble-assoc.com/tourneys/2004/nsc/registered.html and http://www.scrabble-assoc.com/tourneys/2002/nsc/roster.html.

8.3.3. Did the CRA change the frequency of discharge complaints?

Oyer and Schaefer (2002b) present some interesting data on the frequency of EEOC complaints (in cases other than failure to hire) across two-digit SIC industries by race and gender: "in industries where women and blacks have relatively low representation, they file a relatively large number of complaints" per capita. Might this pattern imply that the CRA acted as a drag on employment of protected workers because it led to too many wrongful discharge type suits? One must consider two other possibilities. First, the possibly adverse impact on female hiring could be caused by the sharp increase in *sex harassment* (rather than wrongful termination) cases after the CRA was adopted. Second, changes in hiring patterns can also be the product of broad economic changes rather than legal developments. Specifically, as Donohue and Siegelman (1991) found, industries with lots of discharge complaints likely have large numbers of involuntary terminations.[116] Therefore, one might expect that a declining industry would experience lots of layoffs, which then lead to increased wrongful discharge claims filed by

[116] In general, tight labor markets will reduce employer-initiated terminations and will also reduce the likelihood of filing employment discrimination complaints since, under such circumstances, the market remedy of seeking another job is often preferable to the legal remedies afforded by federal law. Donohue and Siegelman (1991, 1993). History affords an interesting illustration of the claim that employers will discriminate less when labor markets are tight. During the American Revolution, George Washington countermanded the edict

women and blacks (particularly, under last hired, first fired approaches). In other words, the apparently flagging employment of women and minorities that Oyer and Schaefer note may be the product of declining industries rather than the result of an increased likelihood of discharge litigation induced by the more stringent law.

Oyer and Schaefer (2002a) also explored whether the strengthening of wrongful discharge law brought about by the CRA altered the volume of discrimination suits and had broader impacts on black and female employment. They looked at actual EEOC filings from 1988 to 1995 and limited their analysis to sex cases brought by white women and race cases brought by black males (focusing only on those aged 20–40 to avoid the complications of age-based cases). Roughly 19,000 such wrongful discharge charges were filed each year over their eight-year period. Importantly, they make two very interesting points concerning these cases brought by young white women: (1) if one looks at cases brought in a single year, the number of complaints brought per employee falls with age; so 20-year-old women are most likely to file such complaints and the number declines monotonically through age 40 (the last age in their data set); and (2) even though the age profile is the same in the years 1990 and 1993, there are substantially more cases brought in 1993 (after the adoption of the CRA). Neither of these facts (the downward sloping age-litigation profile and the jump in filings) is found for wrongful discharge cases brought by blacks. Black litigation rates (in terms of EEOC filings) actually rise from age 20–30 and then are flat or trend slightly down thereafter, and there is no obvious difference in filing rates between the two years. One can only conjecture about the reason why young women file wrongful discharge complaints at higher rates than somewhat older women. Might this reflect a harassment effect with the youngest women primarily targeted (a common pattern for harassment cases to reach the courts is that a harassed female quits and uses the harassment as the basis for a claim of constructive discharge)? Ordinarily, one would expect that older workers would be more likely to sue for wrongful discharge since the burdens of dismissal increase with tenure and increased acquisition of firm-specific human capital (which is exactly what we see for black males at least through age 30).[117]

Oyer and Schaefer note:

> The complaint rate is much higher for black men than for white women. Each year, the EEOC received a gender-based wrongful termination claim from approx-

that blacks should not be allowed to serve in the Continental Army. Washington acted not out of a sense of fairness, but out of a sense of urgency, caused by his need for more men to help fight the British. American blacks and whites would not fight side by side again until President Truman integrated the military nearly two centuries later. Ellis (2004). At the opposite extreme from Washington, the Nazis were so dogmatic in their insistence that German women should stay at home that they refused to let them work in the war effort, which greatly decreased their effective supply of domestic labor and created an obvious burden as the Germans then needed to import "huge numbers of workers from Poland and other countries under Nazi occupation." Buruma and Margalit (2002)

[117] All sex discrimination filings alleging disparate treatment would be expected to increase after the CRA owing to its authorization of compensatory and punitive damages.

imately one out of every 2500 to 3500 employed white women, but the proportion is one out of 400 to 600 for black men.[118]

The lack of growth in black male wrongful discharge EEOC filings after the CRA is not surprising since the only element relevant to such cases that changed in 1991 was the increased ability to file Section 1981 discharge cases, which litigants were not required to file with the EEOC (since they could proceed straight to federal court).

9. Discrimination on the basis of sex

During the 1980s and 1990s, the male-female wage gap decreased substantially. Darity and Mason (1998), who are generally more sanguine about the impact of federal antidiscrimination law on female employment than Oyer and Schaefer, argue that three distinct factors contributed to this important change:

- First, two opposing trends were in motion. Men at or below the 78th percentile of the wage distribution experienced absolute decreases in their real wage rate. Meanwhile, women at all points on the wage distribution experienced wage increases.
- Second, the disparity in the level of human capital for men and women was shrinking.
- Third, the level of sex discrimination was decreasing.

Clearly, the fact that women bear children and tend to assume a larger role in child-rearing than men has an important impact on female labor market decisions and outcomes. Waldfogel (1998) finds that childless women aged 24–45 receive 81.3 percent of a man's pay, whereas women of the same age with children receive only 73.4 percent. Waldfogel concludes that this pattern is caused by premarket factors that influence employment, as well as discrimination and institutional barriers in the workplace. But, of course, knowing whether and how to respond to this disparity requires an understanding of the relative importance of these factors.

Darity and Mason also contend that the index of occupational dissimilarity for both 1970 and 1990 demonstrates strong evidence of occupational crowding by gender. Although this index has decreased from 68 percent[119] to 53 percent over this twenty-year period, women are still highly concentrated in lower-paying jobs. Blau and Kahn (1996) looked at the economic performance of women in nine OECD countries and drew the following interesting conclusions: (1) in terms of human capital and occupational distribution, U.S. women compare favorably with women from the other countries; (2) the U.S. has had a longer commitment to employment equality; but (3) the U.S. gender gap in wages is larger than in any other country. Darity and Mason interpret this evidence

[118] Oyer and Schaefer (2002a).

[119] A value of 68 percent implies that 68 percent of women (or men) would have to change occupations to have equal gender representation in all occupations.

as implying that the gender wage gap is governed by the overall degree of inequality in the national economy. Since the U.S. has a high level of income inequality, the wage gap will be high despite the existence of strong antidiscrimination measures. In the case of the United States, a decentralized system for setting wages, a low minimum wage mandate, and weak trade unions account for the greater inequality in wages in the U.S. Thus, policy measures other than enhanced antidiscrimination enforcement might have a greater impact on the earnings differential between male and female workers. Of course, as John Rawls argued, we don't want to enforce greater equality at the expense of those at the low end of the income distribution.[120]

Darity and Mason's discussion is largely focused on the over-representation of women at the low end of the earnings spectrum in the labor market. What more can be said about the under-representation of women at the high end of the market? The CRA of 1991 created the Glass Ceiling Commission whose mission was to identify "those artifical barriers based on attitudinal or organizational bias that prevent qualified individuals from advancing upward in their organization into management-level positions."[121] The Commission hoped to explain phenomena such as the 90 percent male share of top managers at Fortune 500 companies. A recent survey of 120 CEOs, who were predominately male, and 705 female executives, who at the time held positions at the level of vice president and above at major corporations, illustrates the expressed beliefs of CEOs and high-ranking female executives about why so few women make it to the very top of the business pecking order.[122] The authors highlight the following survey results:

- Female executives responded that the following barriers exist: exclusion from informal networks, stereotyping, lack of mentoring, shortage of role models, commitment to personal or family affairs, lack of accountability in their position, and limited opportunities for visibility in the workplace.
- CEOs responded that the primary barriers for women workers were ineffective leadership and lack of appropriate skill sets for senior management positions.
- 79 percent of the female executives and 90 percent of the CEOs responded that the primary obstacle to gaining a top-level position is women's lack of line experience. According to the survey, women do not find themselves on the trajectory for senior management positions because they are not aware that such positions are available to them or they are discouraged from pursuing these roles by colleagues and superiors who do not feel that women can perform well in them. As a result, these women simply are not on the radar screen when succession decisions are made because they do not have the profit-and-loss experience that CEOs most value.
- Two-thirds of the female executives and more than 50 percent of the CEOs responded that a key barrier for women is the failure of senior leadership to assume accountability.

[120] Rawls (1971).
[121] See "Report on the Glass Ceiling Initiative" (1991).
[122] Wellington, Kropf, and Gerkovich (2003).

- Less than 1/3 of the total respondents considered a lack of desire by women to reach senior level positions to be a barrier to women's advancement.
- Of those executive women not already at the very top, 55 percent responded that they aspire to attain the most senior leadership positions.

The article closes with the suggestion that current CEOs must alter business strategies and human resources agendas to ensure that their female workers can gain the appropriate skill sets for senior level positions. These results helpfully describe what some highly talented individuals state is the problem, but, of course, in light of the public relations sensitivity on the part of the CEOs and the potentially self-serving responses of the female executives, one must be cautious before accepting these statements as having established the truth of the matters asserted.

Many of the survey comments suggested that women experienced disparate treatment, which would violate federal antidiscrimination law, but even this is not certain. Female executives, for example, apparently feel that they have been excluded from informal networks and were not mentored. Even if the feeling corresponds with reality, though, we still cannot conclude that disparate treatment of women had occurred unless we know that such mentoring occurred more frequently for men with no greater qualifications. Conceivably, the same percentage of men felt (and were in fact) excluded as well. Note that one of the cited "barriers" to female advancement to top managerial positions is "commitment to personal or family affairs," which would not violate current law because it is not a barrier created by employers. Arguments can be made that governmental action may be appropriate to address this situation but (1) this would be more a matter of affirmative action for women, rather than antidiscrimination law or policy, and (2) one may not want to promote policies that undermine women's "commitment to personal or family affairs."

It is unclear whether other aspects of this survey support a Beckerian notion of employer animus against having women in top jobs, or a view of statistical discrimination based on inaccurate—or even accurate, if one believes Hakim's work discussed below—views of female ability and desire for top jobs. Somewhat over half the women reported that they aspired to the highest level jobs. What was the comparable percentage for men (and can we trust the accuracy of self-reported aspirations)? In any event, one would expect that the market would penalize employer animus against women or inaccurate statistical assessments. Again, one might ask why businesses would not have the appropriate incentives to encourage this human capital development given the value of cultivating top corporate managerial talent.[123] The survey might be thought to give support for an externality-based argument for affirmative action: if women saw more top corporate female role models, then they would pursue these jobs more assiduously, thereby expanding their human capital and the productivity of business. If so, a firm

[123] Norway has just launched an experiment in affirmative action for female business executives by mandating, as of January 2006, that all publicly traded corporations must have 40 percent female representation on their corporate boards. It will be interesting to see whether this will dampen profits of the corporations, as the Becker model would suggest and as Norway's business community has strongly predicted.

might find that hiring a woman for a top job creates a positive externality by stimulating the productivity of other women that will not necessarily accrue to the original hiring firm.

9.1. Differences in male and female behavior and preferences

Other recent academic studies have suggested that the plight of women in the labor market is strongly influenced by their own conduct and attitudes existing independent of the labor market. Babcock and Laschever (2003) argue that part of the failure of women to earn as much and advance as far as men stems from the fact that modern Western culture strongly discourages women from asking and negotiating for what they want in their careers. Specifically, women directly out of an MBA program were found on average to earn $4,000 less than their male counterparts in their first jobs because men were more adept at negotiating their starting salaries. This finding appears to suggest that the requirements (or at least the goals) of the Equal Pay Act (designed to ensure that women receive the same pay as men for identical jobs) are not being met. The finding also suggests that, assuming equal productivities, employers would have an added incentive to hire women because they are willing to work for less. If there is no added incentive, then women are not underpaid from the employer's perspective, either because they impose greater costs on employers (either from Beckerian discrimination or perhaps because of the expected penalty on the employer imposed by female workers who will leave the labor market for child-bearing/child-rearing), or because their modest bargaining strategy for a higher salary correlates with lower success on the job.

Note that an employer could not defend against a wage discrimination lawsuit on the ground that women are more likely to leave the workforce for child-rearing, but might be able to prevail on the second claim if the employer made individualized determinations that particular women did not possess the attributes associated with greater productivity. As a practical matter, however, an employer would be risking substantial civil liability by attempting to justify male-female disparities in earnings or hiring on this basis, even if they were economically justified. Note, too, that if culture or biology inhibits females from negotiating aggressively, women on average will have less success in positions where this trait is rewarded. There is also some evidence that women can be trapped in a Catch-22 situation: those women who do negotiate aggressively may be characterized as "pushy or bitchy or difficult to work with," and thus rejected on this basis.[124] In the absence of employment discrimination law, the market would respond to such non-productivity-based discrimination with greater gender segregation across firms without

[124] In *Price Waterhouse v. Hopkins*, 490 U.S. 228 (1989), a talented female accountant was denied a promotion in part because her conduct was deemed aggressive and abrasive under circumstances that raised a question whether these traits would have been acceptable for male accountants. The Court ruled that given the critical remarks about the woman's dress and makeup, the burden should be on the employer to prove that sex had played no part in the decision to reject her for partnership. This burden-shifting doctrine was legislatively endorsed in the CRA of 1991.

necessarily impairing the earnings or employment of women if Beckerian, rather than search, models of discrimination are correct. Segregation of women who are highly productive but viewed as "pushy" by fellow male workers could conceivably allow the firm to profit from hiring female workers without incurring the cost of having male workers feel discomfort at working with a pushy female executive. Query whether the existence of employment discrimination law reduces the ability of firms to engage in such efficient segregation, thereby impairing the prospects of female workers (and lowering male utility).[125]

Catherine Hakim, a sociologist at the London School of Economics, uses "preference theory" to argue that, contrary to the implicit premise of antidiscrimination law, women do not have the same work aspirations as men.[126] Hakim reports that men are three times as likely as women to view themselves as 'work-centered.'[127] She contends that while women in general want opportunities, they do not want a life dominated by work. According to Hakim, antidiscrimination policy has been premised on the inaccurate belief that both men and women desire full-time employment and that spouses will take equal shares of home responsibility. Instead, many women look for spouses who can provide them with the opportunity to remove themselves from the workforce as much as possible so that they can concentrate on home life. According to Hakim, women simply have different preferences than men and most would rather spend time with their families than in the office. In fact, only one-third of those women in dual-career families even regard their jobs as central to their identity. (Query what the corresponding percentage would be for men.) Hakim's preference theory states that "women's lifestyle preferences tend to determine the pattern of their lives, and that with the benefit of equal opportunities, women continue to make choices that are different from those made by men."

Some contend that Hakim is expressing an antiquated view of female preferences, which themselves have been shaped by the discriminatory practices of the labor market. But new social science research conducted by scholars at the University of Chicago

[125] The legal prohibition on such segregation is likely quite effective because complete segregation would be an easily spotted violation of Title VII and thus would presumably be rare. Title VII would also create incentives to expand the opportunities for women, but this incentive may be less potent because the attainment of the legally mandated nondiscriminatory equilibrium is harder to secure through private litigation. The result is that the law bars the segregation that could conceivably give women higher pay and better opportunities (albeit in gender segregated firms), and forces them into integrated workforces with the attendant friction between men and women, but not so effectively that the legal protections of Title VII compensate fully for the loss of the protections of the unregulated market. The more competitive the labor market, the greater confidence one would have in the market remedies, and the less one would need the remedies supplied by law.

[126] Hakim (2000), Hakim (2003) and Kirby (2003).

[127] Through a series of questions relating to work and home/life preferences, Hakim classifies women in the United Kingdom as work-centered, home-centered, or adaptive. Work-centered women account for 15–20 percent of the population, home-centered account for 15–20 percent, and adaptive women (those whose lives encompass both work and family responsibilities) account for 60–70 percent.

business school has been offered to support the view that male workers seem to have a greater competitive drive, on average, than female workers.[128] In a set of controlled experiments involving rewards for solving a maze puzzle, the authors determine that competition between women and men tends to degrade the performance of women. The experiment, conducted in Israel, consisted of 324 engineering students over a span of 54 sessions. The authors targeted engineering students because they wanted women who were used to competing with men. The experiment consisted of five different treatments:

- *Treatment 1:* Piece Rate. Each participant was anonymously paid two shekels for each maze solved.
- *Treatment 2:* Mixed Competitive Pay. A group of three males and three females was told that the (anonymous) winner of the contest would be paid twelve shekels for each maze solved.
- *Treatment 3:* Mixed Random Pay. A group of three males and three females was told that at random, an anonymous participant would get paid twelve shekels for each maze solved.
- *Treatment 4:* Single Sex Competitive Pay. A group of six males or six females with the same setup as Treatment 2.
- *Treatment 5:* Single Sex Piece Rate. A group of six males or six females with same setup as Treatment 1.

The authors found that in either mixed or single sex piece rate tournaments (i.e., each participant receives two shekels for each puzzle solved), no significant gender difference exists. However, in the mixed tournament scheme in which only one player would win, male participants outperformed females. The increase in this gender gap is driven by the competitive performance of males under competitive pay schemes (though the performance of men does not differ between Treatments 2 and 4). When tournaments only consist of a single sex, the authors note an increase in the mean performance of women and a decrease in the gender gap in mean performance. Thus, women do in fact react to tournament incentives and compete in single sex groups. But, when women compete in a mixed group, they may have negative expectations about their relative ability that impair their performance.

In a second study focused on physical tasks, the same authors found that competition enhances the performance of boys but not of girls.[129] 140 fourth graders—75 boys and 65 girls—were tested running on a track both alone and in pairs. When children ran alone, there was no difference in performance between the boys and the girls. However, in competition, boys but not girls improved their performance.[130] The authors chose younger subjects in this experiment (compared to an average age of 23 in the maze study) to determine if competitiveness is due to socialization or other characteristics that develop at a younger age or is instead shaped by the discriminatory workplace and

[128] Gneezy, Niederle, and Rustichini (2003).

[129] Gneezy and Rustichini (2004).

[130] When girls ran with girls, their performance was worse than when they ran alone. In contrasts, boys' time improved by a large margin when they ran with another.

is therefore something that could provide a basis for a claim of unlawful employment discrimination. No monetary reward was used in this second experiment in order to determine whether males only compete for an extrinsic reward. These results confirmed the authors' hypothesis that competition has a stronger effect on boys than on girls and that the gender composition of the group of competing subjects is important. One can imagine that such evidence in an unregulated market could provide yet another incentive towards greater sex segregation in the workforce. The research also suggests that certain ways of structuring the environment might be more effective for male, rather than female, workers and that, accordingly, an employer who allowed practices to remain in place that had this effect might be the subject of a disparate impact analysis. In such a case, the employer could be found to have violated Title VII unless the employer could establish that the practice was sufficiently justified by business necessity.

In his remarks at National Bureau of Economic Research (NBER) Conference on Diversifying the Science and Engineering Workforce, Lawrence Summers, the President of Harvard, enraged some when he suggested that the relatively small number of women to reach the very top levels in the various disciplines of science might have less to do with discrimination and more to do with drive or innate ability at the extreme tails of the distribution. After cataloguing potential explanations for disparate female performance, Summers concluded that "in the special case of science and engineering, there are issues of intrinsic aptitude, and particularly of the variability of aptitude, and that those considerations are reinforced by what are in fact lesser factors involving socialization and continuing discrimination." The point is a general one—if women and men have equal mean aptitude but men have higher variance, then there will be more men at each tail of the distribution. Employment as, say, a physicist at Harvard means that someone is at the far right tail of the distribution. It has long been observed that men in general seem to have higher-variance life outcomes (men have more Nobel prizes but also more suicides, deaths due to homicide, and spells of incarceration), so the higher-variance hypothesis is worthy of consideration.

John Tierney of the New York Times used Scrabble rankings as an indication that men were willing to put in prodigious effort to reach the top of a ranking scheme at a higher rate than women, even when the number of overall Scrabble players in the country included more women than men.[131] Tierney, picking up on the work of Fatsis (2001), noted that to join the Scrabble elite, intelligence and fluency with words is not enough: "you have to spend hours a day learning words like "khat," doing computerized drills and memorizing long lists of letter combinations, called alphagrams, that can form high-scoring seven-letter words." But he then cites the work of anthropologist Helen Fisher to establish the fact that men will be much more likely to engage in such behavior because of an evolutionary predilection. Thus, "women don't get as big a reproductive payoff by reaching the top. They're just as competitive with themselves—they want to do a good job just as much as men do—but men want to be more competitive with others."

[131] "The Urge to Win" (2005).

The National Scrabble Association is the official organization for nearly 10,000 competitive Scrabble players, which supervises over 180 tournaments in the United States and Canada, including the National Scrabble Tournament held every other year. Before each official tournament, a new rating is calculated for each participant. This score, which currently ranges from 400 to 2100, is intended to serve as a relative benchmark with higher ratings indicating higher skill levels. As of June 2005, only 6 of the top 100 ranked Scrabble players are female, with the highest ranked at 45 (the others are ranked at 46, 48, 72, 89 and 100). A player's ranking simply represents their position in the national list of player ratings. The #1 player (David Gibson) has a rating of 2065 and #100 (Gail Wolford) has a rating of 1810. As Table 2 indicates, overall gender representation at the last two National Tournaments has been fairly even. But, interestingly, the premier Division 1 is dominated by male players (113 men versus 17 women in 2002; 145 men versus 25 women in 2004), while the middle divisions are more evenly matched, and women tend to outnumber men in the lower divisions. Once again, we see significant gender disparities at the most elite level of competition, even in an area involving a skill where women would not appear to be disadvantaged (and might even have an advantage). While it is unclear whether this results from some greater competitive drive or some other human capital trait, it is hard to see how discrimination could play a significant role in success in the National Scrabble Tournament.

9.2. Sex harassment

After the CRA of 1991 provided the first monetary remedy for this cause of action at the federal level, the total number of federal sex harassment cases rose sharply until 1995 and has since remained roughly stable. A number of studies have tried to estimate the prevalence of sex harassment. A 1995 survey of active duty women in the U.S. Armed Forces found that perceived sex harassment was rampant. The researchers distributed 49,003 questionnaires and collected 28,296 responses, of which 22,372 were from women.[132] The survey revealed that 70.9 percent of active duty women had faced some sort of sexual harassment over the previous year. Even adjusting for the response rate with the most conservative assumption that none of the women who did not respond had perceived sexual harassment, this is still a strikingly large perceived level of harassment, which has been corroborated by a second set of studies conducted by the U.S. Merit Systems Protections Board in 1980, 1987, and 1994. In 1994, 13,200 surveys went out to federal employees, with 8,000 returned; the results suggested that 44 percent of female employees and 19 percent of male employees had faced sexual harassment over the previous year. In 1980, the figures were 42 percent of women and 15 percent of men, while in 1987 the figures were 42 percent of women and 14 percent of men.[133] One might be tempted to interpret this time-series evidence as indicative of

[132] Antecol and Cobb-Clark (2002).
[133] USMSPB (1995).

the ineffectuality of the federal ban on sex harassment, which developed in the 1980s and was bolstered by the enhanced capacity to secure damages in the CRA of 1991. This conclusion is unwarranted, though, in that it fails to appreciate the likely defects of this time-series data. Increased sensitivity to the issue of harassment has occurred over time, so one would assume that complaints of sex harassment rose even as the incidence of sex harassment declined.

Grafting the prohibition on sex harassment onto the antidiscrimination regime has the benefit of sanctioning clearly undesirable conduct but, of course, comes at a price. First, as this paper has stressed, litigation-based enforcement schemes are costly and subject to Type I and Type II errors. The social loss from high Type II errors (failing to punish actual harassers) is mitigated to the extent that the costly litigation does put at least some burden on wrongdoers. Nonetheless, Type II errors in sex harassment cases likely impose a considerable psychic if not monetary burden on victims—the monetary burdens of the unsuccessful suits fall on the plaintiff's attorneys who typically get paid only when they win. Of course, without the legal prohibition, all wrongdoers go free. High Type I errors impose all of the same litigation costs but wrongfully sanction innocent conduct, which can have an inhibiting effect on unobjectionable workplace conduct (as workers try to avoid anything that might be misconstrued as harassment, presumably reducing both some unpleasant, albeit non-harassing conduct but perhaps also reducing some pleasant and desired conduct). Moreover, if hiring a woman has some chance of imposing an erroneous large monetary penalty plus the stigma of sex harassment liability, that prospect will serve as another burden associated with hiring American workers in general and women in particular.

Second, there is the doctrinal issue of whether the sex harassment claim should be an independent tort or linked to antidiscrimination law where it does not always fit comfortably. Thus, we see an increasing number of sex harassment claims brought by men, many of which are same-sex harassment cases where the reason for the harassment may stem more from sexual orientation than from gender. Moreover, the sex discrimination framework fits uncomfortably when a boss harasses both male and female employees, which is not unknown since some harassers harass anyone over whom they can exert power.

10. Discrimination in credit and consumer markets

10.1. Housing and credit markets

As noted previously, Congress enacted a number of statutes in the late 1960s and early 1970s extending the reach of antidiscrimination law beyond employment, public accommodations, and schooling. The Fair Housing Act (FHA), passed in 1968, prohibits housing providers and lending institutions from discriminating against consumers based on race, religion, sex, national origin, familial status or disability. The Equal Credit Opportunity Act (ECOA) of 1974 made it illegal, inter alia, for the extension of credit

to be influenced by the racial composition of a neighborhood, and the Home Mortgage Disclosure Act (HMDA) of 1975—later amended in 1989—mandates that lenders report information on their lending activity and the disposition of individual applications. The HMDA has generated voluminous data that has been mined by researchers seeking—and, according to Kenneth Arrow, finding—evidence of discrimination in lending practices.

HMDA data has been subjected to regression analysis designed to detect disparate treatment by showing that being a member of a protected class significantly reduces the probability of obtaining fair terms of trade after controlling for legitimate measures of creditworthiness, such as income, credit history, and existing debt.[134] At the same time, audit studies have attempted to reveal discriminatory business practices in housing and lending *as they occur*.

Yinger (1998) and Heckman (1998) stress that both standard regression and audit studies have strengths and weaknesses. Either omitting necessary explanatory variables or including "illegitimate" controls can influence the ultimate findings of regression studies concerning the presence or absence of discrimination. However, as mentioned above, audit studies can also be marred by errors in design and management. For example, the decision to inform the auditors about the study's objectives or about the presence of his or her partner may influence their behavior and survey responses in ways that are likely to support a finding of discrimination if one assumes that test auditors would likely sympathize with the goals of the antidiscrimination organizations that usually initiate audit tests. Moreover, audit studies are typically narrower in focus than regression analysis; they highlight discrimination in isolated stages of economic transactions rather than reveal the experience of the average member of a protected class who may learn to find more reliable trading partners in active, competitive consumer markets.[135] Also, with the partial exception of the housing context where repeated studies have been undertaken, audit studies are not generally available in time series, which limits their usefulness in analyzing changes over time. As a result, inference and interpretation based on either type of study requires explicit consideration of their competing advantages and disadvantages.

Schafer and Ladd (1981) used data on mortgage applications in California during 1977–1978 and in New York from 1976–1978 to estimate the differential probabilities of loan denial by race, sex and marital status. Controlling for an array of variables, including loan-to-value ratio, income of secondary earners and neighborhood effects, Schafer and Ladd estimated that black applicants were anywhere from 1.58 to 7.82 times as likely to be denied loans as white applicants. Interestingly, they found that the disparate treatment of women subsided over time, whereas for minorities the trend seemed to persist. After the 1989 expansion of the HMDA, the Federal Reserve Bank

[134] As Yinger (1998) underscores, the economic status of credit applicants and consumers may itself be the legacy of previous discrimination.

[135] Yinger (1998). Ross (2003) also notes that audit studies cannot reveal disparate impact discrimination, presumably because all of the neutral factors that distinguish the sets of testers have been held equal.

of Boston analyzed newly available data containing all the components of a lender's information set at the time of the loan decision.[136] The resulting study—published as Munnell et al. (1996)—found that even the rich set of controls could not fully explain the differential treatment experienced by blacks and whites. The paper concluded that blacks experienced a denial rate that was almost twice as high as that for similarly situated whites.

A number of criticisms have been leveled against the Boston Fed finding of discrimination in the market for mortgages. Some (incorrectly, according to Ross and Yinger, 1999) argued that coding errors could account for the results, while others, such as Stengel and Glennon (1994), attacked the Boston Fed's model specification. The debate has continued with Kenneth Arrow concluding that *statistical* discrimination was clearly present and Bostic (1996) and others continuing to argue against findings of discrimination. Specifically, Becker (1993) stated: "Some of the evidence found by the Boston Fed contradicts their claim of discrimination against minorities. For example, average default rates found in this study were about the same on loans in census tracts with a large percentage of blacks and Hispanics as in predominantly white tracts. Yet if the banks had been discriminating against minority applicants, default rates on loans to minorities should have been lower than on loans to whites, since banks discriminate in part by accepting minority applicants only with exceptionally good credit histories and employment records. They would reject marginally qualified minority applicants while accepting marginal white applicants."

Berkovic et al. (1996) and others have tried to follow Becker's suggestion by further examining the rate of loan default by race in order to detect or disprove discrimination in lending behavior. As Becker noted, taste-based discrimination would lead institutions to set higher credit thresholds for minorities, thereby decreasing their probability of default relative to white borrowers. Results that point to higher minority default rates have therefore been interpreted as evidence against discrimination. As Ladd (1998) cautions, however, the use of default data is subject to important methodological limitations. She argues that, unlike the loan application data, which include the full set of factors used by the lender when deciding to approve or deny, default data necessarily omit unobserved factors that contribute to the probability of default. Such unobserved heterogeneity, which can influence the probability of default in both directions, has made it difficult to generate an unassailable conclusion about the existence or nonexistence of discrimination from default data.

Using data from the 1989 Housing Discrimination Study, Yinger (1995) probes the severity of discrimination by examining the rate at which members of racial groups learn about housing opportunities through market interaction. He finds, consistent with discrimination, that "black home buyers learn about 23.7 percent fewer houses than do their white teammates, [and] black renters learn about 24.5 percent fewer apartments . . ."[137] These results imply that in addition to the psychic costs of discrimination blacks

[136] This dataset contained crucial information on the credit history, employment stability and public record of defaults of applicants, all of which were missing from Schafer and Ladd (1981) and previous studies.

[137] Yinger (1998).

suffer, they are also burdened by higher search costs and the consequent potentially inferior housing.

A recent study by Han (2002) that Ross (2003) references may reconcile the apparently contending positions in the issue of mortgage lending discrimination. Reanalyzing the Boston Fed data, Han shows that there are distinctly different patterns between applicants who have a credit history and those who do not. In the former case, lenders seem to treat applicants equally across races since they have valid information on which to make their decisions. In the case where the applicants have no credit history, however, strong racial differences are found. Han concludes that since blacks in the first category had significantly worse credit histories than did whites, lenders make similar assumptions about applicants without credit histories and therefore assume blacks are worse credit risks and treat them accordingly. This is precisely what Arrow concluded—that there is statistical discrimination against blacks. At the same time, if the statistical judgments are correct on average then the lending firms are not making *greater* profits on the loans that they do make to black applicants—which is Becker's point. In essence, Becker is emphasizing his belief that there is no taste-based discrimination against black mortgage applicants, while Arrow is emphasizing that there is still statistical discrimination against such applicants (although Han would suggest, only at the point where richer credit information is not yet available to the lenders). Of course, Arrow would be correct in noting that such statistical discrimination against black applicants would be unlawful. Becker would likely reply that such conduct shouldn't be banned since on average blacks are being treated fairly (and that credit is being allocated and priced more efficiently with such statistical discrimination than it would be without it).

Ross and Yinger (2002, p. 310) argue that lenders who appear to be following a legitimate lending model based on neutral lending criteria that do not include applicant race can still disadvantage black applicants considerably. This can occur if these lending schemes:

> "exploit the correlation between many credit characteristics and minority status to create underwriting weights that serve to help identify minority applicants, not just to measure the impact of credit characteristics on loan performance. We show that the only way to rule out disparate-impact discrimination is to make sure that every element of a scoring scheme improves the ability of the scheme to predict the performance of the applicants within a group (among whites, for example). More research is needed to determine whether the elements of existing scoring schemes meet this test, but we use existing default data to show that disparate-impact discrimination generated by these schemes could severely limit minority households' access to credit under some circumstances."

10.2. Auto sales

Ayres (1991, 1995) and Ayres and Siegelman (1995) have also used the audit approach to document the presence of discriminatory pricing in automobile sales. Carefully con-

trolling for observable differences between audit pairs and instructing auditors on precise bargaining tactics, Ayres and Siegelman collected data from 306 cases at Chicago car dealers. They found that black, male customers paid approximately $1,000 more for cars than white men and black females paid $405 more than white males. Additional results from these car sales audit studies suggest that discriminatory practice does not depend on the race of firm employees and that car dealers statistically discriminate by assuming that black men and all women have higher reservation prices than white males.[138]

Goldberg (1996) uses regression analysis of Consumer Expenditure Survey data from 1983 to 1987 to argue against the claim of discrimination in auto sales prices by arguing that car dealers did *not* significantly reduce the price of cars below list value for white males relative to minorities or women. Goldberg's sample of nearly 1,300 households included less than 5 percent minority males or females, which is probably a smaller amount of data than one would ideally like to have in resolving such an important question. Moreover, Goldberg's paper is not necessarily in direct conflict with the findings of Ayres and Siegelman because of their different geographic focus (national versus Chicago) and units of observation (households versus individuals). Finally, as Siegelman (1998) notes,

> "Even though Goldberg (1996, 624) characterized her results as 'quite different from the ones reported by ... Ayres and Siegelman,' ... Goldberg's estimates of the discriminatory premiums paid by white females and 'minority' females are virtually identical to ours. The only difference ... is that Goldberg found black males paying a much smaller premium than we did, and none of her results are statistically significant, whereas ours were, at least for the black testers. Because there are at least six dimensions on which our audit data allowed for more precise measurement and better controls than the survey Goldberg used, her failure to obtain statistically significant results is not surprising and should not be taken as evidence against the existence of discrimination in new car sales."

11. Criminal justice and racial profiling

As crime fell starting in the early to mid-1990s, the ACLU launched a highly successful campaign designed to reduce racial profiling in all aspects of American policing—from drug enforcement by state troopers and customs and immigration officials to the implementation of the death penalty to local policing efforts to disrupt gang activity and simply to enforce motor vehicle laws and criminal law more generally. Racial profiling became a contentious political issue, and a number of prominent cases of apparent police targeting—frequently of African-American men—led to numerous consent decrees and massive increases in the number of departments that collect and retain data

[138] Yinger (1998).

designed to ascertain whether their policing strategies were infected by discrimination. Again, some argued that any disparity in arrest rates across groups should be taken as evidence of discrimination but this, too, is simply another example of the gap between proof of discrimination and evidence of disparities that was discussed earlier.

One pattern that exists in certain towns in the United States that has contributed to this racial profiling litigation is that a largely white suburban area with single family homes is changed in ethnic or racial composition when a low income housing project is built in the town. Suddenly, arrests rise sharply and on a per capita basis, arrests are far more numerous in the high-density area than in the single family part of town. Because of the racially diverse compositions of the two areas, however, evidence of strong statistically significant disparities in arrests rates by race are quickly marshaled as evidence of intentional discrimination.

Ideally, tests for discrimination would develop a behavioral benchmark that corrects for the underlying rate of participation in illegal conduct, which for many crimes is all but impossible. But what if it is shown that blacks commit X percent of a particular crime but make up substantially more than X percent of the arrests for that crime? This pattern could be consistent with intentional race discrimination, but it also could be the product of a neutral practice having a disparate impact. Consider the case where a war breaks out between two gangs vying to gain control over the crack trade in an inner city environment. If the city responds to the mayhem by flooding the area with police, the ability of the police to observe criminal activity in the inner city area will be elevated and may well lead to higher arrests across the board for the residents of the targeted inner city area, who may happen to be members of a racial or ethnic minority. The effect may be that the arrest rate data that are now being routinely collected may seem to show bias on the part of police because of the disproportionate arrest rates of minorities. Ironically, to the extent that the added police activity dampens crime in the flooded area, the benefits of the policing may be disproportionately targeted on law-abiding minority members of the community—even though the political rhetoric may all focus on the discriminatory conduct of the police.[139] Still, where sentences for identical behavior can vary dramatically based on prior convictions, there is a concern about the consequences of severe disparate racial impacts in arrests.

One can imagine a model in which officers have an opportunity to seek contraband or detect criminals by engaging in certain policing actions, such as stops and frisks. If for whatever reason the success rate in these police encounters is higher when blacks are targeted, the police may have an incentive to target blacks more intensively. Efficient policing would then focus on blacks until the success rate from an enforcement action against the marginal black citizen equaled that against the marginal non-black citizen. Indeed, if, say, blacks are more likely than non-blacks to commit crime, it might be rational for the police to focus all their enforcement activity on blacks, since a corner solution may actually define the efficient policing strategy in a particular case.

[139] The claim is frequently made, though, that the police under-enforce the law in black residential areas, and over-enforce against blacks when they are in white areas.

This is precisely the theoretical approach taken by Knowles, Persico, and Todd (2001) in their study of motor vehicle searches along a Maryland highway. Their model of the search process includes a continuum of law enforcement officers and drivers, and the latter are identified by race $r \in \{B, W\}$. All other observable characteristics of motorists are bundled into the variable c. Police officers are free to search vehicles driven by any (c, r) profile and do so with probability $\gamma(c, r)$ but incur a cost of t_r. The event G denotes a search in which drugs are found, and thus the expected payoff to the officer is $P(G|c, r)$. Similarly, drivers receive $v(c, r)$ if they carry drugs and are not searched or $-j(c, r)$ if contraband is found.[140] Therefore, their expected payoff is:

$$\gamma(c, r)[-j(c, r)] + [1 - \gamma(c, r)]v(c, r) \tag{25}$$

Knowles et al. define the event when $t_B \neq t_W$ as racial prejudice since the costs of search differ by race. On the other hand, if $\gamma(B) \neq \gamma(W)$ then there is evidence of statistical discrimination. The equilibrium constructed entails randomization by motorists and police. Setting Equation (25) equal to zero, the equilibrium search rate is given by:

$$\gamma^*(c, r) = \frac{v(c, r)}{v(c, r) + j(c, r)} \tag{26}$$

Officers are willing to randomize whenever $P^*(G|c, r) = t_r$ for all c and r. In the absence of a taste for discrimination, the equilibrium probability of guilt is the same for both races. However, since Equation (26) does not depend on that probability, black motorists will be stopped and searched more often if

$$\gamma^*(c, B) = \frac{v(c, B)}{v(c, B) + j(c, B)} > \frac{v(c, W)}{v(c, W) + j(c, W)} = \gamma^*(c, W) \tag{27}$$

Note that this inequality is satisfied when the value of transporting drugs is higher or when the cost of being found guilty is lower for blacks.

Even though data on c and γ^* are not readily accessible by the econometrician, the authors test for prejudice by calculating the probability of guilt by race *conditional on being searched*. If those probabilities are the same for whites and blacks at the margin, then there is no evidence to support a racial bias claim. Such a test could be implemented by testing the null hypothesis

$$\Pr(G = 1|c, r) = \Pr(G = 1) \quad \text{for all } c, r \tag{28}$$

In order to avoid specification problems with logit and probit models, Knowles et al. opt for a nonparametric test based on the Pearson χ^2 statistic.

Their data set includes over 1,500 motor vehicle searches along Interstate 95 in Maryland between 1995 and 1999. Of those searches, 63 percent of the motorists were African-American, 29 percent were white and 6 percent were Hispanic. A first glance at the data also revealed that the percentage of African-American drivers searched had

[140] If the driver is not transporting drugs, then his payoff is zero regardless of the officer's actions.

decreased in the late 1990s, while whites were searched more often in the same time period.

Tests for equality of guilt rates across race (as well as sex, time of day and car type) are carried out according to different thresholds for measuring guilt. Knowles et al. emphasize the criteria in which any form and amount of illegal substances found constitute guilt (Definition 1) or when seizures of less than two grams of marijuana are excluded (Definition 2). When guilt is measured according to Definition 1, the hypothesis of equal conditional guilt rates is not rejected for whites and blacks but is for Hispanics and the other two race categories. Similarly, under Definition 2 there is no evidence of bias against African-American drivers. Interestingly, when the definition of guilt includes drugs in large quantities, Knowles et al. find potential signs of bias against *white* drivers.

These results are interpreted as evidence of maximizing behavior on the part of law enforcement rather than racial prejudice. As suggested by their model, differences in search rates may arise even without discriminatory preferences. Indeed, they argue that "searching some groups more often than others may be *necessary* to sustain equality in the proportions guilty across groups."

In a recent extension of the Maryland search analysis, the model of Persico and Todd (2004) allows for heterogeneous payoffs for officers and drivers and then tests for bias using data from Wichita. This version permits drivers a third option of delegating criminal activity to a member of another (r, c) group and within each group there is a joint probability distribution over v, j and d, the cost of hiring a delegate. The racial bias of police officers p now enters through an extra benefit, $B(p)$, if a successful search involves an individual of the minority race.

If the police are unbiased and both race groups are searched in equilibrium, then Persico and Todd note that their respective crime rates must be equal, or $\kappa_r = \kappa_R$, where r and R represent the minority and majority race, respectively. However, if $B(p) > 0$, i.e. officers are prejudiced, then it must be the case that the crime rate of the group subject to bias is lower, or $\kappa_r < \kappa_R$. Although characterization of equilibrium in this model is more complex than in Knowles, Persico, and Todd (2001), its implications for empirical analysis are just as straightforward. In fact, the simple analysis above of crime rates carries over into the fully specified model of interaction between police and drivers: the hit rate, or the success rate of searches, will be equal across races at the margin if police are unbiased, and the hit rate of the preferred race will be higher when police are biased.

Persico and Todd apply this test to over 2,000 vehicle searches between January and September 2001 conducted by the Wichita police department. As in the Maryland data, the percentage of blacks searched by law enforcement (32 percent) is higher than their share in the population of drivers (11 percent) while the opposite holds for whites (63 versus 65 percent). Their primary finding once again indicates that police officers are not biased in their search behavior: the hit rates for whites and blacks were 22.03 percent and 22.69 percent. Indeed, they observe equal hit rates across all three race groups. Finally, Persico and Todd summarize the results of 16 other city and state-level racial profiling studies, which hint at an empirical regularity of no police bias against black

drivers. In contrast, Gross and Barnes (2002) conclude from their analysis of the Maryland data that the Maryland State Police do use race to decide who to stop and who to search. This disparate treatment stems from the police effort to increase the minute percentage of stops that lead to drug seizures, they conclude. Gross and Barnes view the discriminatory treatment to be pointless since it has no discernible impact on the drug trade.

In general, using race to target policing activity in the absence of a specific racial description of a perpetrator will violate constitutional doctrine of equal protection under the law, but the disparate impact standard will only govern certain types of policing activities—e.g., where Congress has instituted a broader definition of discrimination for those departments that receive federal funding. One consequence of the racial profiling movement is that far more data about the racial composition of stops and arrests are now collected by the police, which presumably has some opportunity cost since officers must devote time to filling out reports. In addition, the data are costly to evaluate and does create an opportunity for knowingly or unwittingly presenting results that appear to demonstrate racial bias when none in fact exists. The data may be most valuable in reining in the misconduct of particularly biased officers, but even then the fear remains that these bad apples can avoid detection simply by not filling out the forms when they stop but do not arrest blacks. Procedures are then implemented to address that problem, but one can see that rooting out discriminatory conduct is not a trivial task, whether it is in the workplace, the police force, or other arena of social life.

The massive increase in incarceration of young black men clearly signals a social problem, although it may have less to do with discrimination in policing than with the harsh war on drugs. Even if this war is conducted in a race neutral manner, it will enmesh into the criminal justice system a disproportionate number of young males with low socio-economic status and fewer options in the legitimate labor force. Of course, an anti-drug policy directed at the demand side (rather than the current supply-side approach) would have far less racial impact since blacks make up a much smaller share of drug users than of drug sellers.[141] The latest figures show that 12 percent of black men aged 20–34 are incarcerated while the comparable figure for white men is 1.6 percent.[142] No change in policing will radically alter these numbers, although a change in drug policy clearly would. It is worth asking whether our society has done enough to try to alter this situation, or whether it is willing to accept such high levels of black incarceration because of indifference emanating from discriminatory attitudes.

12. Conclusion

We know that discrimination has been an enormous blight on the history of this country. The scholarly consensus is also clear that the enactment of the Civil Rights Act of

[141] Loury (2002).

[142] "Prison Rates among Blacks Reach a Peak, Report Finds" (2003).

1964 was a major step towards addressing this problem, and, in particular, aided the employment and earnings of blacks relative to whites for the decade from 1965 to 1975. This tells us that law was needed to stimulate demand for black labor if society was to be true to the ideal that every person should be judged by their talents and not by the color of their skin. The market alone had not given this protection, despite the claims to this effect by some very prominent economists, such as Milton Friedman. Even the libertarian Richard Epstein now concedes that the Civil Rights Act was required to break the logjam of Jim Crow. Indeed, to the extent that this federal legislation reduced the discriminatory attitudes of southern (and even non-southern) racists, the efficiency gains from reducing these Beckerian costs would be enormous. Just as de Tocqueville writing in 1833 understood that slavery was not only cruel to the slave but deeply harmful to the masters, federal antidiscrimination law revealed a century and one-half later that lifting the oppression of intense discrimination from blacks helped the citizens of the South, both black and white, immensely.[143]

There is much less consensus, though, about where things stand today. As in so many areas of the law—for example, medical malpractice, which kills more than the total victims of homicide and car accidents each year; and antitrust, where the costs of egregious acts in restraint of trade can be enormous—it is easy to point out examples of objectionable conduct, but it is also easy to see that a system of private litigation creates many problems of costly lawsuits and high rates of error. The audit studies described in Sections 6 and 10 remind us that employers and housing agents acting in a discriminatory manner are still common, but by no means dominant. The Urban Institute study of Chicago employers found that black and white testers were treated identically 85.8 percent of the time, while whites were favored in 9.6 percent of the tests and blacks were favored in 4.5 percent of the tests.[144] When one compares those figures to the percentage of Chicago employers who held negative views about the work ethic of black, white, and Hispanic employees (37.7 percent ranked blacks last), one realizes that the combination of competition in the market and the existence of employment discrimination law leads to much lower effective discrimination than one might fear.[145] Ideally, one would like to know the relative importance of law in this equation. Clearly, the economy is more competitive today than ever before, which implies that concerns about employer discrimination should be less pressing than might have been true even 20 years ago.

Heckman may well be correct that efforts to further ratchet up enforcement efforts of the current litigation-based system of antidiscrimination law would elevate costs far beyond likely benefits. In his view, the best policy would be to direct resources more heavily into education and human capital development rather than further antidiscrimination activity or affirmative action, although Loury (2002) argues that the full array

[143] De Tocqueville found the comparison of the contiguous slave state of Kentucky and the free state of Ohio to be dispositive on this issue. The first was marred by poverty and idleness, the second hummed with industry, comfort, and contentment. See Donohue (2003) quoting de Tocqueville.

[144] Donohue (2003).

[145] Donohue (2003).

of approaches will be needed to produce greater racial equality. It may be worth exploring whether it would be sensible to diminish the reliance on private litigation and place greater emphasis on programs such as the federal contract compliance program, under which government contractors are pressed to be sure to avoid "underutilization of women and minorities." Such efforts have the potential not only to redress overt imbalances in hiring procedures but also to mitigate negative, subconscious attitudes about race and sex, of which people in positions of authority may not be aware. As suggested by the empirical findings of sociologists and psychologists, latent, negative attitudes toward racial minorities have persisted despite decades of antidiscrimination legislation.[146] It is therefore likely that the problem of racial discrimination will continue to be widespread and difficult to combat.

The first phase of federal antidiscrimination law was designed to achieve color blind treatment of all workers. In its second phase, however, antidiscrimination law was harnessed as a means of improving the economic status of those who would remain disadvantaged in the marketplace by color-blind treatment: blacks, women, Hispanics, the elderly, and the disabled.[147] This was done in a way that was, arguably, less socially divisive than explicit welfare legislation that could more efficiently target benefits to these groups. Supporters of this implicit affirmative action will assert that it was social welfare-enhancing even if no longer efficient and even if somewhat disingenuously couched in the language of remedying discrimination (rather than promoting fairness or distributive justice). Over time, however, the opponents of such policies have become increasingly unhappy with the perceived excesses of such aggrandized antidiscrimination law, and we have begun to witness this trend in recent legislative initiatives designed to cut back on affirmative action in education and other governmental functions.[148]

Another goal of antidiscrimination law is to prevent the type of racial and ethnic conflagrations that persistently lead to such unhappy consequences around the world. Wise antidiscrimination law and policies may serve to dampen down such antagonisms and prevent the rigid forms of segregation that can allow biased attitudes to percolate into

[146] The experimental study of implicit attitudes in Cunningham, Preacher, and Banaji (2001) provides some interesting evidence of this phenomenon. Test subjects were shown faces of black and white individuals on a computer screen followed by words that were clearly positive or negative in connotation. In one trial, subjects pressed the same key to identify white faces and "good" words and another for black faces and "bad" words. In a second trial, the key for black faces was the same as for good words while bad words were identified with the same key as white faces. Their results revealed a statistically significant slower response time in the second trial suggesting stronger associations between the pairings in the first test. However, participants scored below the midpoint of the Modern Racism Scale (suggesting lower than average levels of racism) based on a questionnaire on explicit attitudes toward race, which indicated a disconnect between implicit and explicit feelings.

[147] See Donohue (1994).

[148] As Card and Krueger (2004) noted: "Between 1996 and 1998, California and Texas eliminated the use of affirmative action in college and university admissions. At the states' elite public universities admission rates of black and Hispanic students fell by 30–50 percent and minority representation in the entering freshman classes declined."

an unhealthy brew. Richard Posner has speculated that the violence initiated by French Muslims in November 2005 resulted from insufficient reliance on American–style antidiscrimination and affirmative action efforts.[149] On the other hand, social science evidence suggests that when affirmative action programs are pushed too aggressively, they can generate angry backlashes. Finding the correct balance, then, becomes an important element of antidiscrimination law and policy. These tensions are always bubbling beneath the surface as evidenced by the fact that Timothy McVeigh, who bombed the Oklahoma City federal building, was involved with the Aryan Republican Army and supported its white supremacist agenda.

The economic analysis of law, especially with respect to antidiscrimination measures, has endured much criticism for its "reduced form" approach to complex social and legal issues. In his denunciation of the neo-classical paradigm, Ramirez (2004) argues that the field of law and economics promotes a "truncated microeconomic analysis of race that is founded on what can only be termed pseudo-economics,"[150] citing the arguments of Arrow (1998) as justification for his position. Arrow does indeed believe that non-market-based accounts of discrimination such as social networks deserve more attention, but this is not to say that economics has little to offer those studying antidiscrimination law. In fact, Ramirez's skepticism of the field echoes precisely the issues that motivated economists like Arrow to formulate alternatives to the Beckerian theory of discrimination, such as the models presented in Section 3. There is even evidence of the multi-disciplinary approach to discrimination that Ramirez contends are woefully missing.[151] Moreover, Ramirez opines that law and economics primarily consists of theoretical analysis. As this chapter has shown, though, empirical studies investigating the effect of legal interventions on racial prejudice and the actual behavior of economic

[149] Posner (2005) states: "Another factor in the recent French riots may be the French refusal to engage in affirmative action. The French are reluctant even to collect statistics on the number of people in France of various ethnicities, their incomes, and their unemployment rates. No effort is made to encourage discrimination in favor of restive minorities (as distinct from women, who are beneficiaries of affirmative action in France) and as a result there are very few African-origin French in prominent positions in commerce, the media, or the government. Affirmative action in the United States took off at approximately the same time as the 1967 and 1968 race riots, and is interpretable (so far as affirmative action for blacks is concerned) as a device for reducing black unemployment, creating opportunities for the ablest blacks to rise, promoting at least the appearance of racial equality, and in all these ways reducing the economic and emotional precipitants of race riots. Of particular importance, affirmative action was used to greatly increase the fraction of police that are black, while the "community policing" movement improved relations between the police and the residents of black communities. French police, traditionally brutal, have by all accounts very bad relations with the inhabitants of the Muslim slums. The French riots are a reminder that affirmative action, although offensive to meritocratic principles, may have redeeming social value in particular historical circumstances."

[150] Ramirez (2004).

[151] For example, Lang (1986) proposes a theory of discrimination based on the transaction costs that accompany the emergence of distinct language or speech communities in the labor market. Lang clearly states his claim that this idea "is a distinct improvement over the existing theoretical literature on discrimination, which either relies on tastes ... or on statistical discrimination having implications generally contrary to factual evidence."

agents have provided valuable insight into the causes and consequences of discrimination.

Indeed, economic analysis has helped to identify some of the unintended consequences of antidiscrimination law, such as the fact that, as employment discrimination litigation changed from being largely about failure to hire to being primarily about wrongful discharge, the law developed from a tool that opened up new areas for minority employment to one that created some incentives against hiring minorities. The potential drag on minority employment resulted from the increased cost associated with hiring someone who might need to be fired at a later date.[152] Another example concerns the ability of employers to circumvent the demands of law: if a firm resides in an area with a 40 percent minority population, it may be able to drastically reduce its reliance on black labor by moving to another locale, with a black workforce of only, say, 2 percent. By doing so, both the prejudiced employer and the employer fearful of discrimination suits might be able to avoid the psychic or legal burden of hiring blacks altogether. In either event, the goal of increasing opportunities for blacks would be thwarted. Similarly, an impressive recent study has raised concerns about whether the "reasonable accommodation" requirements of disability law are harming the employment opportunities of disabled workers.[153] More empirical work is needed before we can state with assurance the full extent of the costs and benefits of antidiscrimination law in employment, housing, lending, medical care, and criminal justice policy.

References

Aigner, D.J., Cain, G.G. (1977). "Statistical theories of discrimination in labor markets". Industrial and Labor Relations Review 30 (2), 175–187.

Allegretto, S.A., Arthur, M.M. (2001). "An empirical analysis of homosexual/heterosexual male earnings differentials: unmarried and unequal?". Industrial and Labor Relations Review 54 (3), 631–646.

Altonji, J.G., Blank, R.M. (1999). "Race and gender in the labor market". In: Ashenfelter, O., Layard, R. (Eds.), Handbook of Labor Economics, vol. 3. North-Holland, Amsterdam, pp. 3143–3259.

Antecol, H., Cobb-Clark, D. (2002). "The sexual harassment of female active-duty personnel: effects on job satisfaction and intentions to remain in the military". Claremont McKenna College, mimeo.

Arrow, K.J. (1972). "Some mathematical models of race in the labor market". In: Pascal, A.H. (Ed.), Racial Discrimination in Economic Life. Lexington Books, Lexington, Mass., pp. 187–204.

Arrow, K.J. (1973). "The theory of discrimination". In: Ashenfelter, O.C., Hallock, K.F. (Eds.), Labor Economics, vol. 4. Edward Elgar, Aldershot, pp. 3–33.

Arrow, K.J. (1998). "What has economics to say about racial discrimination?" Journal of Economic Perspectives 12 (2), 91–100.

Ashenfelter, O., Heckman, J.J. (1976). "Measuring the effect of an antidiscrimination program". In: Ashenfelter, O., Blum, J. (Eds.), Estimating the Labor Market Effects of Social Programs. Princeton University Press, Princeton, NJ, pp. 46–89.

Autor, D., Scarborough, D. (2004). "Will job testing harm minority workers?" NBER Working Paper No. 10763.

[152] Donohue and Siegelman (1991).

[153] Jolls and Prescott (2004).

Averett, S., Korenman, S. (1996). "The economic reality of the beauty myth". Journal of Human Resources 31 (2), 304–330.
Ayres, I. (1991). "Fair driving: gender and race discrimination in retail car negotiations". Harvard Law Review 104, 817–872.
Ayres, I. (1995). "Further evidence of discrimination in new car negotiations and estimates of its causes". Michigan Law Review 94 (1), 109–147.
Ayres, I. (2001). Pervasive Prejudice? Unconventional Evidence of Race and Gender Discrimination. The University of Chicago Press, Chicago.
Ayres, I., Siegelman, P. (1995). "Race and gender discrimination in bargaining for a new car". American Economic Review 85 (3), 304–321.
Ayres, I., Vars, F., Zakariya, N. (2005). "To insure prejudice: racial disparities in taxicab tipping". Yale Law Journal 114, 1613.
Babcock, L., Laschever, S. (2003). Women Don't Ask: Negotiation and the Gender Divide. Princeton University Press, Princeton, N.J.
Bayer, P., Ross, S., Topa, G. (2005). "Place of work and place of residence: informal hiring networks and labor market outcomes". NBER Working Paper No. 1019.
Becker, G. (1957). The Economics of Discrimination. University of Chicago Press, Chicago.
Becker, G. (1968). "Discrimination, economic". In: Sills, D.L. (Ed.), International Encyclopedia of the Social Sciences, vol. 4. Macmilian, New York, pp. 208–210.
Becker G. (1993). "The evidence against banks doesn't prive bias". Business Week, 14 April.
Bendick, M. Jr., Jackson, C.W., Reinoso, V.A. (1994). "Measuring employment discrimination through controlled experiments". Review of Black Political Economy 23 (1), 25–48.
Berkovic, J.A., Canner, G.B., Gabriel, S.A., Hannan, T.H. (1996). "Mortgage discrimination and fha loan performance". Cityscape: A Journal of Policy Development and Research 2 (1), 9–24.
Berlin, I. (1969). "Two concepts of liberty". In: Four Essays on Liberty. Oxford University Press, London, pp. 118–172.
Bertrand, M., Mullainathan, S. (2004). "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination". American Economic Review 94 (4), 991–1011.
Bhagwati, J. (2004). In Defense of Globalization. Oxford University Press, Auckland.
Black, D.A. (1995). "Discrimination in an equilibrium search model". Journal of Labor Economics 13 (2), 309–334.
Black, D.A., Makar, H.R., Sanders, S.G., Taylor, L.J. (2003). "The earnings effects of sexual orientation". Industrial and Labor Relations Review 56 (3), 449–469.
Black, S.E., Brainerd, E. (2002). "Importing equality? The impact of globalization on gender discrimination". NBER Working Paper No. 9110.
Blank, R., Dabady, M., Citro, C. (Eds.) (2004). Measuring Racial Discrimination. National Academies Press, Washington, D.C.
Blau, F.D., Kahn, L.M. (1996). "Wage structure and gender differentials: an international comparison". Economica 63 (250), S29–S62.
Borjas, G.J., Bronars, S.G. (1989). "Consumer discrimination and self-employment". Journal of Political Economy 97 (3), 581–605.
Bostic, S.G. (1996). "The role of race in mortgage lending: revisiting the Boston fed study". Division of Research and Statistics, Federal Reserve Board of Governors, Washington, D.C., Working Paper.
Buruma, I., Margalit, A. (2002). "Occidentalism". The New York Review of Books, January 17, 4–7.
Cain, G.G. (1986). "The economic analysis of labor market discrimination: a survey". In: Ashenfelter, O., Layard, R. (Eds.), Handbook of Labor Economics, vol. 1. North-Holland, Amsterdam, pp. 693–785.
Card, D., Krueger, A.B. (2004). "Would the elimination of affirmative action affect highly qualified minority applicants? Evidence from California and Texas". NBER Working Paper No. 10366.
Carneiro, P., Heckman, J.J., Masterov, D.V. (2005). "Labor market discrimination and racial differences in premarket factors". Journal of Law and Economics 48 (1), 1–40.
Chay, K. (1998). "The impact of federal civil rights policy on black economic progress: evidence from the equal employment opportunity act of 1972". Industrial and Labor Relations Review 51 (4), 608–632.

Chiswick, B.R. (1973). "Racial discrimination in the labor market: a test of alternative hypotheses". Journal of Political Economy 81 (6), 1330–1352.
Conroy, M. (1994). Faded Dreams: The Politics and Economics of Race in America. Cambridge University Press, Cambridge.
Couch, K., Daly, M. (2002). "Black-white wage inequality in the 1990s: a decade of progress". Economic Inquiry 40, 31–41.
Crosby, F., Bromley, S., Saxe, L. (1980). "Recent unobtrusive studies of black and white discrimination and prejudice: a literature review". Psychological Bulletin 87, 546–563.
Cunningham, W.A., Preacher, K.J., Banaji, M.R. (2001). "Implicit attitude measures: consistency, stability, and convergent validity". Psychological Science 12 (2), 163–170.
Darity, W.A., Mason, P.L. (1998). "Evidence on discrimination in employment: codes of color, codes of gender". Journal of Economic Perspectives 12 (2), 63–90.
Donohue, J.J. (1986). "Is Title VII efficient?" University of Pennsylvania Law Review 134, 1411–1431.
Donohue, J.J. (1987). "Further thoughts on employment discrimination legislation: a reply to Judge Posner". University of Pennsylvania Law Review 136, 523–551.
Donohue, J.J. (1989). "Prohibiting sex discrimination in the workplace: an economic perspective". University of Chicago Law Review 56, 1337–1368.
Donohue, J.J. (1994). "Employment discrimination law in perspective: three concepts of equality". Michigan Law Review 92, 2583–2612.
Donohue, J.J. (2002). "The search for truth: in appreciation of James J. Heckman". Law and Social Inquiry 27 (1), 23–34.
Donohue, J.J. (2003). Foundations of Employment Discrimination Law. Foundation Press, New York.
Donohue, J.J., Heckman, J.J. (1991). "Continuous versus episodic change: the impact of civil rights policy on the economic status of blacks". Journal of Economic Literature 29 (4), 1603–1643.
Donohue, J.J., Siegelman, P. (1991). "The changing nature of employment discrimination litigation". Stanford Law Review 43, 983–1033.
Donohue, J.J., Siegelman, P. (1993). "Law and macroeconomics: employment discrimination over the business cycle". University of S. Calif. L. Rev. 66, 709.
Duleep, H., Zalokar, N. (1991). "The measurement of labor market discrimination when minorities respond to discrimination". In: Cornwall, R., Wunnava, P. (Eds.), New Approaches to Economic and Social Analyses of Discrimination. Praeger Press, New York, pp. 181–198.
Dworkin, R. (1993). "Women and pornography". The New York Review of Books, October 21, 36–42.
Economic Report of the President (2003). States Government Printing Office, Washington, D.C.
Ellis, J.J. (2004). His Excellency: George Washington. Alfred A. Knopf, New York, N.Y.
Epstein, R. (1992). Forbidden Grounds: The Case against Employment Discrimination Laws. Harvard University Press, Cambridge, Mass.
Fatsis, S. (2001). Word Freak: Heartbreak, Triumph, Genius, and Obsession in the World of Competitive Scrabble Players. Houghton Mifflin Co., Boston.
Ferguson, R.F. (1995). "Shifting challenges: fifty years of economic change toward black-white earnings equality". In: Clayton, O. (Ed.), An American Dilemma Revisited: Race Relations in a Changing World. Russell Sage Foundation, New York, pp. 76–111.
"Financial Firms Hasten Their Move to Outsourcing" (2004). The New York Times, August 18, C1.
Finkin, M. (1996). "Employce privacy, American values and the law". Kent Law Rewiew 72, 221.
Freeman, R.B., Gordon, R.A., Bell, D., Hall, R.E. (1973). "Changes in the labor market for black Americans, 1948-72". Brookings Papers on Economic Activity 1973 (1), 67–131.
Friedman, M. (1962). Capitalism and Freedom. University of Chicago Press, Chicago.
Fryer, R.G., Levitt, S.D. (2004). "The causes and consequences of distinctively black names". Quarterly Journal of Economics 119 (3), 767–805.
Fryer, R.G., Levitt, S.D. (2005). "The black-white test score gap through third grade". NBER Working Paper No. 11049.
Gneezy, U., Niederle, M., Rustichini, A. (2003). "Performance in competitive environments: gender differences". Quarterly Journal of Economics 118 (3), 1049–1074.

Gneezy, U., Rustichini, A. (2004). "Gender and competition at a young age". American Economic Review 94 (2), 377–381.
Goldberg, P.K. (1996). "Dealer price discrimination in new car purchases: evidence from the consumer expenditure survey". Journal of Political Economy 104 (3), 622–654.
Goldin, C., Rouse, C. (2000). "Orchestrating impartiality: the impact of 'blind' auditions on female musicians". American Economic Review 90 (4), 715–741.
Greenwald, A.G., Banaji, M.R., Rudman, L.A., Farnham, S.D., Nosek, B.A., Mellott, D.S. (2002). "A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept". Psychological Review 109, 3–25.
Gross, S.R., Barnes, K. (2002). "Road work: racial profiling and drug interdiction on the highway". Michigan Law Review 101 (3), 653–754.
Hakim, C. (2000). Work-Lifestyle Choices in the 21st Century: Preference Theory. Oxford University Press, Oxford.
Hakim, C. (2003). Models of the Family in Modern Societies: Ideals and Realities. Ashgate, Aldershot, England.
Hamermesh, D.S. (2006). "Changing looks and changing 'discrimination': the beauty of economists". Economics Letters 93 (3), 405–412.
Hamermesh, D.S., Biddle, J.E. (1994). "Beauty and the labor market". American Economic Review 84 (5), 1174–1194.
Hamermesh, D.S., Parker, A.M. (2005). "Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity". Economics of Education Review 24 (4), 369–376.
Han, S. (2002). "Learning and statistical discrimination in lending". Unpublished manuscript.
Heckman, J.J. (1998). "Detecting discrimination". Journal of Economic Perspectives 12 (2), 101–116.
Heckman, J.J., Payner, B.S. (1989). "Determining the impact of federal antidiscrimination policy on the economic status of blacks". American Economic Review 79 (1), 138–177.
Heckman, J.J., Siegelman, P. (1993). "The urban institute audit studies: their methods: response to comments by John Yinger". In: Fix, M., Struyck, R. (Eds.), Clear and Convincing Evidence: Measurement of Discrimination in America. The Urban Institute Press, Washington, pp. 271–275.
Heckman, J.J., Wolpin, K. (1976). "Does the contract compliance program work? An analysis of Chicago data". Industrial and Labor Relations Review 29, 544–564.
Holzer, H.J., Ihlanfeldt, K.R. (1998). "Customer discrimination and employment outcomes for minority workers". Quarterly Journal of Economics 113 (3), 835–867.
Jolls, C., Prescott, J.J. (2004). "Disaggregating employment protection: the case of disability discrimination". NBER Working Paper No. 10740.
Kirby, J. (2003). Choosing To Be Different: Women Work and the Family. Center for Policy Studies, London.
Knowles, J., Persico, N., Todd, P. (2001). "Racial bias in motor vehicle searches: theory and evidence". Journal of Political Economy 109 (1), 203–229.
Ladd, H.F. (1998). "Evidence on discrimination in mortgage lending". Journal of Economic Perspectives 12 (2), 41–62.
Lang, K. (1986). "A language theory of discrimination". Quarterly Journal of Economics 101 (2), 363–382.
Lang, K., Manove, M. (2004). "Education and labor-market discrimination". Boston University, mimeo.
Leonard, J. (1984a). "The impact of affirmative action on employment". Journal of Labor Economics 2, 439–463.
Leonard, J. (1984b). "Employment and occupational advance under affirmative action". The Review of Economics and Statistics 66 (3), 377–385.
List, J.A. (2004). "The nature and extent of discrimination in the marketplace: evidence from the field". Quarterly Journal of Economics 119 (1), 49–89.
Loury, G.C. (2002). The Anatomy of Racial Inequality. Harvard University Press, Cambridge, Mass.
Lundberg, S., Startz, R. (1998). "On the persistence of racial inequality". Journal of Labor Economics 16 (2), 292–323.
Maxwell, N.L. (1994). "The effect on black-white wage differences of differences in the quantity and quality of education". Industrial and Labor Relations Review 47 (2), 249–264.

McAdams, R.H. (1995). "Cooperation and conflict: the economics of group status production and race discrimination". Harvard Law Review 108, 1003–1084.
Meier, P., Sacks, J., Zabell, S.L. (1984). "What happened in Hazelwood: statistics, employment discrimination, and the 80% rule". American Bar Foundation Research Journal 9 (1), 139–186.
"Military Loses Able Recruits with Gay Rule; ousted linguists' skills badly needed" (2003). Chicago Tribune, January 23, 8.
Mulligan, C.B., Rubinstein, Y. (2005). "Selection, investment, and women's relative wages since 1975". NBER Working Paper 11159.
Munnell, A.H., Tootell, G.M.B., Browne, L.E., McEneaney, J. (1996). "Mortgage lending in Boston: interpreting HMDA data". American Economic Review 86 (1), 25–53.
Nardinelli, C., Simon, C. (1990). "Customer racial discrimination in the market for memorabilia: the case of baseball". Quarterly Journal of Economics 105 (3), 575–595.
Neal, D.A., Johnson, W.R. (1996). "The role of premarket factors in black-white wage differences". Journal of Political Economy 104 (5), 869–895.
Neumark, D., Bank, R.J., Van Nort, K.D. (1995). "Sex discrimination in restaurant hiring: an audit study". Quarterly Journal of Economics 111 (3), 915–941.
Oettinger, G.S. (1996). "Statistical discrimination and the early career evolution of the black-white wage gap". Journal of Labor Economics 14 (1), 52–78.
O'Neill, J. (1990). "The role of human capital in earnings differences between black and white men". Journal of Economic Perspectives 4 (4), 25–45.
Orfield, G., Ashkinaze, C. (1991). The Closing Door: Conservative Policy and Black Opportunity. University of Chicago Press, Chicago.
Oster, E. (2005). "Hepatitis B and the case of the missing women". CID Graduate Student and Postdoctoral Fellow Working Paper (#7). (Available at: http://www.cid.harvard.edu/cidwp/graduate.html.)
Oyer, P., Schaefer, S. (2000). "Layoffs and litigation". Rand Journal of Economics 31 (2), 345–358.
Oyer, P., Schaefer, S. (2002a). "Litigation costs and returns to experience". American Economic Review 92 (3), 683–705.
Oyer, P., Schaefer, S. (2002b). "Sorting, quotas, and the Civil Rights Act of 1991: who hires when it's hard to fire?" Journal of Law and Economics 45 (1), 41–68.
Pager, D. (2003). "The mark of a criminal record". American Journal of Sociology 108 (5), 937–975.
Persico, N., Todd, P. (2004). "Using hit rate tests to test for racial bias in law enforcement: vehicle searches in Wichita". NBER Working Paper No. 10947.
Phelps, E.S. (1972). "The statistical theory of racism and sexism". American Economic Review 62 (4), 659–661.
Posner, R.A. (1987). "The efficiency and efficacy of Title VII". University of Pennsylvania Law Review 136, 513–521.
Posner, R.A. (1989). "An economic analysis of sex discrimination laws". University of Chicago Law Review 56, 1311–1335.
Posner, R.A. (2005). "The French Riots". The Becker-Posner Blog on November 13. (http://www.becker-posner-blog.com/archives/2005/11/.)
Power, S. (2002). "Genocide and America". The New York Review of Books, March 14, 15–18.
"Prison Rates among Blacks Reach a Peak, Report Finds" (2003). The New York Times, April 7, A11.
Ramirez, S.A. (2004). "What we teach when we teach about race: the problem of law and pseudo-economics". Journal of Legal Education 54 (3), 365–379.
Rawls, J. (1971). A Theory of Justice. Belknap Press, Cambridge, Mass.
"Report on the Glass Ceiling Initiative" (1991). US Department of Labor, Washington, DC, USA.
Ross, S.L. (2003). "What is known about testing for discrimination: lessons learned by comparing across different markets". University of Connecticut, Department of Economics, Working Paper.
Ross, S.L., Yinger, J. (1999). "Does discrimination in mortgage lending exist? The Boston fed study and its critics". In: Turner, M.A., Skidmore, F. (Eds.), Mortgage Lending Discrimination: A Review of Existing Evidence. The Urban Institute, Washington, D.C.

Ross, S.L., Yinger, J. (2002). The Color of Credit: Mortgage Discrimination, Research Methodology, and Fair-Lending Enforcement. MIT Press, Cambridge, Mass.
Schafer, R., Ladd, H.F. (1981). Discrimination in Mortgage Lending. MIT Press, Cambridge, Mass.
Schauer, F.F. (2003). Profiles, Probabilities, and Stereotypes. Belknap Press, Cambridge, Mass.
Sen, A. (1970). "The impossibility of a paretian liberal". The Journal of Political Economy 78 (1), 152–157.
Sen, A. (1990). "More than 100 million women are missing". The New York Review of Books, December 20, 61–66.
Siegelman, P. (1998). "Racial Discrimination in 'everyday' commercial transactions: what do we know, what do we need to know, and how can we find out?" In: Fix, M.E., Turner, M.A. (Eds.), A National Report Card on Discrimination in America: The Role of Testing. The Urban Institute, Washington, D.C.
Smith, J.P., Welch, F.R. (1989). "Black economic progress after myrdal". Journal of Economic Literature 27 (2), 519–564.
Steele, C., Aronson, J. (1995). "Stereotype threat and the intellectual test performance of African Americans". Journal of Personality and Social Psychology 69 (5), 797–811.
Stengel, M., Glennon, D. (1994). "An evaluation of the federal Reserve Bank of Boston's study of racial discrimination in mortgage lending". Office of the Comptroller of the Currency, Economic & Policy Analysis Working Paper 94-2.
"The Urge to Win" (2005). The New York Times, May 31, A17.
USMSPB (1995). "Sexual Harassment in the Federal Workplace: Trends, Progress and Continuing Challenges, A Report to the President and the Congress of the United States". U.S. Merit Systems Protection Board, Washington, D.C.
Waldfogel, J. (1998). "Understanding the 'family gap' in pay for women with children". Journal of Economic Perspectives 12 (1), 137–156.
Wellington, S., Kropf, M.B., Gerkovich, P.R. (2003). "What's holding women back?" Harvard Business Review 81, 18–19.
Yinger, J. (1995). Closed Doors, Opportunities Lost: The Continuing Costs of Housing Discrimination. Russell Sage Foundation, New York.
Yinger, J. (1998). "Evidence on discrimination in consumer markets". Journal of Economic Perspectives 12 (2), 23–40.
Zimmer, M.J., Sullivan, C.A., Richards, R., Calloway, D.A. (1994). Cases and Materials on Employment Discrimination. Little, Brown and Company, Boston.

Chapter 19

INTELLECTUAL PROPERTY LAW

PETER S. MENELL
School of Law, University of California, Berkeley

SUZANNE SCOTCHMER
Department of Economics and Goldman School of Public Policy, University of California, Berkeley

Contents

Handbook of Law and Economics, Volume 2
Edited by A. Mitchell Polinsky and Steven Shavell

DOI: 10.1016/S1574-0730(07)02019-1

Abstract

This chapter provides a comprehensive survey of the burgeoning literature on the law and economics of intellectual property. It is organized around the two principal objectives of intellectual property law: promoting innovation and aesthetic creativity (focusing on patent, trade secret, and copyright protection) and protecting integrity of the commercial marketplace (trademark protection and unfair competition law). Each section sets forth the economic problem, the principal models and analytical frameworks, application of economic analysis to particular structural and doctrinal issues, interactions with other legal regimes (such as competition policy), international dimensions, and comparative analysis of intellectual property protection and other means of addressing the economic problem (such as public funding and prizes in the case of patent and copyright law and direct consumer protection statutes and public enforcement in the case of trademarks).

Keywords

Innovation, patent, copyright, trademark, trade secret, R&D, government funding, antitrust, licensing, goodwill

JEL classification: K11, L40, O34, Z11

The digital revolution and other technological breakthroughs of the past several decades have brought intellectual property to the forefront of economic, social, and political interest. Much of the value of the leading companies in the world today resides in their portfolio of intangible assets—ranging from the better defined forms of intellectual property (such as patents and copyrights) to the least tangible of the intangibles (trade secrets (know-how) and trademarks (good will associated with a brand)). By one estimate, approximately two-thirds of the value of major industrial companies derives from intangible assets (Swiss Reinsurance Company, 2000). Not surprisingly, there has been a deluge of economic analyses of intellectual property law during the past decade (Menell, 1998a; Landes and Posner, 2003; Jaffe and Lerner, 2004; Gallini and Scotchmer, 2002; Merges, Menell, and Lemley, 2003; Scotchmer, 2004b).

At the outset, it is important to clarify two important issues relating to "intellectual property." Although it draws upon certain characteristics from the law relating to real and personal "property"—most notably, the concept of exclusive rights—and many parallels can be readily identified, the differences between tangible forms of "property" and "intellectual property" are profound and numerous. To take one prominent example, whereas the traditional bundle of rights associated with real and personal property involve perpetual ownership (the classic "fee simple absolute" of real property law), two of the most prominent forms of intellectual property—patents and copyrights—protect rights for limited durations (although in the case of copyrights, the term is quite long). Furthermore, exclusivity in the field of "intellectual property" is far less inviolate than it is in the traditional property domains. Intellectual property law comprises a system of policy levers that legislatures tailor and courts interpret in order to promote innovation and protect the integrity of markets.

Second, the field of "intellectual property" is far from unified or monolithic. The landscape of intellectual property comprises a highly variegated array of quite distinct legal regimes: patent, copyright, trade secret, trademark, and a variety of specialized modes of protection (e.g., mask work protection). Although multiple intellectual property regimes can protect different aspects of the same work—computer software being a prime example—it is important to recognize that each mode of intellectual property protection has distinct characteristics and limitations.

For purposes of exploring the economic dimensions of the intellectual property field, it is important to distinguish between two quite distinct functions. The principal objective of intellectual property law is the promotion of new and improved works—whether technological or expressive. This purpose encompasses patent, copyright, and trade secret law, as well as several more narrow protection systems (e.g., mask works, database, design, and misappropriation). The other purpose of intellectual property law addresses a very different economic problem—ensuring the integrity of the marketplace. Trademark law and related bodies of unfair competition law respond to this concern.

1. Promoting innovation

Economic interest in intellectual property grows primarily out of the critical importance of innovation to social welfare. Solow (1957) demonstrated that technological advancement and increased human capital of the labor force accounted for most (between 80 and 90%) of the annual productivity increase in the U.S. economy between 1909 and 1949, with increases in the capital/labor ratio accounting for the remainder. Denison (1985) extended and refined this analysis, reaching similar results for the period 1929–1982: 68% of productivity gain due to advances in scientific and technological knowledge, 34% due to improved worker education, 22% due to greater realization of scale economies, and 13% attributable to increased capital intensity; these factors were offset by decreases in work hours (−25%), government regulation (−4%), and other influences. It is now widely recognized that technological advancement and enhanced human capital are the principal engines of economic growth in the United States and other industrialized countries (Scherer and Ross, 1990, pp. 613–614).

The role of intellectual property in contributing to innovation, however, has been more difficult to establish. As we will see, the availability of intellectual property for innovation creates incentives for investment as well as potential impediments to diffusion and cumulative innovation. The net effects are quite complex to sort out from both theoretical and empirical perspectives. As a means for surveying and synthesizing the field, we begin in Section 1.1 by clarifying the economic problem that motivates interest in intellectual property protection. Section 1.2 provides an overview of the principal modes of intellectual property protection aimed at promoting innovation and creativity. We then survey the design of intellectual property systems and discuss the principal policy levers available for tailoring such protection, focusing first upon stand-alone innovation before turning to cumulative innovation. Part 1.4 examines administration of intellectual property regimes. We then turn to enforcement of intellectual property, its interaction with competition policy, and applied research on its role in particular industries. Part 1.8 returns to the question with which we begin—the comparative efficacy of intellectual property in promoting innovation. The final section discusses the economics of intellectual property treaties.

1.1. The economic problem

The principal justification for intellectual property derives from a broader economic problem: the inability of a competitive market to support an efficient level of innovation. In a competitive economy, profits will be driven to zero, not accounting for sunk costs such as research and development (R&D) or costs of authorship. From an *ex post* point of view, this is a good outcome, as it keeps prices low for consumers and avoids deadweight loss. But from an *ex ante* point of view, it produces a sub-optimal level of investment in R&D. Most firms would not invest in developing new technologies, and potential creators might not spend their time on creative works, if rivals could enter the market and dissipate the profit.

Unlike tangible goods, knowledge and creative works are public goods in the sense that their use is nonrival (Arrow, 1962; Nelson, 1959). One agent's use does not limit another agent's use. Indeed, in its natural state (cartooned in the digital age as "bits want to be free"), knowledge is also "nonexcludable." That is, even if someone claims to own the knowledge, it is difficult to exclude others from using it. Intellectual property law is an attempt to solve that problem by legal means; it grants exclusive use of the protected knowledge or creative work to the creator. For other forms of property, exclusion is often accomplished by physical means, such as building a fence. Intellectual property is a legal device by which the inventor can control entry and exclude users from intangible assets.

Of course, intellectual property results in deadweight loss to consumers, and that is its main defect. Two other defects are that it may inhibit the use of scientific or technological knowledge for further research, and, from an *ex ante* point of view, that there is no guarantee that the research effort will be delegated efficiently to the most efficient firms, or even to the right number of firms. Commentators have been lamenting the defects of intellectual property since the nineteenth century, in more or less the same terms as today (Machlup and Penrose, 1950).

But intellectual property also has virtues, of which we mention three powerful ones. Probably the greatest virtue is that every invention funded with intellectual property creates a Pareto improvement. No one is taxed more than his willingness to pay for any unit he buys; else he would not buy it. In contrast, funding out of general revenue runs the risk of imposing greater burdens on individual taxpayers than the benefits they receive.

A second great virtue is decentralization. Probably the most important obstacle to effective public procurement is in finding the ideas for invention that are widely distributed among firms and inventors. The lure of intellectual property protection does that automatically. Decentralization is especially important if private inventors are more likely than public sponsors to think of good ideas for innovations.

The third virtue is that intellectual property is an effective screening device. Cf. Long (2002) (emphasizing the role of patents as a signaling device). Since the private value of the invention generally reflects the social value, inventors should be willing to bear higher costs for inventions of higher value. The intellectual property mechanism encourages inventors to weed out their bad ideas.

But these are not determinative, since other incentive mechanisms may share the same virtues while at the same time reducing deadweight loss. Whereas the earlier economics literature proceeded as if intellectual property protection was the self-evident solution to the incentive problem, a more recent literature, beginning with Wright (1983) and discussed below, has tried to understand when that is true, and when other incentive mechanisms might dominate.

This shift in emphasis has led to another realization: The choice among incentive mechanisms, and even the optimal design of intellectual property laws, depends importantly on the nature of the creative process or, in economists' jargon, on the model of

knowledge creation. We mention some of these up front, as our later discussion of the optimal design of intellectual property will refer to them.

Four principal models of technological change have been proposed in the economics literature: the evolutionary model, the model of induced technical change, a production function for knowledge, and an exogenous process of idea formation, with incentives determining investments.

In the evolutionary model proposed by Nelson and Winter (1982) [see also Mokyr (1990, Ch. 11)], technology evolves in an evolutionary process in which R&D investments occur whenever profit drops below a specified level. Hence, the evolutionary model is not set up to investigate incentives at all, since investment is automatic. In the model of induced technical change proposed by Hicks, technical change occurs in response to changes in factor prices: "A change in the relative prices of the factors of production is itself a spur to invention and inventions of a particular kind—directed at economizing the use of a factor which has become relatively expensive" (Hicks, 1932, pp. 124–125; see also Ruttan, 2001). Thus rising energy prices can be expected to spur technological advances in energy conservation (Newell, Jaffe, and Stavins, 1999). In the production-function model of discovery, which is the basis of almost all the literature that studies patent races, there is an exogenously given relationship which determines, as a function of research inputs or the number of researchers, either the quality of invention (de Laat, 1996; Shavell and van Ypersele, 2001; Che and Gale, 2003) or the likelihood of success in each time period (Loury, 1979; Lee and Wilde, 1980; Reinganum, 1982, 1985, 1989; Wright, 1983; Denicolò, 1996, 1997; and many others). In both the induced-technical-change model and the production-function model, the profit opportunities are common knowledge. Decentralization is not important. In contrast, the "ideas" model of O'Donoghue, Scotchmer, and Thisse (1998) [see also Scotchmer (1999) and Maurer and Scotchmer (2004b)] focuses directly on the scarcity of ideas. The basis of research is "imagination," and to achieve an innovation, a researcher must both have the idea for the innovation and an incentive to invest in it.

Although it is the most widely used model, the production-function model does not lead naturally to intellectual property as superior to other incentive schemes. For example, the advantages of decentralization are more important in a model where "ideas are scarce" than where "ideas are common knowledge." A recurrent theme below is that the optimal design of incentives depends on the model of creation that one has in mind.

1.2. An overview of the principal IP regimes promoting innovation and creativity

Patent law affords a strong and broad form of protection for technological works so as to encourage inventors to disclose their inventions and allow them to fend off imitators that seek to copy all essential elements of an invention before the inventor can recoup the costs of invention and compensate for the risk of investment. The patent offsets this power and further encourages cumulative innovation by allowing follow-on inventors to secure rights on improvements and to enable any competitor to build upon the innovation in its entirety within a comparatively short period of time (20 years from the time

of the application). Copyright law, by contrast, wholly excludes protection for ideas and functional attributes of a work but protects creators against direct or near exact copying of even a significant fragment of the whole for a longer duration of time (life of the author plus 70 years).

Trade secret law can also be seen as a means of promoting innovation, although it accomplishes this objective in a very different manner than patent. Notwithstanding the advantages of obtaining a patent—an exclusive right to practice an invention for a designated period of time—many innovators prefer to protect their innovation through secrecy. They may feel that the cost and delay of seeking a patent are too great or that they can more effectively profit from their investment through secrecy. They might also believe that the invention can best be exploited over a longer period of time than a patent would allow (Horstmann, MacDonald, and Slivinski, 1985). Without any special legal protection for trade secrets, however, the secretive inventor runs the risk that an employee (or a thief) will disclose the invention. Once the idea is released, it will be "free as the air" under the background norms of a free market economy. Such a predicament would lead any inventor seeking to rely upon secrecy to spend an inordinate amount of resources building high and impervious fences around their research facilities and greatly limiting the number of people with access to the proprietary information. Under trade secret law, an inventor need merely take *reasonable* steps to maintain secrecy in order to obtain strong remedies against individuals within the laboratory or commercial enterprise and those subject to contractual limitations who misappropriate proprietary information. Although trade secret law does not limit the use of ideas once they have become publicly known, it does significantly reduce the costs of protecting secrets within the confines of the research and commercial environment.

Table 1 provides a concise comparative summary of patent, copyright, and trade secret law—the principal modes of intellectual property protection fostering innovation.

1.3. Policy levers

From an economic perspective, the modes of intellectual property protection as well as the system as whole can be seen as an interrelated set of policy levers. The ones that have been stressed in economics journals are length and breadth, and increasingly, the threshold for protection. More recent literature, especially in law journals, has examined a wider range of rules and institutions affecting the incentive effects of intellectual property regimes.

The policy levers operate differently in different creative environments. They also operate differently in the contexts of stand-alone innovations and innovations that lay a foundation for future innovations—referred to as "cumulative innovation." In order to distinguish the economic effects, we begin with models of the stand-alone environment and then move onto the more important and complex domain of cumulative innovation.

Table 1
Comparative overview of the principal modes of intellectual property protection for promoting innovation and expressive creativity

	Utility patent*	Copyright	Trade secret
Underlying theory	limited monopoly to encourage production of utilitarian works in exchange for disclosure and ultimate enrichment of the public domain	limited monopoly to encourage the authorship of expressive works in exchange for disclosure and ultimate enrichment of the public domain	freedom of contract; protection against unfair means of competition
Source of law	Patent Act (federal)	Copyright Act (federal)	state statute (e.g., Uniform Trade Secrets Act); common law
Subject matter	processes, machines, manufactures, or compositions of matter	literary (including software), musical, choreographic, dramatic and artistic works	formula, pattern, compilation, program, device, method, technique, process
Limitations	excludes laws of nature, natural substances, printed matter (forms), mental steps	*limited by* idea/expression dichotomy (no protection for ideas, systems, methods, procedures) and useful article doctrine (only expressive elements of useful articles may be protected); no protection for facts/research	
Reqts for protection	utility; novelty; non-obviousness; adequate written description	originality (low threshold); authorship; fixation in a tangible medium	information not generally known or available; reasonable efforts to maintain secrecy; commercial value
Process for obtaining protection	Examination by the Patent Office. Limited Opposition process. Reexamination process. Maintenance fees.	now automatic, but optional registration process at the Copyright Office confers some benefits	none
Rights	exclusive rights to make, use, sell innovation as limited by contribution to art	rights of performance, display, reproduction, derivative works	protection against misappropriation—acquisition by improper means or unauthorized disclosure

(*continued on next page*)

Table 1
(*Continued*)

	Utility patent*	Copyright	Trade secret
Scope of protection	extends to literal infringement (embodiments of all materials elements of a claim) and "equivalents" (non-literal but close imitations; subject to prosecution history estoppel (no protection for reasonably foreseeable "equivalents" if claim narrowed during prosecution))	extends to substantial similarity to protected expression (even a small but significant part of an overall work)	proprietary information
Duration	20 yrs from filing (utility); extensions up to 5 yrs for drugs, medical devices and additives	life of author + 70 years; "works for hire": minimum of 95 years after publication or 120 yrs after creation	until becomes public knowledge
Disclosure	right to patent lost if inventor delays too long after disclosing before filing application; full disclosure is required as part of application; notice of patent required for damages	prior to 1978, publication without proper copyright notice (©) resulted in forfeiture; copyright notice no longer required	loss of protection (unless *sub rosa*)
Rights of others	narrow experimental use exception; limited prior user right (for business methods); reverse doctrine of equivalents potentially allows radical improvements to be practiced without license of embodied patented technology	fair use; compulsory licensing for musical compositions, cable TV, et al.; independent creation	independent discovery; reverse engineering; if someone else obtains patent, then trade secret owner may be enjoined from continued use
Costs of protection	filing, issue, and maintenance fees; litigation costs	none (protection attaches at fixation); suit requires registration; litigation costs	security expenses; personnel dissatisfaction; litigatin costs

*The Plant Patent Act and the Plant Varieties Protection Act separately provide exclusive rights for distinct asexually reproducing plant varieties and sexually reproduced varieties respectively; Design patent protection (affording 14 years of protection for non-functional ornamental designs) largely overlaps copyright protection.

(*continued on next page*)

Table 1
(*Continued*)

	Utility patent*	Copyright	Trade secret
Ownership, licensing, and assignment	Inventors may assign inventions. Licensing encouraged by completeness of property rights, subject to antitrust constraints; non-exclusive (and possibly exclusive) licenses cannot be assigned without licensor consent.	"work made for hire" doctrine—works prepared within scope of employment and commissioned works (from designated categories and evidence by written agreement) are owned by employer *ab initio* termination of transfers—assignor has inalienable termination right between 36th and 41st years (if notice given) assignment—non-exclusive licenses (and exclusive licenses in the 9th Circuit) cannot be assigned without licensor consent	Licenses must ensure that trade secret is not disclosed; trade secret licenses cannot be assigned without authorization from licensor
Remedies	injunctive relief and damages (potentially treble if willful infringement); attorney fees (in exceptional cases)	injunction against further infringement; destruction of infringing articles; damages (actual or profits); statutory damages ($200–$150,000) (w/i court's discretion); attorney fees (w/i court's discretion); criminal prosecution	civil suit for misappropriation; conversion, unjust enrichment, breach of contract; damages (potentially treble) and injunctive relief; criminal prosecution for theft

1.3.1. Stand-alone innovation

Much of the early economic modeling of the role of intellectual property in promoting innovation posed the following question: what system of incentives or rewards would best promote the attainment of a particular invention. Such models provide the basis for analyzing legal protection for a distinct and relatively narrow class of inventions which do not ultimately generate follow-on innovation. Examples from this class include the safety razor, the ballpoint pen, and pharmaceutical innovations for which the scientific mechanism is poorly understood (Nelson and Winter, 1982; von Hippel, 1988, p. 53). Even where inventors depend on prior knowledge, which is almost inevitable, the lag may be such that prior rights have expired, so that the incentive system treats inventions as stand-alone. Examples are the bicycle and the early development of the light bulb (Dyson, 2001).

Models focusing on stand-alone innovation can also be helpful in analyzing legal protection for expressive creativity. Although such works often draw upon prior art for inspiration or common reference points for the work's audience, most authors, musicians, and artists have not traditionally built so extensively upon the work of prior creators as to require express permission.[1] This proposition obviously turns on the underlying right structures—copyrights tending to be relatively narrow in comparison to patents—but it also reflects a fundamental difference between the fields of technological and expressive innovation: "Science and technology are centripetal, conducing toward a single optimal result. One water pump can be better than another water pump, and the role of patent and trade secret law is to direct investment toward such improvements. Literature and the arts are centrifugal, aiming at a wide variety of audiences with different tastes. We cannot say that one novel treating the theme, say, of man's continuing struggle with nature is in any ultimate sense 'better' than another novel—or musical composition or painting—on the same subject. The aim of copyright is to direct investment toward abundant rather than efficient expression" (Goldstein, 1986).

The critical inquiry in seeking to promote stand-alone innovation is how much profit an inventor or creator should receive and how it should be structured. The focus for stand-alone innovation, therefore, is upon ex ante incentives. As we will emphasize below [also emphasized by Gallini and Scotchmer (2002)], all of the results in this area depend sensitively on what is assumed about licensing. Collaboration and the exchange of technological knowledge across firm boundaries encounter substantial transaction costs. Arora, Fosfuri, and Gambardella (2001) find evidence that changes in the technology of technical change—most notably the growing use of digital information technologies—facilitate greater partitioning of innovation tasks activities across traditional firm boundaries. They foresee markets for technology—licensing and specialized technology transfer and innovation service firms—playing a more significant role in the production of innovation. When we turn to cumulative innovation, ex post incentives enter the analysis. The principal categories of policy levers affecting incentives to invent are the threshold for protection, duration, breadth, rights of others (and defenses), remedies, and channeling doctrines (for determining priority where intellectual property regimes overlap).

1.3.1.1. Threshold for protection As noted above, intellectual property protection results in deadweight loss to consumers. Therefore, it should only be available for significant innovation—works that are new and would not be readily forthcoming without legal encouragement. Works already in the public domain should not be protectable and the threshold for protection should be sufficiently high (or the rights sufficiently narrow) to prevent easily achieved ("obvious") advances from being insulated from free market competition.

[1] The ease of incorporating prior musical, graphic, and audiovisual works into "new" digital works has fostered greater cumulative creativity in the content fields (Lessig, 2004).

Intellectual property regimes erect several types of threshold doctrines limiting protection: (i) subject matter rules—categorical limitations on protection; (ii) substantive requirements—minimum criteria for protection; and (iii) formal requirements—administrative and technical rules that must be complied with in order to obtain and maintain protection. Patent law applies broadly to all classes of innovation (i.e., few subject matter limitations), but applies relatively rigorous standards (utility, novelty, non-obviousness (or inventive step), and adequate disclosure) through a formal examination system. By contrast, copyright applies a very low threshold for protection—a work need only be fixed in a tangible medium of expression and reflect a modicum of originality—and does not require examination (registration is optional). As we will see, such a low threshold is counterbalanced by a relatively narrow scope of protection. Trade secret law requires merely that information derive economic value from not being generally known or readily ascertainable (by proper means) by others and be the subject of efforts that are reasonable under the circumstances to maintain its secrecy.

Patent law's threshold requirements have received the most economic scrutiny. Due to the relatively uniform nature of patent protection, some have argued that certain classes of innovation (such as computer software and business methods) that may not require such lengthy protection should be subject to a sui generis form of protection [Menell, 1987 (software); Samuelson et al., 1994] or excluded from intellectual property protection altogether [Thomas, 1999 (business methods); Dreyfuss, 2000]. The basic contours of patent law were established during an age of mechanical innovation and were designed with this model (and the guild system that predominated) in mind (Merges, Menell, and Lemley, 2003, pp. 105–111). Mechanical innovation continued to comprise the bulk of patent applications well into the 20th century. During the past half century, however, various newer fields—such as chemistry, software (and business methods), and biotechnology—have increasingly come into the patent system (Allison and Lemley, 2002), calling into question the premises on which patent law was built. If specialized protection systems are not developed to address new and distinctive fields of innovation (as was partially done in the case of semiconductor chip designs—see: Semiconductor Chip Protection Act of 1984), the challenge remains of reshaping the relatively uniform patent system to accommodate the growing heterogeneity of inventive activity (cf. Burk and Lemley, 2002, 2003).

Patent law's novelty requirement—what it means to be "first"—turns on the location of the "finish line" in the race to invent. Most patent systems in the world apply a first-to-file standard; the United States determines the winner on the basis of who was the first to invent. In principle, the first-to-invent system rewards the first inventor to discover new knowledge, even if they lack the specialized patent filing resources of others. Thus, many small inventors defend the first to invent system as a means of leveling the playing field relative to large companies which may have more resources available and personnel in place to file applications more expeditiously. The first-to-file system significantly reduces the administrative costs of operating a patent system—priority depends solely on the time and date stamped on an application. Evidentiary disputes over the subtle nuances of who was first to grasp

an invention can be quite costly to resolve (Macedo, 1990). Empirical studies cast doubt on the notion that small inventors tend to do better under a first to invent system, likely reflecting the high costs of resolving priority disputes (Mossinghoff, 2002; Lemley and Chien, 2003).

The first-to-invent system also has incentive effects as to the choice between trade secrecy and disclosure (Scotchmer and Green, 1990). Inventors may be inclined to delay their applications in order to effectively extend the expiration date of a patent (20 years from the date of filing). In order to counteract this effect and promote prompt filing, U.S. patent law adds an additional layer of legal complexity (and hence uncertainty and cost): requiring that an inventor file an application within one year after the invention is disclosed (either through patenting or publication in the anywhere in the world or in public use or on sale in the United States). This reduces the delay in disclosure of new knowledge, but does not eliminate it. The first-to-file system promotes earlier disclosure of technological advances. Grushcow (2004) finds that the growing interest in patenting by academic institutions since 1980 has delayed the publication of research, potentially increasing the risk of wasteful duplication of research.

From an economic standpoint, patent law's non-obviousness standard plays the most important role in determining which innovations qualify for protection (and hence what type of innovation patent law encourages). Patent law specifies that a claimed invention must go beyond readily predictable or conventional solutions to technical, engineering, or business problems. Articulating an objective and determinative standard for non-obviousness, however, has proven elusive. In the 1940s, U.S. courts interpreted the law to require a "flash of creative genius" test [Cuno Engineering Corp. v. Automatic Devices Corp., 314 U.S. 84 (1941)]. Such a demanding formulation generated a backlash within the patent community, leading Congress to frame the standard in the following manner: a patent may not be obtained "if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious to at the time the invention was made to a person having ordinary skill in the part to which said subject matter pertains" (17 U.S.C. §103). What raises the non-obviousness hurdle above the novelty standard is that the patent examiner may consider multiple references simultaneously where there is a suggestion, teaching, or motivation to combine elements across these references. The examiner must also consider circumstantial evidence of non-obviousness (so-called "secondary considerations")—long-felt but unsolved need, commercial success of the claimed invention, failed efforts by others, copyright by others, praise for the invention, unexpected results, and disbelief of experts—but only to the extent that such factors are connected to the inventive aspects of the claim.

In its actual formulation and application, the non-obviousness rule falls short of implementing the economic gatekeeping principle. Whereas an economist would consider paramount among relevant considerations the level of research and development expense in making an invention (Merges, 1992), the U.S. Patent Act states that patentability "shall not be negatived by the manner in which the invention was made," implying that inventions requiring minimal effort (and hence likely to obtain even without pro-

tection) may nonetheless qualify for protection. In addition, research expense and effort are not listed among the traditional secondary considerations, although several court decisions on non-obviousness take note of such factors [Merges, 1992 (noting that the threshold for patentability should be lowered with regard to high-cost research); Oddi, 1989, p. 1127 (recommending that courts expressly consider "qualitative and quantitative investment in research and development" among the secondary factors)]. The legal standard for non-obviousness does consider the level of uncertainty involved in research. The fact that a research project is "obvious to try" does not render a resulting discovery "obvious" unless there was little or no ex ante uncertainty about the outcome—i.e., those skilled in the art could readily predict the outcome of the experiment. In practice, the test depends on the number of parameters and the extent to which the relevant prior art guides the experimentation process.

The role of commercial success in the non-obviousness determination has produced conflicting economic analyses and prescriptions. Drawing upon historical and empirical research on the innovation process, Merges (1988) finds commercial success to be a poor proxy for technical advance. What succeeds in the market tends to reflect product strategy and marketing more than technical advances over the prior art. Hence, Merges (1988) argues for downplaying this factor and scrutinizing the connection between market success and the technical advance. By contrast, Kitch (1977, p. 283) sees market success as consistent with the prospect theory. By using subsequent economic success as a factor favoring patentability, the patent law increases "the security of the investment process necessary to maximize the value of the patent." Both analyses support the idea that the consideration of market success in assessing non-obviousness promotes commercialization (Merges and Duffy, 2002, pp. 727–736), although it is not clear that a patent system is needed to achieve this end. Where adequate incentives exist to invent, free market forces should be adequate to promote commercialization. [But cf. Kieff (2001b) (articulating a commercialization theory of patent law).]

Several observers of the patent system perceive that the Federal Circuit has significantly lowered the non-obviousness hurdle over the past two decades (Desmond, 1993; Barton, 2003; Lunney, 2000–2001). Based on both statistical and doctrinal analysis, Lunney (2004) concludes that the Federal Circuit has effectively "eviscerated" the non-obviousness requirement, routinely affirming decisions upholding patents and frequently overturning decisions invalidating patents as obvious. See also Dunner, Jakes, and Karceski (1995). Since 1982, the year that the Federal Circuit was created, the rate at which patents have been held invalid has plummeted (Lunney, 2004). Allison and Lemley (1998) find that non-obviousness remains the most frequent ground for invalidating patents at the trial and appellate levels (42% of invalidity judgments), but often fails when raised (63.7%). As we discuss in the section on administration of intellectual property (I.D), however, empirical studies have not discerned a significant decline in patent quality.

Some commentators believe that it is now far too easy to obtain a patent, particularly with regard to business method and DNA sequence patent (Barton, 2003; Hall, 2003). They recommend raising the non-obviousness hurdle, although articulat-

ing a standard that appropriately balances the concerns of over- and underprotection has proven elusive. In fact, Scotchmer (2005b, Chapter 3) argues that the low bar to patentability is a misplaced worry, and shifts the discussion from the standard for patentability to the breadth of the right. Unnecessary patents are not harmful to competition if the patents are narrow, so that similar products compete in the market.

The non-obviousness standard may have some perverse collateral effects on the nature and timing of disclosure of new knowledge. By preemptively publishing work in progress, a firm that is ahead may induce a shake-out among rivals by raising the level of prior art to render a rival's subsequent invention obvious (Scotchmer and Green, 1990); a laggard in a patent race may be able to reduce the likelihood that a leader will be able to obtain a patent by raising the level of the prior art sufficiently to defeat patentability by a leader (Lichtman, Baker, and Kraus, 2000; Parchomovsky, 2000; cf. Bar-Gill and Parchomovsky, 2003). This possibility might also lead competitors to collude or collaborate to maximize patent opportunities. Under this theoretical account, a higher standard of non-obviousness increases the viability of a preemptive patenting strategy. The likelihood that such a strategy would be pursued by rivals has been questioned on doctrinal and practical grounds (Merges, 2004b, pp. 195–196; Eisenberg, 2000; cf. Hicks, 1995).

Much of the economics literature on trade secrets addresses the optimal level of expenditures to maintain secrecy, i.e., what constitutes "reasonable efforts" under the circumstances. Kitch (1980) argues that all such "fencing costs" are inefficient and would require only such expenses as are necessary to provide evidence of the existence of a trade secret, i.e., a notice or marking function. Friedman, Landes, and Posner (1991) make the related point that trade secret protection should be available when it is cheaper than the physical precautions that would be necessary to protect a particular piece of information.

1.3.1.2. Duration Nordhaus (1969) offered the first formal model of the optimal duration of intellectual property protection. Nordhaus asked why the life of the intellectual property right should be limited, since a longer right leads to more innovation, and more innovation creates social benefit. He argued that there is a counterveiling cost. The longer right might increase innovation, but it also increases deadweight loss on all the inframarginal innovations that would occur even with shorter protection—i.e., innovations that would be forthcoming even in the absence of the longer right. The optimal duration of a patent or copyright should balance the incentive effect against the deadweight loss in order to maximize social welfare. Many economists believe that copyright duration (life of the author plus 70 years) is much longer than justified to provide an appropriate ex ante incentive for creation of new works. See Akerlof et al. (2003); but cf. Landes and Posner (2003, p. 218) (noting that the deadweight loss from copyright protection is relatively small due to the narrow scope of copyright protection).

To see the Nordhaus argument, suppose that there is a universe of "ideas" available for investment. Let an idea be a pair (s, c) where s measures the value of the resulting innovation and c is its cost. An idea with higher s can be interpreted as leading to a larger market; a higher s means that the demand curve is shifted out. Let $\Pi(s, T)$ be

the profit available to a rightholder with an intellectual property right of length T and an idea of quality s. Π is increasing in both T and s. Let $W(s, T)$ be the corresponding social welfare associated with investment in the idea. The welfare $W(s, T)$ is the sum of consumers' surplus for the infinite life of the innovation, sold at the competitive price, minus the deadweight loss during the period of protection. Thus, W is increasing in s and decreasing in T. Finally, suppose that for each R&D cost c, the distribution of "ideas" is given by a distribution function F with density f, where $F(s|c)$ is the fraction of ideas with cost c that have value less than s.

Then the social value of investment in ideas with cost c is $\hat{W}(T, c)$ defined below, where $\sigma(T)$ is the minimum value that will elicit investment ($\Pi(\sigma(T), T) = c$). That is,

$$\hat{W}(T, c) = \int_{\sigma(T)}^{\infty} [W(s, T) - c] f(s|c)\, \mathrm{d}s$$

Notice that $\sigma'(T) < 0$. A marginal increase in T will increase investment in amount $-W(\sigma(T), T)\sigma'(T)$. However, even though investment goes up with T, total social welfare $\hat{W}(T, c)$ may go down. The change in social welfare is

$$\frac{\partial}{\partial T}\hat{W}(T, c) = \int_{\sigma(T)}^{\infty} \frac{\partial}{\partial T} W(s, T) f(s|c)\, \mathrm{d}s - W(\sigma(T), T)\sigma'(T)$$

The last term represents the welfare due to new innovations called forth by longer protection, but the first term, which is negative, represents the loss in consumers' surplus on all the inframarginal innovations that would have been achieved even with shorter duration. As T becomes large and $\sigma(T)$ becomes small, it is reasonable to think that the first term becomes large relative to the last term. Increasing the duration T beyond that point will not be in the social interest.

Of course the best length T must be established by adding up the marginal effects for all c. Depending on the distributions of (s, c) in different product classes, the one-size-fits-all nature of the patent system may provide excessive protection in some product classes, and deficient incentives in others.

Races for the intellectual property right introduce another inquiry as to how profitable the intellectual property right should be, regardless of how the profit is achieved.[2] Unlike the Nordhaus argument, the inquiry leads to an argument for limited duration that applies even if the profit is given as a prize out of general revenue and involves no deadweight loss. The argument concerns the optimal amount of R&D effort. A more

[2] Innovation races are more suited to patents and patentable subject matter than to copyrights and creative works. Such races can only occur if several rivals are vying for a right that only one of them will receive. Rights to creative works are generally narrow enough in scope that several authors can obtain protection for works that have some similarity (and hence can compete). Thus, an author may fear a reduced market due to competition from another author, but does not generally fear that he or she will be wholly excluded from the market through a rival completing their work first.

profitable right will encourage more entry into the race (the extensive margin) or more collective effort as each participant accelerates its effort (the intensive margin).

The potential benefits of inciting more effort by offering more profit depend on the creative environment—the nature of the R&D process. Nordhaus implicitly addressed a creative environment where "ideas are scarce" so that duplication of costs is not the focus. Suppose, however, that more than one potential innovator can serve the same market niche. Then there is a second reason to limit duration. Not only will there be excessive deadweight loss on inframarginal innovations, but the disparity between profit and cost will also lead to duplication of R&D cost as firms vie for the very profitable rights.

Thus, part of the inquiry into the optimal strength (profitability) of an intellectual property right concerns the extent to which additional effort is duplicative. This issue takes us back to the question, What is the right model of the creative environment? If "ideas are scarce," then races are not an issue. But if all investment opportunities are commonly known, then races may or may not be efficient, depending on the "production function for knowledge." If successes and failures in the R&D process are perfectly correlated, then a race is duplicative. If successes and failures are independent, then a race increases the probability of at least one success, or in another interpretation, accelerates progress (Loury, 1979; Lee and Wilde, 1980; Reinganum, 1982, 1985, 1989). Further, if the creative environment is one in which different firms have different unobservable ideas for how to address a given need, then entrants to a race need not be the most efficient firms or those with the best ideas (Scotchmer, 2005b, Chapter 2).

The number of entrants in a race may be too large or too small, as compared to the efficient number, depending on the size of the private reward. Suppose, for example, that two firms have different ideas about how to fill a market niche with value s. Suppose that each firm's cost is c, and that each has probability $1/2$ of succeeding. Suppose that the value of the property right will be $\Pi(s, T)$, and that the firms' prospects for success are independent. If both firms succeed, each will receive the property right with probability $1/2$.

Then a second firm will enter the patent race if $\Pi(s, T)(3/8) > c$, since its probability of receiving the patent is $3/8$. On the other hand, entry by the second firm is only efficient if $W(s, T)(3/4) - 2c > W(s, T)/2 - c$ or $W(T, s)/4 > c$. Thus, if $\Pi(s, T)(3/8) > c > W(s, T)/4$, there will be excess entry to the patent race—the second firm will enter even though that is not efficient—and if $\Pi(s, T)(3/8) < c < W(s, T)/4$ there will be too little entry.

Entry into a race may provide a private value to the entrant that is greater than the social value of the entry, and always provides a private value that is greater than the increment to private value of both firms. The latter is because of the "business-stealing effect." The second entrant's expected profit is $\Pi(s, T)(3/8) - c$, while the increment to joint profit is only $\Pi(s, T)(1/4) - c$. The second entrant's chance of winning the race and getting the patent comes partly at the first entrant's expense. It is this externality that may lead to excessive entry into a race. It also implies that, if the reward were as large as the social value, there would be too much entry. In fact, the only thing that is clear

in this environment, without imposing additional structure, is that the optimal reward is smaller than the social value of the innovation. But this is not a very useful design principle, because rewards given as intellectual property will have that attribute almost inevitably.

Landes and Posner (2003, pp. 222–228) suggest that some works may be diminished by a congestion externality. They illustrate their point by reference to the Disney Corporation's self-imposed restraint on commercialization: "To avoid overkill, Disney manages its character portfolio with care. It has hundreds of characters on its books, many of them just waiting to be called out of retirement .... Disney practices good husbandry of its characters and extends the life of its brands by not overexposing them ... They avoid debasing the currency" (Britt, 1990). Landes and Posner (2003) assert that this concern justifies perpetual protection for some works. To balance the costs of protection, they advocate a system of indefinitely renewable copyright protection, with the renewal fee acting as policy lever for diverting works not subject to congestion externalities into the public domain. They note that a similar over-saturation can arise with regard to some rights of publicity (use of persona in advertising) and trademarks.

1.3.1.3. Breadth The breadth or scope of an intellectual property right has critical bearing on its economic value, and hence its incentive effect. A broader right preempts more substitutes than a narrow right.

The scope of a patent is determined by the language of the claims (which define the boundaries of literal infringement) and the extent to which such boundaries will be stretched to cover similar, but not quite literal, embodiments. Under the "doctrine of equivalents," courts will find infringement where the accused device "performs substantially the same function in substantially the same way to obtain the same result" Graver Tank & Mfg. Co. v. Linde Air Products Co., 339 U.S. 605, 608 (1950) [quoting Sanitary Refrigerator Co. v. Winters, 280 U.S. 30, 42 (1929)]. *See also* Warner-Jenkinson Co. v. Hilton Davis Chem. Co., 520 U.S. 17 (1997). The scope of copyright is determined by the substantial similarity test in conjunction with copyright's limiting doctrines (e.g., originality, scenes a faire, non-protectability of ideas and facts, fair use)—does the defendant's work embody *substantial similarity* to *protected elements* of the plaintiff's work? In practice, a copyright is quite narrow with regard to newly created works. It is unlikely that different authors operating from the same ideas will produce substantially similar novels. Breadth issues arise, however, with regard to works built on copyrighted works, such as sequels and film adaptations. We deal with these issues in the context of cumulative innovation. Breadth does not generally arise in the context of trade secrets.

These legal tests do not map directly onto the economic concepts of breadth that economists have developed. Economic models of breadth have been developed for two market contexts: where an innovation is threatened by horizontal competition, and where an innovation might be supplanted by an improved innovation. We take up the latter question in the analysis of cumulative innovation (Section 1.3.2). For horizontal competition, breadth has been modeled in two ways: in "product space," defining how

"similar" a product must be to infringe a patent, and in "technology space," defining how costly it is to find a noninfringing substitute for the protected market.

In the first notion of breadth, introduced by Klemperer (1990) using a spatial model, the size of the market for the patented product depends on the closeness of noninfringing substitutes. A broader patent covers more of the product space, meaning that more substitutes infringe. The right to keep a substitute out of the market is profitable for the patent holder in two ways: by shifting the demand for the protected good outward (where the intellectual property owner excludes the substitute from the market) or by allowing the intellectual property owner to charge higher prices for both the patented good and the infringing substitute.

The second notion of breadth for horizontal substitutes, due to Gallini (1992), is that it determines the cost of entering the market. In this conception, the goods are exact substitutes, and breadth implicitly refers to the technology of production (process innovation) rather than closeness of substitutes in the market. Entry by a second firm does not cause demand curves to shift, but instead causes the firms to compete in a given market. A narrower patent (lower cost of entry) will lead to more entry and lower prices. Entry stops when the cost of entry can no longer be covered by competing in the market.

In both conceptions of breadth, a narrower patent leads to lower per-period profit. Thus, breadth might be conceived as a policy lever that governs profit, as described above, in a one-size-fits-all system where the duration of protection cannot be tailored to the cost of innovation. In the patent system, such tailoring is not generally done in any systematic way by the Patent Office. Examiners focus solely on ensuring that the application meets the threshold criteria and that the claims are clear. They do not adjust the "breadth" of claims. The courts exercise a modest degree of tailoring. In applying the doctrine of equivalents, courts accord "pioneering" inventions greater scope than more modest inventions. Such a rule increases the reward for major breakthroughs. The copyright system does not systematically vary the scope of protection with the cost or importance of the work.

Even within the one-size-fits-all system, there is a policy question as to whether, on average, rights designed to give a pre-specified reward should be structured as long and narrow or short and broad. The inquiry into how market rewards should be structured has led in several papers to a ratio test: a policy reform is desirable if it increases the ratio of profit to deadweight loss. The ratio test was devised by Kaplow (1984) in the patent/antitrust context, and also used by Ayres and Klemperer (1999) in the enforcement context. It reappears in the cited discussions of patent breadth. The basic notion is that deadweight loss is the consumer cost of raising money through proprietary pricing. If the ratio of profit to deadweight loss is higher, the money being raised through proprietary pricing is raised more efficiently.

In a broad class of demand curves including linear ones, any price reduction from the monopoly price will increase the profit-to-deadweight-loss ratio, but will also reduce profit, thus necessitating a compensation such as longer protection. This can be seen in Figure 1, where the monopoly price is $1/2$ and the lower price $1/3$ is the duopoly price.

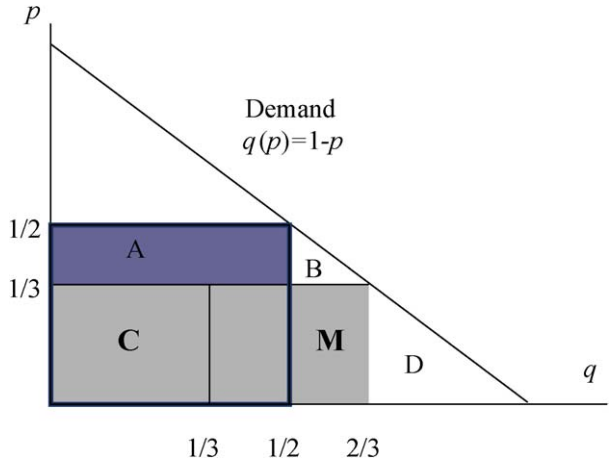


Figure 1.

At the price 1/3, the ratio of profit to deadweight loss is the ratio of the cross-lined areas (of size 2 × C) to the triangle D. At the monopoly price 1/2, the ratio of profit to deadweight loss is the ratio of the outlined box that represents monopoly profit to the triangle (B+M+D). One can see by inspection that the ratio of profit to deadweight loss is smaller at the monopoly price 1/2 than at the lower price 1/3. In fact, with the linear demand curve, this argument generalizes for any reduction in price: the lower the price, the higher the ratio of profit to deadweight loss. This is the argument given by Tandon (1982), arguing for compulsory licenses to lower prices, and Gilbert and Shapiro (1990), arguing for narrow patents, which they interpret as lower prices, although they do not say how price reductions in a given market might flow from narrower scope.

How, though, do narrow patents lower the price in a given market? In Gallini's conception, breadth determines the cost that an imitator must pay to enter a proprietary market. Entry is only tempting if the market will be protected long enough so that the entrant, in competition with the patent holder, can still cover the cost of entry. If entry occurs, competition between the entrant the rightholder will lower the price.

Figure 1 can be used to compare a relatively short period of protection where entry by an imitator is not tempting, with a longer period of protection, where entry is tempting even though the imitator must pay a cost. With the shorter period of protection, say T^M, consumers will pay the monopoly price 1/2, but with the longer period of protection, say $T^D > T^M$, they will pay the duopoly price 1/3. Suppose that T^M and T^D are chosen so that the patent holder makes the same discounted profit in both regimes, and the cost of entry is such that exactly one imitator will enter if the patent lasts for length T^D. Then by the above argument, consumers would be better off in the duopoly regime, despite the longer period of protection, because of the lower price.

However that argument does not account for the fact that the imitator must pay real resource costs to enter the market. Gallini argues that the duplication of costs is severe

enough to overturn the above argument. Given that the price can only be reduced by costly entry, it is better for society as a whole—including consumers, the patent holder and the imitator—to have a short period of monopoly pricing than a longer period that attracts entry.

However, we have already stressed that the best design of intellectual property rights depends importantly on what one assumes about licensing. In this case, licensing again overturns the conclusion. Maurer and Scotchmer (2002) argue that the patent holder will anticipate entry, and offer a license instead of tolerating unlicensed entry. In this way the patent holder can increase his own profit without reducing the profit of the entrant, and at the same time can eliminate the wasteful duplication. The narrow patent thus has the effect of lowering price without imposing the social cost of duplication, and the above welfare analysis is restored. The better policy is a narrow patent for a relatively long time.

1.3.1.4. Rights of others (and defenses) The rights afforded others in protected works directly affect the profits from intellectual property. Many of these rules—such as blocking rights (patent and copyright) and exceptions for experimental use (patent), fair use (copyright), and reverse engineering (copyright and trade secret)—find their economic justification in the cumulativeness of innovation, and therefore we take them up in Section 1.3.2. Doctrines relating to independent invention, prior user rights, and "first sale" (or exhaustion of rights) relate to stand-alone invention, as do proposals about extending user rights to limited sharing of copyrighted works.

Rights arising from independent invention. A right of independent invention means that, provided the independent inventor was actually an "inventor" (and, in particular, did not learn the invention from any other party, such as a prior inventor), he or she is free to practice the invention. Both copyright law and trade secret law immunize independent inventors from liability, but patent law does not. In the case of trade secrets, it would be impossible for an independent inventor to know what had previously been invented. In the case of copyrights, which protect expression, any re-expression escapes liability (broadly speaking). In the case of patent law, the right is defined with respect to claims, and (broadly speaking) not with respect to how a potential infringer achieved the potentially infringing innovation. These principles have doctrinal nuances, some of which are mentioned below.

Scholars have made three types of economic arguments about independent invention. First, in the context of trade secrecy, the absence of an independent-invention defense would stifle innovation because inventors would be uncertain as to whether they could practice the new knowledge they create.

Second, a right of independent invention can reduce the duplication of R&D costs in patent races (La Manna, MacLewod, and de Meza, 1989; Blair and Cotter, 2002; Maurer and Scotchmer, 2002; Leibovitz, 2002; Ottoz and Cugno, 2004). If the value of an exclusive right in the market is $100, and the R&D cost is $20, five firms may enter a race. But if all five firms have rights ex post, competition will reduce the private value

of the right below $100, and fewer than five firms will enter. The right of independent invention reduces the duplication of costs, and at the same time affords lower prices to users, all without undermining the incentive to invent.

Landes and Posner (2003, pp. 361–362) make a similar argument for trade secrets. They compare the American rule, under which the owner of a trade secret loses his right to the invention if someone else patents it [W.L. Gore & Assocs. v. Garlock, Inc., 721 F.2d 1540 (Fed. Cir. 1983)], to the prior-user-right that prevails in some other nations. The prior-user-right divides the entitlement, enabling multiple independent inventors to share its value through an effective oligopoly structure. As in the foregoing argument, duplicative entry will only occur to the extent that all firms cover their costs.

Third, giving rights to independent inventors can induce patent holders to license ex post on terms that reduce market price, in order to discourage ex post entry through independent invention (Maurer and Scotchmer, 2002). Suppose that a single patent holder is in the market. Then, whether or not the patent holder licenses, a right of independent invention will reduce the price in the market below the proprietary price. Without licensing, the price will fall due to entry by independent inventors. Instead, the patent holder can license at a fee equal to the cost of independent invention. Then independent inventors are indifferent to paying the license fee or paying the costs of independent invention, but the patent holder prefers licensing. The price reduction in the market (determined by the terms of license and number of licensees) must be large enough to deter further entry.

The market price with licensing will thus depend on the cost of independent invention. If the cost is high enough, the right of independent invention can benefit users without undermining the incentive to invent. In fact, in plausible models, the cost of independent invention only needs to be greater than half the cost of the original innovator (Maurer and Scotchmer, 2002; Ottoz and Cugno, 2004). Nevertheless, Blair and Cotter (2002) rightly point out that the economic consequences depend critically on the relative costs of first inventors and imitators, which will differ across technologies. Giving a right of independent invention can have harmful consequences if imitation or independent invention is too cheap.

Lichtman (1997) made a similar argument in the context of unpatented inventions, advocating on grounds of cost that independent inventors be allowed to copy but not clone them. Armond (2003) proposed that independent discovery be available as a defense to a preliminary injunction motion.

Although independent inventors are not generally exempted from liability under U.S. patent law, the law does, in fact, recognize user rights in two circumstances: (1) prior secret use of business methods—as a narrow statutory exception with regard business method patents (17 U.S.C. §273); and (2) shop rights—under state law governing employment agreements and the employment relationship, an employer obtains a royalty-free, non-exclusive, non-transferable license to use an employee's invention where the employee makes a patented invention using the employer's facilities. In most research environments today, employers require employees involved in research-related activities to assign their inventions to the employer, although some state laws limit such agree-

ments to inventions developed within the scope of employment or developed using the employer's facilities (e.g., Cal. Lab. Code § 2870). Even where no express agreement has been signed by an employee, patents invented by the employee may nonetheless be deemed to have been assigned where an employee has specifically been employed to invent in the field in which the invention was made. In these circumstances, a court may imply an assignment clause into the employment contract.

Rights after sale. Under what is commonly referred to as the "first sale" or "exhaustion" doctrine, the intellectual property owner "exhausts" the legal monopoly in a product by selling it to the public, thereby enabling the purchaser to use the work and resell it without infringing. Such a default right structure reduces transaction costs for subsequent transactions. Similarly, purchasers of patented products are deemed to have an implied license to make repairs, although this license does not extend to "reconstruction" of the patented product. Intellectual property owners can, subject to anti-competitive restrictions, circumvent the first sale doctrine by imposing licensing restrictions upon the conveyance of a product.

Rights to share copyrighted works. Even though the purpose of copyright law is to prevent copying, a controversial idea that keeps resurfacing is that copying or sharing is less harmful to creators than meets the eye, at least where the sharing of each legitimate copy is limited. Where it is unlimited, such as in peer-to-peer networks or when users make copies of copies, sharing poses a greater threat to appropriability.

The argument is that the proprietor will price in a way that anticipates sharing. Sharing allows the proprietor to charge a higher price, since demand is determined by the willingness to pay several parties. Limitations on sharing may arise because copies of copies degrade [Liebowitz, 1982, 1985; Liebowitz and Margolis, 1982 (arguing in the era of analog copies)] or because it is less costly to facilitate sharing than to produce a copy for every user, as in a video rental market (Varian, 2000), or because the probability of detection increases with the size of the sharing group.

The earlier set of papers in this vein relied on the fact that copying is costly. Novos and Waldman (1984) and Johnson (1985) argued that proprietors may reduce price to avoid copying, but that the cost of copying will nevertheless preserve the proprietor's market. The market price will be lower than without copying, reducing the deadweight loss of excluding users, but the per-period reward to creative works will also be reduced, especially when there is heterogeneity in tastes as well as in copying costs. The welfare effects are different according to whether the cost of copying is per-copy or per-user, as when it requires the purchase of a copying device. Scholars have also argued that copying can have an affirmative benefit for rightholders because it builds network effects (Conner and Rumelt, 1991; Shy and Thisse, 1999).

A second set of papers focus on the fact that prices can be tailored to the groups that form. Liebowitz (1985) emphasized price discrimination according to whether the purchaser will make the copy available to many users, as libraries do. See also Ordover and Willig (1978). Bakos, Brynjolfsson, and Lichtman (1999) argued that, depending on the groups, sharing might actually be more profitable than selling to individual users.

This is true if, first, the willingnesses to pay within groups are negatively correlated, or, second, if there is variance in the sizes of groups. Thus, whether sharing enhances profit depends on what governs group formation. However, Scotchmer (2005a) argued that sharing groups will not be formed exogenously or even randomly, and if they form in a way that is efficient for the group members conditional on the proprietors' prices, then group formation has no effect at all on profit opportunities. Sharing is neither profit-reducing nor profit-enhancing.

Given that copying can have salutary effects as well as deleterious effects, these arguments have led authors to consider an additional set of policy levers specific to the copying context, such as taxes and subsidies on prices of legitimate copies, and taxes and subsidies on copying devices, as well as the optimal mix of enforcement activities and other incentives. See Besen and Kirby (1989a, 1989b); Chen and P'ng (2003); Netanel (2003); Fisher (2004).

1.3.1.5. Remedies As in other bodies of law, the remedial opportunities in intellectual property law are injunctions and damages. There are two branches of thought about the relative efficacy of these rules, one branch focusing on whether remedies will lead to efficient use of the property ex post, and the other branch focusing on the ex ante effects.

The first set of arguments (Calabresi and Melamed, 1972; Polinsky, 1980; Kaplow and Shavell, 1996) for the general framework, and Blair and Cotter (1998) for the intellectual property context considered whether property rules (injunctions) are more or less likely than judicially imposed liability to encourage bargaining to an efficient outcome ex post. For example, property rules (injunctions) may be preferred when transaction costs of exchange are low and the costs of valuing violations of rights by courts are high. For the intellectual property context, Ayres and Klemperer (1999) add the consideration that "soft" remedies, which do not actually restore the proprietary price, can be socially beneficial because they increase consumers' surplus without impinging much on profit, at least for small price reductions.

The second set of arguments are not concerned with what would happen in the out-of-equilibrium event of infringement, but focus instead on how potential remedies affect equilibrium profits and the ex ante incentives for R&D (Schankerman and Scotchmer, 2001; Anton and Yao, 2007). In these arguments, remedies are only important because they do or do not deter infringement, and because they determine the terms of an ex ante license. The terms of license that will be accepted by a potential licensee/infringer depend on the consequences for infringement, and this threat has an affect on the ex ante division of profit. Schankerman and Scotchmer argue that if infringement leads to profit-eroding competition between the infringer and rightholder, a wide range of remedies will deter infringement, at least for stand-alone innovations, and are therefore equivalent from an ex ante point of view. However, this is not necessarily true for research tools and other potentially licensed intellectual property where infringement does not dissipate profit (see below).

Merges and Duffy (2002) and Blair and Cotter (1998) argue, again from the ex post perspective, that patent and copyright law are better suited to a property-rule para-

digm than a liability-rule paradigm. Since intellectual property rights are relatively well defined, disputants or potential disputants should have little trouble resolving their differences by negotiating licenses against the backdrop of an injunction. In contrast, if the setting of damages (an ex post "compulsory license") is left to a generalist judicial institution under a liability rule, the court may have difficulty placing a value on the intellectual property or on the injuries caused by infringement. Further, judicially imposed licenses can undermine the prospect function of patent law (Kitch, 1977). Merges (1996) argues that for complex transactions involving many players, a property rule will facilitate the creation of private exchange institutions, such as patent pools, that can evolve in response to changing circumstances and draw upon industry and institutional expertise.

Although infringed rightholders generally have a right to enjoin unauthorized use, injunctions are backed up by compensatory damages for past violations. These may include enhanced (punitive) damages for patent infringement, statutory damages for copyright infringement, and attorney fees and costs in "exceptional cases." In several areas such as the covers of musical compositions, juke boxes, cable television broadcasts, and webcasting, copyright law provides for compulsory licensing. These regimes arguably economize on transaction costs, although commentators are divided on the economic effects of such compulsory licenses. Cf. Merges (1996) and Netanel (2003).

Consistent with traditional economic analysis of damages (harm internalization), patent and copyright law award intellectual property owners the greater of lost profits or a reasonable royalty for the defendant's unauthorized use of the protected works (Blair and Cotter, 1998). Calculating these measures, however, is quite complex, involving numerous subtle determinations of how markets would have evolved had infringement not taken place (Blair and Cotter, 2001).

It follows from general economic principles that enhanced damages should be awarded where improper behavior is costly to detect and where full compensatory damages are costly to prove (Polinsky and Shavell, 1997; Cooter, 1982; Cotter, 2004). Excessive damages (i.e., where expected damages exceed actual damages) could lead to overdeterrence in the sense that parties may exercise caution in order to avoid a risk of liability. Several recent studies indicate that courts may well be overdeterring patent infringement based on the high rate of enhanced damage awards [Federal Trade Commission, 2003; Cotter, 2004; Moore, 2000; (finding that requests for enhanced damages were fully litigated in nearly half (45%) of cases and awarded in 64% of cases between 1983 and 1999, thereby resulting in awards in 29% of all cases)].

Due to the very different nature of trade secret protection, the remedies available for unauthorized use and dissemination of a trade secret are more limited. Where the secret has not been disclosed to the public, courts will generally enjoin further use of the secret by a misappropriating entity. But where the secret has been disclosed, the trade secret owner will be limited to damage remedies or limited injunctive relief against the entity that misappropriated the trade secret (such as a "head start" injunction which excludes the misappropriator from the market for a designated period). Disclosure to the public destroys the secret and therefore it would be inappropriate (and infeasible) to enjoin use

of the information by others. Nonetheless, Sidak (2004) argues an injunction against a misappopriating entity should be perpetual in order to encourage efficient post-litigation bargaining over the value of continued use. Such a rule would also avoid the expense and difficulty (error costs) of having courts adjudicate the option value of a trade secret.

1.3.1.6. Channeling doctrines The various modes of intellectual property provide overlapping protection. For example, patent, copyright, and trade secret law all cover computer software. As noted earlier, however, copyright doctrines exclude ideas, processes, and methods of operation from copyright protection. Thus, software developers cannot gain copyright's long duration of protection for the functional aspects of computer software. Such inventions must comply with patent law's formal examination requirements and surpass patent law's higher thresholds in order to obtain legal protection. In this way, intellectual property law prevents inventors from obtaining protection for functional features through the "backdoor" of the copyright system.

The relationship between patent and trade secret law is somewhat more complicated. Both regimes cover technological innovation. Where an inventor obtains a patent before a subsequent researcher invents the same technology, the patent trumps the subsequent inventor, regardless of whether the subsequent inventor seeks to protect the invention as a trade secret. (Such secrecy may well conceal infringement, particularly in the case of process inventions, but that does not suggest that the trade secret would have any validity vis á vis the patent.)

A somewhat more complicated issue arises where the first inventor chooses to protect the technology as a trade secret. If a subsequent inventor independently discovers the same invention and obtains a patent, two overlap issues arise: (1) does the trade secret invalidate the patent (on novelty grounds); and (2) if the patent is valid, can the trade secret owner continue to practice the invention—in essence, does the first inventor enjoy a prior user right. As suggested earlier, the trade secret will not invalidate the patent because it does not fall within the body of prior art that may be considered in judging novelty. Therefore, assuming the second inventor meets the other requirements of patentability, she will obtain a valid patent. As regards the rights of a trade secret owner, U.S. patent law holds that the trade secret owner infringes upon the patent by continuing to practice the invention. Some nations recognize a prior user right in this circumstance, which places the technology under duopoly rather than monopoly control. The profits available to the patentee are reduced accordingly. It can be argued, however, that such a structure of rights might partially improve the screening function of patent law—inventions that have been independently developed may not have needed as much of an ex ante incentive in the first place. To the extent that ex ante incentives are more than sufficient to generate the innovation, duopoly improves social welfare by reducing the deadweight loss from exclusive exploitation.

1.3.2. Cumulative innovation

In the context of stand-alone inventions or creations, intellectual property rewards reflect the social value of the contribution, since the profit is determined by demand. That is one of the main virtues of intellectual property as an incentive system. However, when innovation is cumulative, the most important social benefit of an innovation may be the boost given to later innovators, and this may make the benefits harder to appropriate (Scotchmer, 1991). Moreover, the innovation may enable rivals to enter with improved products. In that case, social success may mean private failure: the boost given to the rivals may cause the innovation's own demise (Scotchmer, 1991, 1996; Green and Scotchmer, 1995; Chang, 1995; O'Donoghue, Scotchmer, and Thisse, 1998; O'Donoghue, 1998; Besen and Maskin, 2002; Hunt, 1999, 2004). Merges and Nelson (1990) give a rather opposite perspective on the cumulative problem. Instead of worrying that later improvers pose a threat to earlier innovators, they worry that earlier innovators (earlier patents) pose a threat to later improvers.

The intellectual property system must resolve these contradictions. In general, the problem of appropriating benefits has two facets: the overall level of profit, and how it is divided among the sequential innovators. As we will see, the roles of the policy levers are closely intertwined in the cumulative context, and the best design of the system will depend on the transaction costs of licensing.

Many scholars have emphasized the importance of cumulativeness in the process of knowledge creation, especially economic historians. As expressed by David (1985, p. 20), "Technologies ... undergo ... a gradual, evolutionary development which is intimately bound up with the course of their diffusion." Secondary inventions—including essential design improvements, refinements, and adaptations to a variety of uses—are often as crucial to the generation of social benefits as the initial discovery. See, e.g., Nelson and Winter (1982); Taylor and Silberston (1973); Mak and Walton (1972); Rosenberg (1972). Cumulative technologies tend to involve multiple components, serve as building blocks for further incremental innovation, and often spur wide-ranging applications. Automobiles, aircraft, electric light systems, semiconductors, and computers fall within this category. Some chemical technologies are hybrids of discrete and cumulative models. New chemical compounds are typically discrete in terms of the product market that they serve, but can suggest promising new lines of research (e.g., penicillin, Teflon) (Nelson and Winter, 1982). The biotechnology field reflects several cumulative features. The development of research tools provides the means for decoding genomic information. Research decoding genomes provides the input for downstream biomedical research.

Cumulativeness also extends to expressive creativity. All authors and artists draw, to some extent, on prior works. Sequels, translations, and screenplays build directly upon prior works. Parodies and satires comment on or employ other works. Most musical compositions reflect rhythm and other elements of established genres. The hip-hop and rap musical genres embody prior recordings through the use of digital sampling.

Computer software, which is written work intended to serve utilitarian purposes, falls between the technological and expressive realms. Cumulativeness plays a particularly important role here, whether in operating systems, technical interfaces, peripheral devices or application programs (Menell, 1987, 1989; Lemley and O'Brien, 1997).

1.3.2.1. A preliminary model: the virtues of licensing One of the lessons that emerges powerfully below is that the best design of the intellectual property system depends on how fluid the market for licenses is. Before turning to a more detailed analysis of design issues and how licensing affects them, we illustrate the importance of licensing by modifying the "reduced form" model of Landes and Posner (2003), where a variable z is taken as the "strength" of a right. For example, the strength of the right may be affected by breadth or exemptions such as fair use.

Let $q(\cdot)$ be the demand function for a protected innovation, where $q(p)$ is decreasing with p. Referring back to our discussion of copying, and how the threat of copying affects the market price, let $y(p, z)$ be the supply of illicit copies. Then the net demand faced by the proprietor is $q(p) - y(p, z)$. Assuming for convenience that the marginal cost of copies is zero, the proprietor maximizes $p[q(p) - y(p, z)]$, and sets a profit-maximizing price $p^*(z)$ that depends on the strength of protection through the threat of copying.

Now consider how the profit-maximizing price and the proprietor's profit depend on the strength of protection, z. Assume that the supply $y(p, z)$ of illicit copies increases with p and decreases with z. Because the supply of imitations $y(p^*(z), z)$ depends on the proprietor's price as well as the level of protection, there is a potential indeterminacy in the model. An increase in protection could conceivably lead to a decrease in price and an increase in illicit copying. However, under reasonable assumptions we can assume that the profit-maximizing price increases with the level of protection, even though the increase in price has a feedback effect of increasing imitation or copying.

Nevertheless, the profitability of creations and hence their supply will not necessarily increase with the level of protection z. This is because the cost of creation can also depend on z, e.g., by creating a burden to innovate outside the scope of other rights. Suppose, in fact, that each potential innovator faces an R&D cost k plus additional costs $e(z)$ that reflect the burdens imposed by other intellectual property rights. The creation is profitable if

$$p^*(z)[q(p^*(z)) - y(p^*(z), z)] - k - e(z) > 0.$$

If potential creations differ in their markets (for example, if we introduce a quality variable s into the demand function q), then the more valuable ideas will call forth investment, while the less valuable ones will not. Without formalizing this idea, we will let $N(z|k)$ describe the number of profitable creations with cost k. The supply of new creations $N(z|k)$ is not monotonic in the strength of protection z because raising z increases the creator's costs. The punch line of this model is that too much protection can be bad for creators as well as for imitators.

The point we would like to make, however, is that the punch line is largely reversed if firms can license to avoid conflicting property rights, rather than being forced into the costly activity of avoiding them. Following the perspective in O'Donoghue, Scotchmer, and Thisse (1998), assume that each innovating firm will initially be in the position of paying license fees on the discoveries of its predecessors, and then in the position of collecting license fees from its followers. The effect of strong rights, z, is to increase the licensing obligations, rather than to increase the real resource cost of avoiding prior rights. The essential point is that licensing also creates claims against future innovators.

Suppose, in fact, that all innovators are in symmetric positions: they pay for the same number of licenses, and are paid by the same number of licensees. Then, since all the money must go somewhere in the end, symmetry means each innovator pays as much in licensing fees as it earns in licensing fees, say, $\ell(z)$ on both sides of the ledger. The profit-maximizing decision is then to invest in the potential creation if

$$p^*(z)[q(p^*(z)) - y(p^*(z), z)] + \ell(z) - k - \ell(z) > 0$$

or equivalently,

$$p^*(z)[q(p^*(z)) - y(p^*(z), z)] - k > 0$$

Hence (assuming that the revenue $p^*(z)[q(p^*(z)) - y(p^*(z), z)]$ is increasing in z), the ambiguous effect of strong protection on innovation disappears. A strengthening of protection leads to more creations.

Thus, with licensing, we are cast back to the same consideration as with stand-alone inventions, namely that there is a tradeoff between deadweight loss and innovation, but no tension between protecting early innovators and protecting the later innovators who use the knowledge they create. Licensing will largely resolve that tension, to everyone's benefit.

These points about the salutary effects of licensing have mostly been made in models that distinguish between the policy levers of intellectual property, rather than lumping the policy levers into a single variable called the "strength" of the right. We now turn to that more disaggregated discussion.

1.3.2.2. Duration Although the length of protection has an obvious effect on the overall level of profit, the *statutory* length may be irrelevant. When innovations are under threat of being supplanted by improved innovations, market incumbency only lasts until that happens. O'Donoghue, Scotchmer, and Thisse (1998) define a notion of "effective patent life" that focuses on the rate of market turnover, and argue that the effective life of the patent may be determined by the breadth of the right, rather than its statutory length. This is because breadth determines how long it will take before the product is supplanted (see below).

In fact, there is considerable evidence that the "effective" lives of most patents are shorter than their statutory lives. Mansfield (1986) reported, using survey evidence, that in some industries 60% of patents are effectively terminated within four years. The literature on patent renewals (Pakes and Schankerman, 1984; Schankerman and Pakes,

1986; Pakes, 1986; Schankerman, 1998; Lanjouw, 1998) carries a similar message. For example, Schankerman (1998) reports that half of patents in France are not renewed beyond the tenth year, even though renewal fees are very low.

Besen and Maskin (2002) present a model of sequentialness where this endogeneity of patent life is absent (because the products do not compete in the market) but argue that statutory life should be shorter when innovators learn from previous innovators. Their model has sequentialness in innovation (because innovators learn from each other) but the resulting products are "stand-alone" and live out their statutory lives. They argue that the optimal statutory life should be shorter if innovators learn from each other than if not, because the loss from impeding future innovation is greater. (This result depends sensitively on the absence of licensing; see below for an argument that all results in this arena turn on what is assumed about licensing.)

For the case of basic and applied research, Green and Scotchmer (1995) argue that patent lives must last longer if the research is divided between sequential innovators rather than concentrated in a single firm, because of the problem of dividing profit in a way that respects the costs of both parties.

To the extent that transaction costs may impede licensing and first stage inventors do not need large ex ante rewards to induce innovation, then a shorter duration of intellectual property protection promotes cumulative innovation. Legal protection for computer software fits this profile (Menell, 1987). There are relatively strong non-intellectual property incentives for developing operating system and other platform technology for many product markets. Interoperability with widely adopted platforms is often critical to secondary innovation, such as application programs and peripheral devices. Owners of the intellectual property rights in widely adopted proprietary platform technologies can exercise tremendous market power due to network effects and consumer lock-in. Shortening the duration of protection for such technologies is one mechanism for constraining such market power and better equilibrating the incentives of first and second generation innovators.

1.3.2.3. Threshold requirements and breadth The status of an invention that builds on other inventions can be: (1) protected and noninfringing, (2) unprotected and noninfringing, (3) protected and infringing, or (4) unprotected and infringing (Scotchmer, 1996; Denicolò, 2002a). Thus, the economic effects of threshold (patentability) and breadth are hard to disentangle. Scenario (1) gives the best incentive for second generation innovators, but does not force the second-generation innovator to share the profit with the prior innovator. Scenario (2) will clearly stymie second-generation innovation unless there is a mechanism other than intellectual property to protect the innovator. In scenario (3)—which is possible under patent law—the works are considered "blocking": the later work infringes the prior innovation and cannot be exploited without a license; and the later work is protected and cannot be exploited by the pioneering inventor without license. Such a scenario encourages the inventors to share profit from the subsequent invention. Scenario (4), which approximates the treatment of derivative works under copyright law, discourages improvements or adaptations by subsequent

inventors in the absence of ex ante bargaining. However, there is less difference between scheme (3) and scheme (4) than meets the eye. Even if the later product is not protectable, it can be protected by an exclusive license on the innovation it infringes.

The literature draws widely differing conclusions about the optimal way to organize the rights of sequential innovations, largely because authors make different assumptions about when and whether licenses will be made, and who can be party to the negotiation. Kitch (1977) was the earliest, and perhaps most extreme, licensing optimist.

Licenses can be made either *ex ante*, before the follow-on innovator invests in his project, or *ex post*. If licenses must be negotiated *ex post*, after both innovations have been achieved, scheme (4) may stifle innovation, since the second innovator will fear that the first innovator will simply appropriate it (Green and Scotchmer, 1995). On the other hand, if the second innovator can approach the first innovator for an *ex ante* license before investing in his idea, the second innovation is not in jeopardy under either of scheme (3) or (4). However, the first innovator will typically collect more of the profit in scheme (4) because the second innovator will have less bargaining power (Scotchmer, 1996).

Denicolò (2002a) considered a model where ideas are common knowledge, and asked how the various scenarios affect patent races, assuming that there will be *ex post* licenses, but no *ex ante* licenses. He finds that the choice should depend on the relative costs of the innovators. If, for example, the cost of the first innovation is low and the cost of the second is relatively high, it may be better not to let the first innovator share in the second innovator's profit. Of course, this also depends on whether the first innovation can earn profit in the market, or only through licensing. Chang (1995) also considered a context where the firms could make *ex post* licenses, but not *ex ante* licenses, and concluded that the choice between (1) and (3) should depend on the relative costs of the innovators.

The worst situation is when licensing may fail entirely, and when, in any case, the earlier innovator does not need to profit from licensing in order to cover his costs. Merges and Nelson (1990) draw on many actual examples to argue that such circumstances are quite plausible. A defect of the one-size-fits-all intellectual property regime is that it cannot distinguish cases where blocking rights are unnecessary for cost recovery from cases where earlier innovators would not invest unless they can profit from licensing. In part on this basis, Burk and Lemley (2002) argue that it would be better to make intellectual property protection more finely attuned to industrial contexts.

But even in those cases where earlier innovators should be allowed to profit from the later innovations they enable, blocking rights are a blunt instrument for dividing profit. Profit shares will not necessarily reflect the cost shares. This is especially true if the licenses will be negotiated *ex post*, after all innovators' costs are sunk (Green and Scotchmer, 1995), although Chang (1995) argues that the problem can be mitigated by making the infringement (breadth) responsive to cost. In copyright law, blocking rights are not available as a tool at all. Follow-on creators may not prepare infringing derivative works without permission of the owner of the copyright in the underlying

work. Copyright law therefore has less flexibility than patent law in how it balances the incentives for sequential innovators (Lemley, 1997).

In the case of product improvements, there is a question of whether a patent's breadth can extend to slightly better products that have not yet been invented, or only to inferior products. O'Donoghue, Scotchmer, and Thisse (1998) define a notion of "leading breadth" that determines when there will be blocking rights in the cumulative context, and also establishes the "effective life" of the patent right. To see why leading breadth is useful as a policy lever, suppose to the contrary that every trivial infringement is noninfringing, even if patentable. A potential improver may discard ideas for small improvements because they lead to price-eroding competition between close, vertical substitutes. It is only the relatively big ideas that will become innovations. This problem can be solved by making the small improvements infringing. Firms may then be willing to invest in them, since control of the improvement and its predecessor can then be consolidated through licensing in the same firm. Instead of competing, both will be marketed together. Further, if the small improvements are infringing, and if it takes a relatively long time for large ideas to come along, the "effective life" of each patent is prolonged. These effects cannot be achieved by choosing the patentability standard alone; the opportunity to consolidate successive improvements in the hands of a single firm arises because the patents are infringing.

In the "ideas" model, there is not much role for a nuanced patentability standard. However, in a "production function" model, a patentability standard can make each successive innovator more ambitious in the size of improvement he invests in. This may be socially beneficial (O'Donoghue, 1998).

Hunt (1999, 2004) presents a production-function model motivated by the Semiconductor Chip Protection Act, in which eligibility for protection coincides with noninfringement of previous innovations (as with copyright). He argues that the standard for protection should be increasing in the "dynamicness" of the industry. Since noninfringement coincides with protection, it is hard to sort out their respective roles. Whereas the effective patent life is determined by breadth in the model of O'Donoghue, Scotchmer, and Thisse (1998), it is determined in the Hunt model by both.

Finally, we return to the idea of Kitch (1977), who argued that strong patent rights should be given to pioneers, as he calls them, so that the pioneers can coordinate the subsequent development of the technology. These are called "prospect" patents. The theory is not focused on the reward purpose of the patent, since it would apply even if the pioneer innovation were costless to achieve, and no incentive for R&D were required. The theory thus rejects the line of reasoning that says intellectual property is at best a necessary evil, due to deadweight loss.

The prospecting theory rests on the premise that social interests and the private interests of the patentholder are aligned. Scotchmer (2005b, Section 5.6) shows that this may be true in some ways, but is not true in other ways, and in particular, that strong pioneer patents can pre-empt competition policy. As in later theories of cumulativeness, the prospector's profit comes from getting the intellectual property into use. For this reason the pioneer has an incentive to encourage use at a fee. Further, the pioneer can

profit from delegating research effort to the most productive researchers, and avoiding bad projects, as are socially efficient.[3] These are ways in which the pioneer's interest is aligned with the broader public interest.

However, the pioneer can preempt competition policy in two ways: by avoiding competition in the "innovation market" for second-generation products, and by avoiding competition among second-generation innovators once the second-generation innovations exist. The first of these may or may not be socially harmful (see below, the section on the patent/antitrust conflict), but the second is clearly harmful to competition, assuming that patent law is well designed in the first place. If these harms to competition are important, it might be better to avoid pioneer patents even if second-generation innovators must then duplicate the cost of achieving the pioneer innovation. As with all intellectual property, the case for pioneer patents is strongest if the pioneer innovation is costly to achieve, and the patent is actually needed as a reward.

For radical improvements that read on existing patents, Merges (1994) suggests that it may be socially advantageous to exempt an improver from infringement under the "reverse doctrine of equivalents." Under this doctrine, a radical improvement may be deemed non-infringing even though it literally reads upon an existing patent. The doctrine would allow the radical improver to avoid a holdup by the underlying patent holder and to avoid a potential bargaining breakdown. See also Lemley (1997). This is again an argument that relies on difficulties in licensing. Such a doctrine avoids over-rewarding a first-generation inventor (who did not foresee a much greater advance) while providing strong encouragement to visionary subsequent inventors. The rule has rarely been applied in actual cases, although the possibility of its application may well have fostered licensing.

Evolving doctrines regarding subject matter. We turn now to how these economic arguments have been reflected, if at all, in legal doctrines. Notwithstanding the general applicability of the patent system to "anything under the sun made by man," [Diamond v. Chakrabarty (1980)], courts have, since the early history of patent system, barred inventors from claiming patents on natural physical phenomena (e.g., the properties of lightening), laws of nature (e.g., the theory of general relativity), mental processes, and abstract intellectual concepts (e.g., algorithms) [Gottschalk v. Benson (1972)]. Courts have noted that allowing exclusive rights for such fundamental discoveries would unduly impede future inventors—a cumulative innovation rationale [O'Reilly v. Morse (1854)]. Implicit in this justification is the notion that transaction costs could impede licensing. The courts have thus realized, as is more explicit in the economic models discussed above, that licensing plays an important role in balancing the rights of sequential innovators.

Court decisions over the past 25 years scaling back if not effectively removing the traditional jurisprudential limitations on patentable subject matter have led scholars to

[3] In fact this is always true if the prospector can make deals with consumers as well as follow-on inventors, for example, if consumers can pay the prospector not to retard progress.

consider the merits of imposing categorical subject matter exclusions, the development of sui generis protective regimes tailored to particular technological fields, as well as technology-specific rules within the patent system in order to better promote cumulative innovation—particularly in the software, bioscience, and business method fields (Menell, 1987). Samuelson et al. (1994), and Cohen and Lemley (2001) have argued that in environments of rapid turnover where costs are relatively low, for example, computer software, strong intellectual property rights can impede the rate of technological advance. As we have seen, this depends on both the design of the rights and the ease of licensing. Heller and Eisenberg (1998) have argued that patenting of gene sequences generates a tragedy of the anticommons, a fragmentation of rights which vastly increases transaction costs, thereby impeding downstream research for medical advances. This is also, at root, an argument about the ease of licensing [see Walsh, Arora, and Cohen, 2003 (reporting survey data revealing indicating that university research has not been impeded by concerns about patents on research tools as a result of licenses, inventing around patents, infringement (often informally invoking a research exemption), developing and using public tools, and challenging patents in court)].

The opening of the "business method" patent flood gates has raised concerns about whether patent protection is needed at all to promote such innovation and, more troubling, whether such protection is chilling innovation and competition. The idea of protecting business plans runs counter to a core premise of the free market system by offering a form of antitrust immunity for business models. As the rising tide of prior art raises the threshold for protection, such adverse effects may abate and innovation could well produce valuable new business methods. Nonetheless, several scholars find that the costs of extending patent protection to business methods as a class significantly outweigh the benefits (Dratler, 2003; Dreyfuss, 2000), especially when pro-patent biases and patent quality considerations are factored into the analysis (Meurer, 2002; Merges, 1999a). Drawing upon the approach of foreign patent systems, Thomas (1999) calls for limiting the subject matter of patent law to fields of industrial applicability.

Copyright law applies a different rights structure than patent law, with significant implication for the direction of cumulative innovation (Lemley, 1997). Unlike patent law, which allows anyone to patent improvements to patented technology, copyright owners have the exclusive right to prepare derivative works. Therefore, a novelist can prevent others from translating their work into another language, adapting the story for the stage or a motion picture, selling the story in narrated form (e.g., books on tape), and developing sequels that draw extensively on the protectable elements (including possibly character names and attributes—e.g., Rocky IV or the next James Bond film). As we will see below, some borrowing is tolerated under the fair use doctrine, but pioneers generally have exclusive authority to pursue the further development of their expressive work. Copyright law wholly excludes protection for ideas and functional attributes of a work but protects creators against direct or near exact copying of even a significant fragment of the whole for a tremendously long duration (life of the author plus 70 years), reflecting the notion that society prefers to have one hundred different war novels embodying similar themes, ideas, and facts than one hundred versions of "War and Peace"

that differ only in their final chapter. Consequently, copyright protection for an author's *expression* of ideas and the relatively long period of its duration effectuates a different balance than patent law. Patent law encourages cumulative innovation by allowing follow-on to secure rights on improvements and to enable any competitor to build upon the innovation in its entirety within a comparatively short period of time (20 years from the time of the application). By contrast, copyright law, with its narrow scope of protection, allows subsequent creators to pursue competing works using the same ideas as the "pioneer," but allows the pioneer exclusive rights over the development of the expressed ideas (Karjala and Menell, 1995).

Adequacy of disclosure requirements. Both patent law and copyright law provide for the disclosure and dissemination of knowledge, which promotes cumulative innovation. Patent law requires disclosure as the quid pro quo for patent protection. In the software field, however, disclosure of source code is not required under the best mode requirement—the inventor need only disclose the functions of the software on the grounds that a person of ordinary skill in the art could write software to perform the functions without undue experimentation [Fonar Corp. v. General Electric, 107 F.3d 1543 (Fed. Cir. 1997)]. In practice, however, the knowledge protected by many software patents would be difficult and costly to decipher without access to the source code, which is usually maintained as a trade secret (Burke, 1994; Burk and Lemley, 2002). This slows the process by which follow-on inventors can build upon earlier generations.

In the case of most copyrighted works, the knowledge contained in the work may be comprehended from direct inspection. Therefore, the publication of a work discloses and disseminates. Furthermore, the Copyright Act requires those who register a work, which is optional, to deposit a copy with the Library of Congress.

Some copyrighted works, however, do not lend themselves to visual inspection and comprehension. Piano rolls, for example are not human readable, although even that technology can be deciphered by the trained ear or a mechanical translation system. Musical or audiovisual works stored on magnetic tape and digital media can also be perceived directly. Computer software, however, cannot typically be perceived unless it is available in source code. Copyright Office regulations, however, do not require disclosure of the entirety of source code in order to secure copyright registration. Thus, as with patent law, copyright law allows protection of software without providing access to the underlying knowledge. As a result, follow-on invention is stifled, although such rules may deter infringement that would otherwise be difficult to detect.

1.3.2.4. Rights of others (and defenses) Several doctrines provide safety valves, beyond the limitations embodied in scope of protection, for promoting cumulative innovation. These include the experimental use doctrine (patent law), the fair use doctrine (copyright law), and the reverse engineering doctrines (of trade secret and copyright law). In addition, copyright law provides for several exemptions for educational and related purposes which can be viewed as promoting basic education for new authors and artists.

Experimental use. A subsequent inventor who wants to improve a patented technology may benefit from experimenting with it. U.S. patent law has had a common law exemption for "philosophical experiments" and research to ascertain "the sufficiency of [a] machine to produce its described effects" [Whittemore v. Cutter, 29 F. Cas. 1120 (C.C.D. Mass. 1813)].[4] Subsequent cases, however, have declared the defense to be "truly narrow" and applicable solely to activities "for amusement, to satisfy idle curiosity, or for strictly philosophical inquiry," [Roche Prods., Inc. v. Bolar Pharm. Co., 733 F.2d 858, 863 (Fed. Cir. 1984); Madey v. Duke Univ., 307 F.3d 1351 (2002), *cert. denied* 123 S. Ct. 2639 (2003)], prompting several scholars to warn that patent law unduly hinders academic and basic research and unduly supplants academic and scientific norms promoting progress, disclosure, and cumulative innovation.

Eisenberg (1989) proposed to exempt research that could potentially lead to improvements or design-arounds of patented technology, but she also pointed out the inherent contradiction that arises when further research is the main use of the patented invention. A broad exemption could entirely undermine the profitability of the patent. However, the effect of a research exemption depends on an ancillary doctrinal question, namely, whether the invention that is achieved by using the prior invention will infringe the prior patent (Scotchmer, 2005b, Chapter 5). If so, the exemption may (counterintuitively) increase the patent holder's profit. Exercising the research exemption can put the improver in the position of bargaining for a license ex post (after he has sunk his costs) rather than ex ante. This strengthens the bargaining position of the first patentholder.

Kieff (2001a) contends that the exclusive rights provided by patents promote university research by increasing private investment in research and improving the efficiency of academic research environments. His analysis does not, however, directly address whether a more permissive experimental use doctrine would adversely affect the flow of private research investment into universities. Based on survey research, Walsh, Arora, and Cohen (2003) find little evidence that patents on research tools have significantly impeded university research. There is one notable exception: the field of genetic diagnostics. Cai (2004) notes that the chilling effect of a narrow experimental use defense may not be very significant due to patent holders' rational forbearance in enforcing against universities as well as legal constraints (sovereign immunity) on suing state actors. Cf. Menell (2000).

Many legal scholars in addition to Eisenberg have proposed alterations to existing law. Mueller (2001) endorsed a broadened experimental use defense along the lines of the European system as well as compulsory licensing. Feit (1989) proposed a compulsory license for patents infringed during experimentation to improve patented technology. These authors, like Eisenberg, raise particular concerns about patents on upstream research tools, particularly in the bioscience field. O'Rourke (2000) proposes a fair-use doctrine for patents that would go beyond the similar doctrine for copyright by allowing courts to judge permissible conduct and impose compulsory royalties. Dreyfuss

[4] Article 27(b) of the European Patent Convention exempts "acts done for experimental purposes relating to the subject matter of the patented invention."

(2003) proposes to allow experimental use if the investigator's institution promptly waives patents on subsequent discoveries, subject to a "buyout" provision. In cases where a patentee has refused to license to a non-profit on reasonable terms, Nelson (2003) proposes a research exemption, provided the researcher agrees to publish his results and agrees either not to patent his own results or to license them on nonexclusive and reasonable terms. Strandburg (2004) proposed to exempt improvement patents and to provide for compulsory licensing of research tools. Other authors (Epstein, 2003; Merges, 1996) have countered that such proposals would entail significant administrative costs and have complex effects upon licensing markets and the formation of licensing institutions.

Fair use. The fair use doctrine in copyright law exempts a user from liability for infringement when copyrighted works are used for criticism, comment, news reporting, teaching, scholarship, and research (17 U.S.C. §107). In applying this doctrine, courts balance the purpose and character of the use (including whether such use is of a commercial nature or is for nonprofit educational purposes), the nature of the copyrighted work, the amount of copying, and, most importantly, the effect of the use upon the potential market for or value of the copyrighted work. Transformative (or more radical "improvements") are more likely to be permissible, whereas uses that merely supplant the underlying work are disfavored. In this way, the fair use doctrine promotes significant creative advances while protecting the pioneer from direct market competition. The fair use doctrine can also be seen as an efficient means for permitting uncompensated use of copyrighted material where the transactions costs of licensing or other means of exchange would prevent a transfer through the market (Gordon, 1982).

Courts have applied the fair use doctrine to enable software developers to make copies of protected programs for purposes of learning how such software functions (Samuelson and Scotchmer, 2002). With such code deciphered, the rivals could discern unprotectable elements (e.g., interoperability specifications) which they were then free to incorporate in their own commercial products. In this way, the fair use doctrine operates as a form of "experimental use" exemption.

Reverse engineering. As with copyright law, trade secret law allows others not bound by contractual constraints to reverse engineer technology in order to determine how it functions. To the extent that they decipher trade secrets, they undermine the inventor's advantage. By disclosing the information, they destroy the trade secret. The reverse engineering limitation on trade secret protection thus exposes the trade secret owner to free riding by others. Nonetheless, most commentators believe that it strikes a salutary balance between protection on the one hand and competition and the dissemination of knowledge on the other (Landes and Posner, 2003; Samuelson and Scotchmer, 2002). The trade secret owner can "purchase" greater protection against this risk by investing in higher levels of security (e.g., more effective encryption for software-encoded technology). The inventor can also pursue patent protection, which proscribes reverse engineering, although only for the limited duration of the patent, and mandates disclosure of the invention to the public. By declining to pur-

sue patent protection (or failing to satisfy the requirements thereof), however, inventors should not be able to secure potentially perpetual rights in technologies merely by encrypting them or otherwise obscuring how they function. To do so would undermine the larger balance of the federal intellectual property system.

Compulsory licenses. Copyright law uses a statutory compulsory license mechanism for cover versions of musical compositions, which spurred cumulative innovation in sound recordings. Once a musical composition has been released (with consent of the copyright owner) as a sound recording, any sound recording artist may record and distribute copies of that composition without consent of any copyright owner. This privilege is made possible by limiting the scope of the sound recording copyright to exact duplication[5] and establishing a compulsory license rate for copies made of the underlying musical composition (currently 8.5 cents per copy) (17 U.S.C. §115). The creative freedom associated with this privilege, however, is constrained by the statute—the follow-on recording artist may make "a musical arrangement of the work to the extent necessary to conform it to the style or manner of interpretation of the performance involved, but the arrangement shall not change the basic melody or fundamental character of the work" [17 U.S.C. §115(a)(2)]. This privilege also does not extend to the use of prior sound recordings—as, for example, in digital sampling—without the consent of both the musical composition copyright owner and the sound recording copyright owner. Nonetheless, this compulsory license has likely fostered a wider body of interpretations of musical compositions than would have occurred if musical composition copyright owners held exclusive (blocking) rights. Due in part to this privilege, more than 100 cover versions of the popular song "Mack the Knife" are available, with performances ranging from Louis Armstrong to Bobby Darin and instrumental artists.

1.3.2.5. Remedies As noted above, the main remedies for patent and copyright infringement are injunctions and compensation for past injury, possibly compounded to treble damages in case of willfulness. As these laws are interpreted, courts operate from a baseline of prospective injunctive relief and compensatory damages for past injury. Hence, they do not generally adjust the level of damages as a policy lever, except in the context of enhanced damages, which we discuss below.

With regard to awarding damages for past infringement, the court will often be in the position of having to decide whether, absent the infringement, the rightholder would have licensed. If not, the lost profit may be the value lost to the patentholder because the follow-on product was preempted by the infringer. If licensing would (should) have occurred, the lost profit is lost royalty. These are two rather different inquiries.

[5] "The exclusive right of the owner of copyright in a sound recording . . . is limited to the right to prepare a derivative work in which the actual sounds fixed in the sound recording are rearranged, remixed, or otherwise altered in sequence or quality. The exclusive rights of the owner of copyright in a sound recording . . . do not extend to the making or duplication of another sound recording that consists entirely of an independent fixation of other sounds, even though such sounds imitate or simulate those in the copyrighted sound recording . . ." 17 U.S.C. §114(b).

Lost royalty is even more speculative in the cumulative context than in the stand-alone context, for a sound theoretical reason (Schankerman and Scotchmer, 2001). On one hand, the potential damage award establishes the maximum license fee. On the other hand, the equilibrium license fee establishes the damage award. Hence there is an inherent circularity that leads to multiple equilibria. Because of multiple equilibria, the profitability of the patent is unknowable in advance to a researcher investing in it. This problem is especially acute for research tools, and will be less acute for inventions where infringement leads to competition and dissipates total profit. Because of the dissipation, infringement is its own punishment, and infringement is more easily deterred.

The awarding of enhanced damages in patent law (up to treble) and statutory damages in copyright law can be a policy lever, although it is restricted to penalizing willful infringement. It is not generally seen as a way of addressing the cumulative innovation problem. The patent law standard for awarding enhanced damages produces a deleterious effect upon cumulative innovation. To reduce exposure for treble damages (which is based upon a finding of willfulness), patent attorneys routinely advise their clients (including research engineers and scientists) to avoid reading patent prior art, in effect negating a valuable aspect of the disclosure function of the patent system (Lemley and Tangri, 2003). This is a form of overdeterrence of socially beneficial behavior—learning from prior discoveries. The inability to consult patent prior art undoubtedly results in duplication of research and may lead a researcher to overlook valuable potential solutions to scientific and technical problems. In order to alleviate this undesirable effect, the Federal Trade Commission (2003, pp. 28–31) recommends that the legal standard be tightened to require either actual, written notice of infringement from the patentee or deliberate copying of the patentee's invention, knowing it to be patented, as a predicate for willful infringement. Similarly, the National Research Council (2004, pp. 118–120) recommends elimination of the use of a subjective test for determining willfulness.

1.3.2.6. Channeling doctrines In the context of cumulative innovation, patent law serves a particularly important antidote to the non-disclosure aspect of trade secret law. By penalizing inventors that rely upon trade secret law—by subjecting them to the risk that a later inventor will obtain a patent on what they hold as a secret and thereby be able to block further use of the invention by the trade secret owner—patent law promotes disclosure.

1.4. Administration

Apart from the substantive rules, the administration of intellectual property law significantly affects the efficacy of the overall system. Administration plays the most significant role in patent law, which affords protection only for those inventions judged by a patent examiner to clear the validity hurdles. Therefore, we will focus our attention there. Two sets of issues have attracted substantial economic inquiry—the quality of patent analysis conducted by the Patent Office and administration of judicial deci-

sionmaking. Although optional, registration of copyrights plays some role in licensing. There are no formal requirements for trade secrets.

1.4.1. Registration/examination

Registration and examination processes serve several potential functions in systems aimed at promoting innovation. They can screen out undeserving or defective applications, disclose knowledge to the public at large (published patents, deposit of copyrighted works with the Library of Congress), create a public record of intellectual property titles that can be valuable for licensing and using intellectual property as collateral for debt, and, through the use of application fees, impose some of the system-wide costs of administration on those the most significant beneficiaries.

Due to the relatively high standards for patent protection, ex ante screening of patents through examination by experts trained in technology fields plays a vital role in minimizing disputes as to the validity of patents. An elaborate system of reexamination, reissuance, and interferences (for determining priority among inventions making the same or similar claims) uses skilled and specialized examiners and administrative law judges to resolve disputes.

In view of the low threshold for obtaining copyright protection, examination serves a rather modest role in the overall system. The copyright registration, which is optional, serves primarily as a title registry. Copyright registration also augments the public availability of knowledge through the deposit function. With the move away from formal requirements for obtaining copyright protection (including copyright registration and renewal) to an unconditional system, it is more difficult to identify copyright owners, thereby increasing the transaction costs of licensing (Sprigman, 2004; Landes and Posner, 2003, Ch. 8) (suggesting a system of periodically renewable copyrights to reduce ownership tracing costs).

Renewal requirements and maintenance fees also function as policy levers. Such fees have traditionally been relatively low, being designed to cover the costs of administration, but nevertheless can cause rightholders to let their rights lapse (Lemley, 1999). A large literature, based partly on European data (where such fees have been in effect longer) indicates that about half of patents lapse by the tenth year. [See, in particular, Schankerman, 1998; see also Lanjouw, 1998; Pakes and Schankerman, 1984; Schankerman and Pakes, 1986; Pakes, 1985, 1986]. Cornelli and Schankerman (1999) show how renewal fees can be used to give higher incentive for investment to higher-ability inventors, and Scotchmer (1999) shows how a renewal system can be used as a screening device to give more rewards to inventors who, although they may have higher costs, also have much higher value innovations.

1.4.2. Quality

Most commentators take as a measure of quality the probability that the patent would survive a legal challenge (National Research Council, 2004; Shapiro, 2004), although in

a legal environment with unresolved patentability issues, issues of quality shade into issues of design (Scotchmer, 2004c). Poor quality patents may result from inadequate review of prior art during examination, poorly drafted claims, or lax standards (the height of the non-obviousness threshold). They may undermine economic efficiency by restraining competition, raising transaction costs, and increasing litigation without promoting innovation. A proliferation of poor quality patents can choke entry and cumulative innovation.

Ensuring quality patents, however, comes at a cost. Emphasizing the large number of patent applications (up to 300,000 per year), the administrative and human resource costs of comprehensive patent examinations, and the relatively small number of patents having significant economic value (under 2,000 patent suits filed per year), Lemley (2001) asserts that the costs of improving patent quality ex ante (through more careful examination) exceed the benefits—the Patent Office is, in his view, rationally ignorant. Kieff (2001b) takes Lemley's insight quite a bit further, arguing for abandonment of patent examination in favor of a pure registration. The United States experimented with such a system in its early history (1793–1836), producing chaotic results. The Senate report accompanying the 1836 legislation reinstating formal examination commented that the registration system left the nation "flooded with patent monopolies, embarrassing to bona fide patentees, whose rights are thus invaded on all sides ... Out of this interference and collision of patents and privileges, a great number of lawsuits arise ... It is not uncommon for persons to copy patented machines in the model-room; and, having made some slight immaterial alterations, then apply in the next room for patents." S. Rep. No. 338, 24th Cong.., 1st Sess. 2 (1836). See Landes and Posner (2003, p. 309); Merges (1999a).

But what about Lemley's more modest suggestion of shifting more resources toward screening at the litigation stage rather than the examination stage? As others emphasize, low patent standards raise litigation and licensing costs, impede cumulative innovation where dubious patents stand in the way of improvements, deter entry into promising markets, especially by small entrants that cannot easily withstand costly patent litigation, and encourage more filings, which may hamper the operation of the PTO [Ghosh and Kesan, 2004; Rai, 2003 (benefits of certainty in patent examination); Thomas, 2002]. The relative merits of screening at the examination versus the enforcement stage likely turns on the field of innovation. For example, given the large expenses needed to bring a drug from the testing laboratory to the market, the pharmaceutical industry cares deeply about knowing that patents are screened thoroughly prior to such investments. Uncertainty about patent validity could undermine investment in that sector. By contrast, software and business method patents likely do not require large up-front investments. The number of such patents is quite large and growing. Therefore, post-issuance screening—where selection is based upon which patents are litigated—may make more sense (so long as the chilling effects on new ventures are too great).

Ayres and Klemperer (1999) offer a less conventional argument for preserving uncertainty in the patent system. Based upon the fact that marginal price increases near the monopoly price only benefit the monopolist a small amount while producing dispropor-

tionately large deadweight losses to consumers, they show that even a small amount of uncertainty regarding the enforceability of a patent can alleviate monopoly pricing ex post without materially reducing incentives to innovate. They suggest that such uncertainty can be introduced into the system by having more lax patent examination (as well as other policy instruments, such as the standard for granting preliminary injunctions and application of the doctrine of equivalents). Their model, however, works only where there is uncertainty as to valid patents. Uncertainty as to invalid patents increases deadweight loss (Rai, 2003). More generally, introducing uncertainty by loosening patent quality at the examination stage would impose significant additional costs upon third parties for prior art searches, opinion letters, and transaction costs (Merges, 1999a).

Several considerations suggest that patent quality has declined during the past several decades. The expansion of patentable subject matter to include software and business method patents has been cited a particular concern. Since much of the technological knowledge in these fields subsists in the form of trade secret, products, and services, as opposed to more readily searchable sources (such as patents and published scientific journals), it is far more difficult to search these fields than the traditional patent fields (Galler, 1990; Samuelson, 1990; Merges, 1999a). Furthermore, as patenting of software and business methods took off, the Patent Office lacked examiners adequately trained in these fields. A second reason why patent quality may have declined is as a result of relaxation of substantive validity requirements. Since the consolidation of federal patent appeals in one court in 1982, observers of the patent system have discerned a clear lessening of the non-obviousness requirement (Lunney, 2000–2001, 2004; Barton, 2000; Desmond, 1993). The incentive structure and mission statement of the Patent Office may also have contributed to a lessening of standards. In general, compensation levels at the PTO may not be high enough to retain experienced examiners. More specifically, patent examiners are undertrained, overworked, and subject to distorted incentives that bias the system toward allowances. The bonus system—by which examiners get additional compensation for "dispositions"—favors allowances over rejections. Since rejected applications are easily revived through the continuation process, an examiner can more quickly and confidently secure a disposition through a grant than a rejection [Merges, 1999a; Rai, 2003 (noting further that detailed explanations need to be provided for rejections but not allowances); cf. Lemley and Moore, 2004 (advocating substantial limitations upon continuation practice); Quillen and Webster, 2001–2002]. By contrast, there are no systematic compensation penalties for errors. On a macro level, the Patent Office is supported by fees paid by applicants. The Patent Office has over the past decade shifted its mission toward "customer service"—with the "customer" being the patent applicant, not the public (Corcoran, 1999).

Several sources of evidence support the concern over patent quality. Lunney (2000–2001) documents the easing of the non-obviousness requirement by the Federal Circuit, finding that non-obviousness is far less likely to be cited as the basis for invalidating a patent. Several frequently cited patents—such as those issued for one-click ordering, a system for restroom queuing on airplanes, and crustless peanut butter and jelly sandwiches—reinforce the perception that inventions not need be particularly "inven-

tive" to be patentable. U.S. allowance rates are believed to be in the 70 to 80 percent range, substantially higher than allowance rates at the Japanese and European patent offices (National Research Council, 2004, pp. 43–45; Clarke, 2003; Quillen and Webster, 2001–2002). These discrepancies, however, may be explained by the higher costs of obtaining protection in the United States. U.S. applicants may engage in more prescreening of applications.

Empirical studies of patent quality, however, have not discerned a significant decline in patent quality. According to U.S. PTO quality assurance audits, error rates have fluctuated between 3.6 and 7 percent since 1980, trending upward through the 1990s, but declining since that time (National Research Council, 2004, p. 40). This data, however, is quite limited and has been questioned by external inspectors (National Research Council, 2004, p. 40). Cockburn and Henderson (2003) find that resources devoted to patent examination (as reflected in examiner hours and actions) have kept pace with the increased number of applications during the 1985–1997 time frame. But see National Research Council (2004, pp. 41–43) (reporting a 20 percent drop in the number of examiners per 1,000 applications from 1999–2002). Using a data set of 182 patents litigated at the Federal Circuit between 1997 and 2000, Cockburn and Henderson (2003) do not find evidence that examiner experience or workload predict invalidity decisions. Several commentators have suggested that patent quality can be proxied by the number of prior art citations disclosed in the application: fewer citations indicate that the applicant has not provided the examiner with the range of existing knowledge against which to evaluate non-obviousness. Focusing on the subject matter area that has received the most heavy criticism on quality (as well as other) grounds—business methods—Allison and Tiller (2003a, 2003b) found that business method patents examined through the end of 1999 cited more patent and nonpatent prior art references than a large, contemporaneous, random sample of all patents. Focusing on the success ratio of patent applications as a measure of patent quality, Lesser and Lybbert (2004) find that patent standards have been relatively stable over the 1965–1997 time horizon.

Even if patent quality has not declined, there could be substantial benefits from reforming the patent system to improve patent quality. Proposals have focused on different stages of the patent review pipeline: (1) initial patent examination; (2) incentives for disclosure of relevant information (including oppositions); (3) substantive standards; and (4) litigation-related reforms.

Patent examination. Applying insights from the field of personnel economics to the design of the Patent Office, Merges (1999a) advocates reforming of Office procedures and practices to emphasize examiner training, employee, and more systematic quality review of applications. He also recommends changes in the bonus system to balance rewarding application processing efficiency with quality review. Responding to public criticism, the Patent Office adopted a second stage of review for business methods.

Eliciting information. Several scholars have focused on the problem of eliciting good prior art information early during initial examination and possible later review of patent validity. Although good prior art research data bases are available in many fields, patent

applicants and their competitors often have the best access to the most current and relevant prior art. The duty of candor imposed upon applicants encourages disclosure of such knowledge, but it may not be adequate. Various proposals aim to elicit more and better prior art disclosure.

Under current practice, the duty of candor requires only that the applicant disclose all material prior art of which they are aware to the Patent Office. They need not explain how such art relates to the patent. As a means of encouraging more expansive applicant disclosure, Kesan (2002) advocates limiting the presumption of validity in subsequent litigation over the patent to art that the patentee disclosed and explained during prosecution.

Another approach to the information problem focuses on third party disclosure—mechanisms to enable competitors to participate in the examination or post-issuance review process. As currently constituted, the U.S. patent examination is largely ex parte—i.e., involving only the applicant and the Patent Office. Often, competitors of the applicant are in a good position to identify weaknesses in an application. For this reason, many observers recommend expanded use of inter partes (opposition) proceedings in order to bring forward stronger resistance to weak patents, as occurs in the European Patent Office (National Research Council, 2004; Federal Trade Commission, 2003; Levin and Levin, 2003; Rai, 2003; Kesan, 2002; Merges, 1999a; Soobert, 1998). Publication of applications can also generate third party input to the examination system (Duffy, 1998). The primary motivation for opposition is to prevent a competitor from gaining an unfair advantage in the marketplace through obtaining a questionable patent. This can prevent or limit commercial advantages in the marketplace as well as avoid subsequent litigation costs.

The efficacy of an opposition may well be hindered by a collective action problem. Even though competitors of a patent applicant collectively may gain more from defeating the patent than the costs of pursuing an opposition, each individual competitor may not have sufficient incentive to initiate the process (Thomas, 2001). An alternative approach to eliciting pertinent prior art provides a financial reward to those who come forward with information to invalidate unwarranted patents [Thomas, 2001 (examination stage bounty); Kesan, 2002 (one-way fee shifting rule); Miller, 2004 (litigation stage bounty)]. The threat of such a bounty being exacted from the patentee creates a further deterrent effect on filing questionable patents.

Substantive standards. A relatively direct means of increasing patent quality is to raise patentability standards. Some have proposed narrowing the scope of patentable subject matter—for example, by requiring that an invention have "industrial applicability," which would exclude some business methods claims (Thomas, 1999); or excluding business methods or DNA sequences as a class. In 2001, the PTO issued written description and utility guidelines intended to signal that it would be applying uniform and stringent guidelines for DNA-related patents (United States Patent and Trademark Office, 2001). Both the FTC and NAS reform studies recommend reinvigorating the non-obviousness bar for all inventions (Federal Trade Commission, 2003; National Research Council, 2004).

Litigation-relation reforms. As an alternative to increasing the costs of patent examination, Lemley (2001) advocates that patent litigation rules be altered to allow better ex post quality review of those relatively few economically significant patents that become the subject of litigation. In particular, he recommends that the presumption of validity be removed. Rai (2000, 2003) comes at the problem from another direction, recommending greater deference to PTO rejections made on grounds of obviousness.

1.4.3. Judicial administration

The patent system relies heavily upon the judiciary to interpret the Patent Act and adjudicate infringement cases. Although Congress has written several quite specific exceptions into the statute over the past two decades, the courts have played a more central and active role in determining validity requirements and infringement standards. The expansion of the patent domain into the software, business methods, and biotechnology has been driven almost entirely by judicial interpretation as opposed to new legislation. The non-obviousness standard, written description requirement, utility threshold, aspects of novelty (e.g., inherency doctrine), claim construction, and infringement analysis continue to be calibrated by the judiciary.

The design of judicial institutions governing patents can have significant effects on the operation of the patent system. It is useful to distinguish between the trial and appellate levels. In some nations, patent trials are heard by specialized and technically trained tribunals. By contrast, patent cases are heard in the first instance in the United States by courts of general jurisdiction, i.e., non-specialist courts. The location of technology industries (e.g., Silicon Valley (Northern District of California)), incorporation patterns (favoring Delaware), and attorney preferences for plaintiff-friendly juries (Eastern District of Texas) among other factors affect the geographical incidence of patent cases. Thus, several district courts have become relatively more specialized in the handling of patent cases. No systematic comparative international study has yet been made analyzing the advantages of specialization and technically trained jurists on patent adjudication. Several empirical studies highlight the high reversal rate for claim construction decision by district courts (Chu, 2001 (finding that the Federal Circuit modified claim construction decisions by district courts in 44% of cases); see also Moore, 2001 (finding 33% modification rate for a somewhat different sample period). Based upon institutional analysis, Rai (2003) advocates the use of specialized trial courts for patent cases in order improve the quality of technologically oriented fact-finding within the judicial system. See also Rai (2002); Wiley (2003) (advocating greater use by courts of technology experts as special masters)).

Specialization at the appellate level has garnered substantial scholarly attention. Prior to 1982, appeals of patent infringement cases in the United States were heard in the regional circuit courts in which the district courts were located. This system produced inconsistency in patent law as well as high rates of invalidation in some courts of appeals. Not surprisingly, it also produced a good deal of forum shopping (Dreyfuss, 1989; Commission on Revision of the Federal Court Appellate System, 1975, pp. 217–

221). In an effort to improve administrative efficiency, Congress centralized appeals of all patent cases (from both the Patent Office and district courts) at the Court of Appeals for the Federal Circuit in 1982. As some observers at the time surmised, such a move would likely go beyond harmonizing the law. Institutional considerations—such as tunnel vision, political influence in selection of jurists for the Federal Circuit, and socialization effects among the members of the court—would likely produce a pro-patent bias (Posner, 1985).

Several studies have borne out this prediction. Since 1982, patent law has become both more unified and more favorable to patentees. Cf. Wagner and Petherbridge (2004) (reporting mixed results on unity). Federal Circuit decisionmaking has generally resulted in an expansive interpretation of the subject matter of the Patent Act, narrow interpretation of limitations (e.g., experimental use), lower thresholds for protection, higher infringement damage awards, and greater average patent scope [Lunney, 2004 (finding that prior to 1982, courts were more likely to reject claims on invalidity grounds than non-infringement by a ratio of nearly three to one; since 1982, the ratio nearly reversed, with non-infringement becoming the dominant (68.1%) explanation for cases being dismissed and invalidity becoming comparatively rare); Landes and Posner, 2003, 2004 (using regression analysis, finding that Federal Circuit has had a positive and significant impact on the number of patent applications, the number of patents issued, the success rate of patent applications, the amount of patent litigation, and possibly the level of research and development); Lanjouw and Lerner, 1997; Kortum and Lerner, 1998 (finding that, from 1982 to 1990, the Federal Circuit affirmed 90% of district court decisions holding patents to be valid and infringed, and reversed 28% of judgments of invalidity and non-infringement); see also Allison and Lemley, 2000 (finding that appellate judges appointed to Federal Circuit since 1982 have been significantly more likely to uphold patent validity)].

The normative implications of these effects are complex. The harmonization of patent law has reduced uncertainty about the law, discouraged forum shopping, and possibly promoted research and development spending in some sectors (Landes and Posner, 2003, pp. 345–357). The shift away of validity-based policy levers, however, has made the patent system less sensitive to the diversity across the range of technological fields. Lunney (2004) concludes that the Federal Circuit has shifted the patent system toward a more uniform, one size-fits-all regime in which validity has become more routine and scope more narrow. In effect, the court has dampened several critical validity policy levers, limiting the versatility of the patent system to promote the diverse range of new technologies. Several scholars advocate a shift in the Federal Circuit's role, viewing it as the best situated institutions for producing a patent system that responds to the heterogeneity of inventive activity across the growing range of technological fields (Burk and Lemley, 2002, 2003; Rai, 2003; but cf. Wagner, 2003).

1.5. Enforcement

In our discussion of policy levers, we implicitly assumed that the rightholder has little difficulty identifying, pursuing, and excluding unauthorized users. The design conclusions of the literature depend on that assumption. However, enforcement of intellectual property laws in the real world is far more complex than this stylized caricature. The profitability of rights can be changed by uncompensated infringement or by the terms of license on which rightholders are induced to license in the absence of strong rights.

We have already discussed the main remedies to infringement, damages and injunctions, and whether they are likely to deter infringement. We now augment that discussion by saying what is known about the costliness and effectiveness of enforcing intellectual property rights, drawing attention to some additional legal mechanisms.

Evidence on patent litigation. Comprehensive evidence on patent litigation can be found in Lanjouw and Schankerman (2001, 2004), based on litigation data assembled by the PTO (to whom patent litigation is supposed to be reported), patents themselves (which contain information on the technology and characteristics of applicants), the Federal Judicial Center (which assembles information on the disposition of cases, such as whether they settle and when), and Standard and Poor's database on companies that are publicly traded (which contains information on characteristics of the company such as size). See also Allison et al. (2004). Based on this data, the overall litigation rate is about 2 cases per 100 patents, concentrated on high-value patents. An earlier study by Lerner (1994), restricted to biotechnology patents, estimated that 6 in 100 patents were litigated. Litigation increased substantially over the 1978–1999 period, but the increase is attributable to the changing composition of patents, and to the overall increase in patenting. There was a 71 percent increase in patent grants from 1978 to 1995. Most of the increase in patent suits has been in drugs, biotechnology, and computers and other electronics, which have always been highly litigated and have been increasing as a percentage of total patent grants. Thus, litigation has grown faster than patent grants.

The role of small entities (including independent inventors and firms that acquire patent portfolios for purposes of licensing), and particularly firms that do not themselves practice their inventions, in patent litigation has been the subject of growing interest. At least traditionally, small firms were at a disadvantage due to the magnitude of litigation and enforcement costs. Lanjouw and Schankerman showed that patents held by small firms are more likely to be litigated. Lerner (1995) concluded that small firms avoid technology areas where litigation is prevalent, and Lanjouw and Lerner (2001) showed that, in the litigation process itself, preliminary injunctions are used strategically by large firms against small firms. This pattern, however, appears to be shifting. In the dot com age, a proliferation of software and business method patenting has spawned a plaintiff's patent bar that aggressively enforces patents [Federal Trade Commission, 2003 (referring pejoratively to a new class of "patent trolls"); Sandburg, 2001; Meurer, 2003; Allison et al., 2004 (noting the high percentage of patent purchasers (as opposed to inventors or original assignees) instituting patent litigation)]. The asymmetric stakes

of such litigation may in fact favor small enterprises, which have little to lose and much to gain by asserting patents against large enterprises.

Large entities with sizable patent portfolios often prefer to resolve their differences with cross-licenses (often royalty-free) rather than risk mutually assured destruction that can result from high stakes patent battles (Parchomovsky and Wagner, 2005). Hall and Ziedonis (2001) found that between 1979 and 1995, semiconductor firms amassed large patent portfolios in order to deter litigation and to negotiate more favorable access to technology owned by competitors. A follow-up study indicates a spike in semiconductor patent litigation relative to R&D activity (Ziedonis, 2003). More specialized semiconductor design firms—lacking complementary manufacturing assets—have a higher propensity to litigate.

Indirect liability and least-cost enforcement. Courts have long recognized liability for acts that contribute to infringement by others. Congress codified liability for contributory infringement, with limitations, in the 1952 Patent Act (35 U.S.C. §271). Similarly, copyright law extends liability to those who contribute to and vicariously benefit from copyright infringement.

Indirect liability can reduce enforcement costs by allowing an intellectual property holder to cut off infringement at a higher level in the chain of potentially responsible actors—such as suppliers of the means for infringement. It can also provide a more effective sanction when direct infringers are difficult to identify. Of course, the act which contributes to enforcement may also have a lawful purpose, e.g., the sale of a component part used in practicing a patented invention. For this reason, the law does not recognize contributory infringement if the acts or product sales have "substantial non-infringing uses."

Copyright enforcement in the digital age. Above we cited arguments along the lines that limited sharing does not impinge on rightholders provided that they anticipate the sharing in their pricing behavior. These arguments were more suited to the analog age (e.g., photocopying) where a form of "natural" encryption—the lack of availability of reproduction technologies, the degradation of quality of second generation copies, and the relatively high cost of making copies—limited unauthorized reproduction. Furthermore, anyone seeking to mass produce and distribute copies could be easily detected. Although copyright enforcement has long been a problem in some foreign markets (Ryan, 1998), copyright enforcement was not a major worry in the United States during the analog age.

Modern digital technology has brought enforcement to the forefront of copyright policy throughout the world (Menell, 2003). Such technology allows rich media content to be flawlessly copied and redistributed through largely anonymous peer-to-peer digital networks. In this environment, it is less likely that degradation or the cost of copying will protect proprietors, or that sharing groups will be limited in size.

Where infringement is particularly difficult to detect, preventive measures may be a second best means of preventing unauthorized use of intellectual property. In response, film, music, computer software, and computer games proprietors are turning to tech-

nical protections, such as encryption and copy controls. The economic consequences depend on how effective the technical protections are, a matter which is still evolving. As a means of enhancing the effectiveness of such technologies, Congress enacted a set of anti-circumvention provisions as part of the Digital Millennium Copyright Act (DMCA) which largely prohibit decryption of digital locks placed on content. Such preemptive protections, however, have the undesired consequence of preventing some otherwise lawful uses—e.g., fair use of an encrypted work. To reduce such effects and balance both under and over-enforcement, the DMCA contains numerous exceptions, such as for reverse engineering of software products for purposes of creating interoperable programs, security testing, encryption research, etc. The Act also empowers the Librarian of Congress to grant categorical exceptions.

If users will circumvent the protection system when the cost of circumvention is lower than the price, the threat of circumvention will have a moderating effect on the pricing strategy of vendors, which reduces per-period deadweight loss (Conner and Rumelt, 1991). Park and Scotchmer (2004) point out that if the price reductions are achieved through technical protections that can be circumvented at a cost, and that the technical protections continue forever, just as trade secrets can, the net effect can be beneficial for both content providers and consumers. Because the price will be lower than monopoly price, the profit-to-deadweight-loss ratio will be lower. Consumers may be better off due to the lower price, and proprietors may be better off due to the longer protection. Thus, the transition away from the enforcement of legal protections to technical protection has an ambiguous effect on consumer welfare and on the incentives to create.

Because digital sound recording files are widely available (the compact disc encoding technology introduced in 1981 was not encrypted) and relatively small (in comparison to film files), the sound recording industry has been the first content industry to be affected on a large scale by the capabilities of the emerging digital platform (Menell, 2003). Surveys and various other forms of empirical evidence suggest that teenagers (a prime target audience for new music and film releases) consider peer-to-peer networks to be an attractive source for obtaining content. The overall effects on the content industries are complicated to assess, although the most recent studies seem to suggest that peer-to-peer technology is at least partially responsible for the post-2000 decline in record industry revenues. See Liebowitz (2004) (finding that peer-to-peer file sharing has caused harm); but see Oberholzer and Strumpf (2004) (questioning a link between free downloads and CD sales).

In order to combat unauthorized distribution for the purposes of bolstering traditional retail sales and building support for legitimate online distribution (subscription and download services), the music industry initiated a high profile enforcement campaign against distributors of peer-to-peer software. After an initial victory against a centralized peer-to-peer technology (Napster), the record industry has encountered difficulty shutting down more decentralized networks on legal (newer technologies are outside of the software providers' control and have non-infringing uses) and practical (off-shore providers) grounds. As a result, the record industry has begun pursuing indi-

vidual uploaders directly, although this is a costly process due to the relative anonymity of filesharers.

Economic analysis of copyright enforcement in the digital environment involves several complex considerations. Allowing greater leeway for courts to hold distributors of peer-to-peer software indirectly liable for infringement has the advantage of economizing on enforcement resources, but it produces a chilling effect on legitimate uses of such technology and discourages the development and diffusion on new digital technologies that might have substantial societal benefits. Some loosening of the "substantial non-infringing use" defense may be called for to balance the competing effects on aesthetic creativity on the one hand and technological innovation on the other (Menell, 2005; Lichtman and Landes, 2003). Several scholars advocate abandoning direct enforcement in favor of a levy system (fees on technology and Internet services that operate as a compulsory license) as a means of supporting creative enterprise, although such approaches cannot price usage efficiently and introduce administrative costs and rent-seeking behavior (Netanel, 2003; Fisher, 2004; Gervais, 2005; but see Merges, 2004a). Lemley and Reese (2004) suggest that limiting enforcement to actions against direct infringers through a streamlined and lower cost administrative enforcement process would provide the best compromise between deterrence and compensation on the one hand and freedom to innovate on the other. Their proposal, however, would entail substantial administrative cost.

Another dimension of enforcement policy relates to the choice between public and private enforcers and the penalty structure. Public enforcement can offer advantages where the government has easier access to information about infringing behavior, can realize economies of scale not achievable by private enforcers, or can impose sanctions (e.g., imprisonment) that are more effective than civil penalties. Exclusive government enforcement may be appropriate where there is some benefit to be gained from prosecutorial discretion. The federal government has expanded criminal penalties for unauthorized online distribution of copyrighted works.

1.6. Interaction with competition policy

Intellectual property protection can conflict with competition policy. We discuss the principal economic theories bearing on this tension here (for comprehensive analysis of the intellectual property/antitrust interface, see generally Hovenkamp, Janis, and Lemley, 2004). The chapter on antitrust also examines this issue.

There are two stages at which antitrust concerns can be raised in the intellectual property context: in the rivalry to achieve inventions in the first place, and in the licensing that takes place ex post. Licensing is generally thought pro-competitive, since it increases the use of intellectual property. Further, licensing is common. Among members of the Intellectual Property Owners Association, 17.6 per cent of patents are licensed out, and many innovators invest with the sole objective of licensing rather than practicing or manufacturing their innovations (Cockburn and Henderson, 2003). In the content industries, many works are independently produced and distributed by larger companies

that finance and/or license the copyrighted products. Because licensing creates alliances that affect production, distribution, and pricing, such transactions inevitably raise competition issues.

Some of the pro-competitive uses of licensing were discussed in the context of cumulative innovation. These include licensing to resolve blocking patents, and to ensure the widespread use of complementary pieces of technology such as research tools. Here we discuss the more traditional context of horizontal substitutes and the special circumstances of licensing complementary intellectual property.

Licensing is pro-competitive whenever the efficient use of the intellectual property is to share it. However, a problem arises when the licensor and licensee are rivals in the market. The terms of license must facilitate the sharing of technology without at the same time facilitating collusion. This is a fine line to walk, and since the firms will not be inclined to walk it, the proper boundary must be established as a matter of law.

For example, suppose the technology reduces the marginal cost of producing a product. A license from the patent holder to a rival creates a social benefit by cutting the rival's costs. But if the royalties are higher than the cost-saving, the license can result in a market price that is even higher than would prevail in the absence of licensing. In that case, consumers do not benefit from the cost reduction. Should this be allowed?

As in other antitrust areas, the governing principle in U.S. law has increasingly become rule of reason; see Gilbert and Shapiro (1997) for a discussion of the per se rules (the "nine no-nos") that have come and gone in U.S. law and policy. We will not give a comprehensive overview of specific licensing rules that have been in and out of favor, but will instead articulate some of the economic principles that have been suggested as a basis for adjudicating licensing practices.

In general, rule of reason is a test that weighs harms to competition against gains in efficiency. But this is not a very practical test in the intellectual property context, since efficiency can be either *ex ante* or *ex post*. Even if a licensing practice seems collusive ex post, the parties may argue that the prospect of using it is what gave them incentive to invest in the first place. It is hard to see what kind of evidence would either contradict or buttress such a claim, especially in a research environment where, ex ante, success was not assured. In that case, a firm will only invest if it earns substantial profit (higher than cost) in case of success. Even more importantly, it is not clear how such an inquiry respects the presumed right of Congress to set the incentives for research.

A slightly more practical test, which is at least founded on a sensible and clearly articulated principle, is that of Kaplow (1984), who recommends that a licensing practice be approved if it allows the rightholder to earn profit in a way that increases the profit-to-deadweight-loss ratio. The conceit is that Congress anticipates this efficiency principle in setting the other policy levers, such as length, so that the courts are implementing Congress's will. A problem with the principle, however, is that it has no natural boundary. Into which markets is the rightholder allowed to leverage? For example, if it is efficient to raise money by taxing real estate, then shouldn't the intellectual property be licensed collusively to real estate owners? It is not obvious how the principle mediates between the incentive purpose of the patent grant (the patent should not

be lucrative unless it creates value to users) and the problem of raising money through efficient taxation, in whatever market that can be done most efficiently.

Maurer and Scotchmer (2004a) claim that courts have implicitly addressed this problem by applying a principle they call "derived reward," which means that the profit can only be earned by collecting some of the social value created by the invention. In fact they argue that courts have employed (and previous commentators have implicitly endorsed) three principles that jointly constitute a sensible guide for adjudicating license disputes: profit neutrality, derived reward and minimalism. Profit neutrality means that the rightholder should not be penalized for his inability to work the patent efficiently himself. (This principle may, for example, justify price-fixing in the patent context.) Minimalism means that courts should not allow terms that are unnecessary in achieving the first two principles. Unnecessary terms only give opportunity for sham licenses.

The problems are compounded when a user needs many complementary licenses. Both cross-licensing and patent pools can compound the concerns about competition. For a discussion of cross licensing, see Barton (2002); Denicolò (2002b); Merges (1996, 1999b); Gilbert (2002); Lerner and Tirole (2004). Patent pools are generally suspect when they contain substitute technologies, but not when they contain complements. Price-fixing by a pool that contains substitute patents will generally raise the joint price relative to individual licensing, whereas price-fixing by a pool that contains complements will generally lower the joint price relative to individual licensing.

Aside from their effect on prices, cross licenses and patent pools affect the incentives to create and improve technologies in other ways. The division of profit among rightholders in the pool determines their rewards. The division of profit has not been the focus of the literature, but there is no reason to think that profit will be divided as necessary to cover the respective innovators' costs. Looking forward instead of backward, if all members of a pool will share equally in the benefits of new knowledge, then any member's incentive to invest in new knowledge is attenuated. Pooling may reduce the incentives to innovate.

The second concern of antitrust policy is how ex ante alliances (rather than ex post alliances) affect incentives to innovate. The official policy of the antitrust authorities in this regard is articulated in the 1995 *Antitrust Guidelines for Licensing Intellectual Property* (U.S. Department of Justice, 1995). The *Guidelines* distinguish between "technology markets" in which firms license intellectual property that already exists, and "innovation markets" in which firms compete to develop new technologies. The policy with respect to innovation markets addresses two fears: that alliances may retard progress by reducing competition to innovate and reduce the number of substitute innovations, and undermine competition ex post in a product market.

The *Guidelines* assume that rivalry in innovation generally improves welfare—more rivalry will lead to greater investment, which will in turn produce more rapid innovation. Cf. Loury (1979), Lee and Wilde (1980), Reinganum (1981, 1982, 1989), Merges and Nelson (1990). Rivalry might, however, result in duplication of costs without yielding more innovation, dissipating the value of innovation (Barzel, 1968; Kitch, 1977; Grady and Alexander, 1992; Gilbert and Sunshine, 1995a). For this reason, the prospect

theory of patent policy favors non-rivalrous exploitation of innovation opportunities, whereby an initial prospector obtains "breathing room" to develop the claim without fear that rivals will preempt or steal the claim and the inventor will be able to coordinate the development process (Kitch, 1977). The opportunity to license the technology enables the inventor to contract with entities that may be better able to develop the claim. The prospect theory thus turns importantly upon a smoothly functioning technology licensing market and the capacity, foresight, and rationality of prospectors to coordinate the development and diffusion of the technology.

In theory, therefore, the effects of rivalry on economic welfare are ambiguous. Whether competition promotes innovation better than coordination depends, among other things, on the nature of the innovative process and the innovative environment. Lurking behind the disagreement is Schumpeter's classic 1942 book, arguing that market concentration encourages innovation. The *Guidelines* largely reflect the opposite view, that concentration inhibits innovation. There is a large, but inconclusive empirical and theoretical literature on this question, originating with Arrow (1962).

The *Guidelines* reflect the policy of the antitrust agencies, which is not necessarily the law as interpreted by the courts, which apply a rule of reason standard. Cost efficiencies that might be considered include delegating effort to the more efficient firms (Gandal and Scotchmer, 1993), sharing technical information that might be hidden if firms compete (see Bhattacharya et al., 1990, 2000; Brocas, 2004), sharing spillovers of the knowledge created (see Katz and Shapiro, 1987; d'Aspremont and Jacquemin, 1988; Kamien, Muller, and Zang, 1992; Suzumura, 1992; Aoki and Tauman, 2001), or avoiding duplicated costs. See the Appendix to Hoerner (1995) for a compendium of early cases in which courts and the agencies have made judgments about the relative merits of various arguments, and also Gilbert and Sunshine (1995a, 1995b).

Turning to the second concern—that alliances might undermine competition ex post in product markets—mergers or other alliances can lessen competition where the combined enterprise develops a single product where the separate entities would have developed competing products. Evaluating the welfare effects of such alliances puts courts in the difficult position of predicting what types of intellectual property the members of a proposed alliance would develop absent the merger. The firms that propose to merge will presumably not announce that they would otherwise develop noninfringing substitute products. Instead they will argue that competition will be wasteful and duplicative, and that only one firm will, in the end, have a viable product. Given this incentive to dissemble, the agencies and the court might rightfully be skeptical. Cf. Shapiro (2003).

Related issues arise in the rules governing standard-setting organizations, patent pools, and cross-licensing agreements. Such agreements can promote consumer welfare by facilitating innovation in network industries and facilitating the development of products incorporating the most advanced technologies or where different entities hold mutually blocking patents (Barton, 1997; Lemley, 2002 (standard-setting organizations); Shapiro, 2001; Lerner and Tirole, 2004 (patent pools); Merges, 1996 (patent pools; copyright collectives); Besen, Kirby, and Salop, 1992). Given the transaction costs of licensing (including the costs and delays in resolving disputes about intellectual

property rights) and the importance of standardization in many markets, such institutions can play a critical role in promoting innovation and commerce. Nonetheless, like any agreement among competitors that can exclude competitors and potential entrants, such licensing must be scrutinized to ensure that the pro-innovative benefits outweigh the anti-competitive costs.

1.7. Organization of industry

The organization of industry can affect the incentive to do R&D, and, in reverse, the task of doing R&D can be a reason that industry wants to reorganize. In this section we discuss (1) the Schumpeter (1942) hypothesis relating to monopoly and innovation, (2) the role of employment relationships, geographic concentration of innovation resources, and contracting patterns in promoting innovation, (3) patent races and alliances such as research joint ventures, (4) systems competition and network industries, and (5) the open source movement.

Much of the research on the role of industry structure on innovation traces back to Schumpeter's (1942) hypothesis, based largely on empirical grounds, that large, monopolistic firms are more innovative than small, competitive firms because of their superior ability to marshal resources for large R&D programs. The hypothesis, if true, has three important implications. First, if large firms have an exaggerated incentive to do R&D, then R&D perpetuates monopolies rather than controlling them. But this is not necessarily bad if more monopoly means more progress. Second, if monopolists have more incentive than rivals to patent close substitutes, as suggested by Gilbert and Newbery (1982), then the analysis of patent breadth summarized in the section on policy levers may be moot. The analysis is based on competition between rival patentholders, which is not relevant if patents on substitutes are likely to be held by a single firm. Third, if size increases the incentive to innovate, then an antitrust analysis based on rule of reason would be less hostile to merger among innovative firms than otherwise.

Subsequent empirical and theoretical work of the Schumpeter hypothesis has proven inconclusive. Survey research by Levin et al. (1987) and Cohen, Nelson, and Walsh (2000) suggests a much more complicated relationship between market structure and innovation than suggested by Schumpeter. On purely theoretical grounds, Arrow (1962) showed that monopoly can reduce the incentive to invent, while at the same time making invention more valuable. Suppose that the innovation in question is a cost-reducing innovation, and suppose that the cost reduction is so large that the innovator will become a monopolist even if the market was previously competitive. Compare the following two situations: Prior to the innovation, the innovator operates in a perfectly competitive market, or, prior to the innovation, he is already a monopolist. Then, contrary to Schumpeter's hypothesis, the incremental profit that the innovator earns by innovating is larger if he begins as a competitor than if he begins as monopolist. This is because, as a monopolist, he would have earned some profit in any case. On the other hand, the gain in consumers' surplus is larger if the innovator starts as a monopolist, since consumers then started with already high prices. Thus, when the innovator begins as a monopolist,

innovation creates less profit and more consumers' surplus than when he begins as a competitor. As a consequence, it may be optimal to offer greater profit incentives, e.g., through patent life, but there is no way to achieve that, since intellectual property rights cannot depend on market structure.

This inquiry relies on the notion that there are commonly known opportunities to produce knowledge (the "production function" model). If ideas are scarce, then patents on substitute technologies are less likely to become concentrated as a consequence of incentives. Since there is only one monopolist and many rivals, a rival is more likely to think of any given competing product than the monopolist.

Employment conditions as well as the geographic concentration of industry can have a strong effect on the pace of innovation. In comparing Northern California's Silicon Valley to Boston's Route 128 corridor, which were comparably positioned to lead the digital technological revolution, Saxenian (1994) found that Silicon Valley's free wheeling culture of encouraging exchange of information and mobility of labor across companies proved more successful than Route 128's more proprietary, staid, and vertically integrated business ethos. California's legal limitations on non-competition agreements as well as its competitive venture financing network fostered sustained rapid technological progress and relatively stable economic growth, defying the predictions of product cycle theory (positing that regions follow a pattern of innovation, growth, maturation and scale production, and ultimate decline as production shifts to other, lower cost regions) and the production-function model of knowledge creation.

Thus, the organization of industry has important impacts on the success of R&D and the discovery of knowledge. In reverse, intellectual property also affects the organization of industry, in sometimes surprising ways. Of these we mention three: the incentive for competitors to collaborate in research, the organization of network industries, and the open source movement.

Much of the earlier economics literature was devoted to studying patent races, e.g., how many firms would enter a race, how intensively they would compete, and at what point there would be a shake-out. Aspects of the intellectual property environment that affect these matters are the private value of the patent right (Loury, 1979; Lee and Wilde, 1980; Dasgupta and Stiglitz, 1980a, 1980b; Tandon, 1983; Reinganum, 1982, 1985, 1989; Wright, 1983; Denicolò, 1996), legal details such as whether interferences are resolved in favor of the first firm to file or first to invent (Scotchmer and Green, 1990), the degree of spillover in knowledge that will occur after the invention (Katz and Shapiro, 1987; d'Aspremont and Jacquemin, 1988; Kamien, Muller, and Zang, 1992; Suzumura, 1992; Aoki and Tauman, 2001), whether the firms can learn from observing each other's investment strategies, either about the other firm's research efficiency (Choi, 1991) or the other firm's private information about the value of the objective (Minehart and Scotchmer, 1999), and whether licensing would occur to prevent it (Gallini, 1984; Gallini and Winter, 1985; Shapiro, 1985; Rockett, 1990).

We mentioned at the outset that one of the defects of intellectual property as an incentive mechanism is that the investments it incites might not be efficient. First, the

private return to entering a patent race is different from the social return. Second, the patent race does not aggregate or use the firms' private information about their relative efficiency or the value of the investment (Gallini and Scotchmer, 2002).

With respect to the first point, part of the entrant's reward is a transfer from the other participants. When the entrant's probability of winning goes up, the other firms' probabilities of winning go down. To the extent that these effects are offsetting, entry creates a benefit for the entrant, but not for society as a whole. Thus there may be too much entry. There will almost certainly be too much entry if the private value of the right is equal to the social value. There may alternatively be too little entry if the private value of the intellectual property right is low relative to its social value, or if the innovation will create unappropriable, but beneficial, spillovers among firms.

The other inefficiencies that arise in patent races are due to the imperfect sharing of information about cost efficiency or the value of the objective, an unwillingness to disclose intermediate steps of progress (Scotchmer and Green, 1990), and an unwillingness to share technical information (Bhattacharya, Glazer, and Sappington, 1992). Many of these inefficiencies can be solved by forming a joint venture to share information of the various types, and thus allocating R&D effort efficiently. This problem has been studied in various contexts using the methods of mechanism design (Bhattacharya, Glazer, and Sappington, 1992; Gandal and Scotchmer, 1993; Brocas, 2004).

Collaborations among innovative firms through merger or a joint venture can have the beneficial effect of avoiding the inefficiencies of a patent race. However they can also be anticompetitive, and are therefore a matter for antitrust scrutiny, regardless of whether the firms have market power in a product market.

Systems competition has come to play a critical role in the digital technology field (Matutes and Regibeau, 1992; Katz and Shapiro, 1986; Farrell and Klemperer, 2004). A "system" has complementary pieces, such as a computer operating system and compatible software. The distinguishing aspect of a "system," as opposed to other complementary products, is that the two pieces of the product must be made compatible by some kind of interface. There are three features of a system that might be protected with intellectual property: the hardware (platform), the interface, and the software (applications). Which, if any, should be protected?

When the interface is itself proprietary, the system is called "closed," and otherwise "open." Whether or not the platform and applications are also protected, open and closed interfaces lead to different market structures. With open interfaces, firms may enter on both sides of the market to create products compatible with the complementary ones. With closed interfaces, the two sides of the market must be supplied by an integrated firm, namely, the firm that controls the interface. This control may be exercised by licensing the right to make compatible applications, perhaps with an exclusive dealing clause.

Especially in the context of network effects, a proprietary interface may become an important determinant of market structure. Network benefits arise when the value of using the system increases with the number of other users. As a consequence of net-

work benefits, the market may "tip" to a single integrated system, such as the Microsoft Windows operating system and applications. The threat of tipping is reinforced if an entrenched platform owner has more incentive to increase the number of applications because he has more customers.

With an open interface, a system is not likely to remain under integrated ownership, due to entry. In contrast, due to the tipping phenomenon, a proprietary interface can create market power and profit far beyond the value of any social value it provides. Indeed, the interface can be entirely arbitrary, and not have any social value at all. Protection of the interface thus serves a very different economic purpose than protection of the intellectual property in operating systems or applications.

It seems natural to protect the innovations on the two sides of the market (platforms and applications), since they represent the costly and creative endeavors for which intellectual property is intended. If both sides are adequately protected, there is no need to protect the interface as well. The resulting open market structure would be similar to any market with complementary goods. However, this outcome may be difficult to achieve. The interface may be protectable under copyright or patent law, although strong economic arguments can be made on the basis of network economics that the thresholds for such protection should be quite high and that rights should not be exclusive. See Menell (1987, 1989, 1994, 1998b, 2003); Cohen and Lemley (2001); O'Rourke (2000). Even if such protection is not available, interfaces may be protected through encryption (and trade secrecy). Depending on the complexity of the encryption, reverse engineering may be an antidote subject to the limitations of anti-circumvention constraints (Samuelson and Scotchmer, 2002). For an analysis based on protection of interfaces, rather than the two sides of the market, see Farrell and Katz (1998).

The open source movement has developed in part as a response to the constraints of closed systems (McGowan, 2001; Benkler, 2002, 2004). The movement developed in the computer software industry around programming efforts to develop the Apache web server and the Unix operating system, under the name Linux (Raymond, 2001; Lerner and Tirole, 2000). It employs intellectual property protection in an unconventional manner: as a means of precluding innovators building upon the open source platform from asserting intellectual property rights to exclude others. In addition to this precommitment attribute, the decentralized, collaborative innovation process underlying open source development provides advantages in addressing the myriad complex manifestations of flawed ("buggy") computer code. With more eyes and more uses, more bugs will surface, and those who find them can easily rewrite the code to fix them. Users may find it convenient to develop the code for their own idiosyncratic purposes, but the social value is much larger if the same code can also be used more broadly. The open-source movement exploits this potential by ensuring that all innovators in the open-source community make their source code available.

From the point of view of intellectual property, the open source movement is interesting because is uses copyright for a purpose opposite to its customary one, thereby spawning the term "copyleft." Through the use of a form of copyright license, a software developer can make his code available under a license which allows access and does not

require royalties, but requires, for example, that the user make his own derivative product available on exactly the same terms. Such licenses are called "viral"—products are infected with a self-replicating term that cannot be shaken by creating a new product. In this way, the community keeps the code open, observable, and useful to a broad community.

Who, in the end, pays for all this code if no one pays royalties? Lerner and Tirole (2000) stress career concerns coupled with personal uses, but admit that the economic models we currently have are not well adapted to explaining the phenomenon. None of the four models of the creative environment identified at the outset seems suitable. As stressed by von Hippel (2001), the system seems to work best in an environment where at least some developers have in-house uses. It is hard to see how it would work for strictly a mass market. The two most important examples, Apache and Linux, are notably less user friendly than their rivals in the mass market, e.g., the Microsoft server and Microsoft Windows operating system.

Recent developments, however, suggest that traditional industry players may well see support for open source software as a means of dethroning Microsoft from its long-standing monopoly position attributable to its widely adopted proprietary operating system products. The open source community's pre-commitment to non-proprietary software development creates a means for commercial enterprises that can profit from complementary products (hardware) and service businesses (such as consulting and maintenance)—including IBM and Hewlett-Packard, as well as newer companies such as Red Hat which specializes in supporting Linux—to move outside of the shadow of Microsoft's influence and compete more effectively in other computer business sectors (Merges, 2004b). Such "property preempting investments" may be a successful commercial business strategy.

Biomedicine is another realm where industry has organized to cut back on the exercise of intellectual property rights. Biomedicine has become heavily reliant on data and databases which contain gene sequences that may be patented or protected as trade secrets. Science as a whole is more efficient if researchers can share the data produced by others. Licensing such data piece by piece would impose prohibitive transactions costs. Instead, researchers are experimenting with collaborative business models to assemble the data into databases. In the SNP consortium, they have renounced intellectual property rights altogether (Merges, 2004b; Maurer, 2003).

1.8. Comparative analysis: intellectual property versus other funding mechanisms

Public and private funding mechanisms for R&D (and for creative works, to a lesser extent) have always existed side by side. In the U.S., the share of R&D funding provided by the public sector has seldom dropped below a third in the last half of the twentieth century, and in 2000 was about 26% (National Science Board, 2002). In most OECD countries the public share has been closer to half. Between the 19th century and the late 20th century, public funding has shifted from a system of *ad hoc* initiatives to a

routinized system based predominantly on peer review, with researchers competing for large federal budgets allocated before the recipients are named.

Since public sponsorship can reduce the restrictions on use that afflict intellectual property, and perhaps improve the way that R&D is organized, why isn't all R&D funded by the public sector? What accounts for the mix between public and private incentives? We return to those questions after commenting on the variety of public funding mechanisms currently in use.

A public funding mechanism that has been used more or less continuously throughout history is direct government employment of researchers. This is a system with obvious virtues when the sponsor is the only beneficiary of the resulting knowledge, or the benefits cannot be appropriated by a commercial vendor. However the defects of this system are many. Perhaps most importantly, it does not make use of the imagination that is widely dispersed in the population, and does not recognize that, for any given research task, some other party may be better equipped to perform it. It is an odd conception of research that starts from the premise that we know what we want to discover, we know how to discover it, and we know who can achieve it at least cost, namely, our employee. In what sense is that promoting discovery? In-house research will work rather badly in the ideas model, but much better in the model of induced change and the production-function model, since the investment opportunities are commonly known.

Like in-house research, prize systems also have a long history, continuing to the present. Prizes share several important features with patents. On the virtues side, they can attract investments from unexpected quarters, but on the defects side, will not reliably delegate the research effort to the most efficient firms. Prizes avoid deadweight loss, but prize authorities have two challenges that patents automatically avoid: the problem of choosing the value, and the problem of making it credible that they will, in fact, award the prize.

Of course, depending on what is observable, the flexibility to choose prize values can create an improvement over patents. A unifying theme is that, if a prize giver can base the prize on the value of the innovation, then he should do so, and prizes may dominate intellectual property rights (Wright, 1983; de Laat, 1996; Kremer, 1998; Scotchmer, 1999; Gallini and Scotchmer, 2002; Shavell and Van Ypersele, 2001; Abramowicz, 2003). Observations of the value can take many forms. Foray and Hilaire-Perez (2000) discuss how the silk-weaving guild in 18th-century Lyons used a prize committee to make judgments about value direction. Kremer (1998) argues that the value can be observed ex post by auctioning a patent right, and completing deal (transferring the patent right) with small probability, otherwise putting the invention in the public domain and giving a reward. Shavell and Van Ypersele (2001) suggest that the reward giver can link the reward to sales. Of course it may be more sensible to link the prize to the expected costs instead, and in this sense, prizes are more flexible than patents.

Maurer and Scotchmer (2004b) categorize prizes as *targeted* and *blue-sky*. For targeted prizes, such as the ones that were offered for solving the problem of longitude (Sobel, 1995), it is natural to assume that the sponsor, being the one to post the prize,

has a good idea of the social value, and can link the prize to it, or alternatively to the expected costs of achieving a solution. For blue-sky prizes, it is harder to tailor the prize to either cost or value. Blue-sky prizes are given for achievements that were not anticipated by the sponsor, so the prize cannot be established in advance.

Prizes can only work if the prize giver can commit not to renege, and will work best if the prize, like a patent, can increase with the social value of the invention. However costs are extremely hard to measure, especially when different inventions are supported with the same overhead and research projects have risky outcomes. Three possibilities for how to accomplish this are (i) to offer the prize against a backdrop of patents (Kremer, 1998), (ii) to structure the mechanism as a contest (Che and Gale, 2003), and (iii) to link the prize to performance standards as was sometimes done at Lyons. In addition, of course, there must be some means to ensure that the prize giver does not renege.

With patents as a fallback option, an inventor would not accept a prize less than the patent value. The prize value will thus be linked, like patents, to the value of the innovation. However, the prize giver must have some means to discover the value. A scheme suggested by Kremer (1998) is for the prize authority to take possession of the patent. The invention is put up for auction, although it is awarded to the highest bidder with only small probability. In most cases, it is dedicated to the public. But the small probability the patent is transferred to the highest bidder provides incentives for honest bidding, which yields the reward amount paid by the prize authority to the inventor (regardless of whether the invention is granted to the highest bidder).

A contest is a prize coupled with a commitment to give away the money, e.g., through the by-laws of a foundation or a trust. The commitment overcomes the problem of reneging. Nobel prizes are in that category. Contests can be structured so that the reward reflects costs instead of value. In the contest described by Che and Gale (2003), the contestants bid against each other before investing, making contingent contracts with the sponsor for what he will pay, conditional on choosing each contestant's invention *ex post*. The price only depends on which invention is chosen, and is thus easy to enforce. Because the firms compete on the contingent contracts, and will only be paid if chosen, they have an incentive to keep the contingent price low in order to be chosen. On the other hand, if a firm delivers a worthless innovation, the innovation will not be chosen even at a low contingent price. Such contests are sometimes called prototype contests, and they have been used by the U.S. Air Force, for example, in procuring fighter jets.

Prizes and contests share with intellectual property the inconvenience that the inventor is rewarded *ex post* rather than *ex ante*, and must therefore find funding. Government grants are a funding system that overcomes this problem. Of the R&D that is funded by the federal government, only about a quarter is performed in government laboratories. More than half of the National Institutes of Health budget and almost all of the National Science Foundation budget is given out as grants. Even the national labs, which used to be funded directly by the Department of Energy, now compete for funds in peer-reviewed grant processes. The government grant process improves on in-house research in that it taps into the scarce ideas likely to be found elsewhere.

As an incentive mechanism, the problem with grants is that applicants may propose research they cannot accomplish or wastes the funds. Since the whole point is to pay the costs of research as they are incurred, grant-giving organizations do not take the money back if the research fails, and have little recourse if the grantee wastes the funds (other than costly monitoring). But despite the limitations on oversight, the repetitive nature of grantsmanship exerts a discipline. A researcher can be cut out of the system if he does not deliver the research results he promised. For highly productive researchers, this threat will keep them honest, although the system will be more costly than if oversight could be exercised directly [see Maurer and Scotchmer (Ch 8 of Scotchmer, 2005b)].

Finally, again following Maurer and Scotchmer, Chapter 8 in Scotchmer (2005b), we turn to the "hybridization" of public and private institutions in the late 20th century. In year 2000 in the U.S., approximately 75% of total R&D was performed by industry, but only about 68% was funded by industry (National Science Board, 2002). Most of the rest was made up by the federal government. For most of the 20th century, more federal funding has gone to private firms than to universities, mostly from the Departments of Energy and Defense.

Not only is the private R&D sector infused with public money, but the public R&D sector is also infused with private money, at least if one includes universities and national labs as part of the public sector. That is, public and private funds are blended both in private industrial laboratories and in laboratories that have traditionally produced knowledge dedicated to the public. Further, the outputs of federally funded research are increasingly patented and exploited by the private sector under legislation enacted in the 1980's (the 1980 Bayh-Dole Act for universities, the 1980 Stevenson-Wydler Act for national labs, and the 1984 Technology Transfer Act), authorizing the creation of cooperative research and development agreements (CRADAs) (Mowery et al., 2001). This has turned out to be highly controversial. What is the rationale for subsidizing research that will ultimately be subject to intellectual property rights? Why give intellectual property rights on publicly funded research? The purpose stated in the Bayh-Dole Act is "to promote utilization of inventions arising from federally supported research or development ... without unduly encumbering future research and discovery" (Section 200). On the basis of this language, one might guess that the rest of the Act prohibits patenting, since patenting gives the right to restrict use. To the contrary, the point of the Act is to authorize patenting. The Bayh-Dole Act rests on the unlikely premise that the best way to diffuse innovations is to allow exclusions on use, subject to limited and rarely exercised "march in" rights (Eisenberg, 1996; Mikhail, 2000; Kieff, 2001b).

This contradiction is usually reconciled by positing that, without protection of the underlying science, firms will not make the collateral investments required to commercialize it. But it is a well established principle of patent law that improvements and new uses are themselves patentable. If so, this argument has no force. In any case, as many have argued, e.g., David (2003), Scotchmer (2003), Lemley (2004), it would be better to fix the thing that is broken (patent law) than to compromise open science.

We know of only one other justification for the policies that authorize private firms to leverage public money toward ownership of valuable intellectual property. Many gifts from the private sector are like matching funds that industry gives in return for intellectual property rights. Because industry can choose what to match, this system selects the projects that are likely to be commercially valuable, and thus serves the two purposes of allowing the public to subsidize expensive research while at the same time getting the benefit of private expertise in screening investments (Maurer and Scotchmer, 2004a).

The controversy over patenting discoveries in the university is not new. Such patenting has been going on since the late 19th century, at least for discoveries that were not funded by federal grants (Mowery and Rosenberg, 1998).

1.9. International treaties

In studying the optimal design of intellectual property, economists typically assume that the objective is to maximize consumers' surplus plus inventors' profit net of development costs. But whose consumers' surplus and whose profits? Externalities and profit flows across borders change the design problem. Domestic inventions create consumers' surplus abroad, and if protection is available abroad, also generate profit; for empirical evidence, see Alston (2002), McCalman (2001). In reverse, a strengthening of domestic protections will create an outflow of profit. How does this change the design problem?

The profit flows and externalities are governed by international treaties. Treaties create two types of obligations: for national treatment of foreign inventors and for certain harmonized protections. National treatment means that foreign inventors receive the same intellectual property protection as national inventors, while harmonization means that the countries have agreed on at least some aspects of what will be protected. Otherwise all countries could have different protections. These reciprocal obligations affect the rewards to innovation, the balance of trade, and foreign direct investment (Maskus, 2000a, 2000b). About half of American and European patents are issued to foreign inventors, and about 20% of Japanese patents (European Patent Office, 2002a, 2002b). The treaty obligations also extend to copyright, although there are no analogous administrative data that would allow us to asess their importance.

If the only objective is to minimize the deadweight loss of achieving a given amount of profit, then innovations should be protected in markets where the profit-to-deadweight-loss ratio is highest (Scotchmer, 2004b, Chapter 11). If this ratio is the same everywhere, then it does not matter for total deadweight loss where the profits are earned. However it obviously matters for equity. In any case, there is no policy maker with worldwide authority to make these decisions. Three arrangements have been tried: autarky, national treatment with independent choices of protections, and harmonized choices.

The first treaties to create reciprocal obligations for national treatment of copyrighted works and patented inventions were the Berne Convention and Paris Convention in the 1880's. It was not for another 100 years that significant strides toward harmonization were made, culminating in the 1994 TRIPs Agreement (Trade Related Aspects of In-

tellectual Property). In the meantime, the treaties, which had begun with about a dozen members, had grown to about 140 member states.

By autarky, we mean that each country protects only its own inventors. Autarky was the norm prior to the treaties of the 1880's. The main problem with autarky is that the market in any small country may be too small to cover the costs of innovations. If not, however, autarky can be a good system. Because inventors are protected where they are domiciled, and not elsewhere, inventors in different countries create reciprocal externalities. If the countries are more or less commensurate in size, these externalities more or less offset each other. On the other hand, autarky may not provide enough incentive. Reciprocal national treatment is a solution.

However, with national treatment, the fair solution where each country protects its own innovators is no longer possible. A jurisdiction can either protect a subject matter for both domestic and foreign inventors, or it can free ride, letting its inventors be rewarded abroad. There is no intermediate possibility (Scotchmer, 2004a). For a subject matter that a country chooses not to protect, its own consumers get the benefit of competitive supply, not only of its own inventions, but also of inventions made abroad. In the meantime, its own inventors collect profit in foreign markets. Free riding eventually led to the harmonization effort of TRIPS.

Suppose then, that the jurisdictions embark on a harmonization effort to coordinate their protections. One possibility is that they will simply harmonize to the efficient regime that a global optimizer would choose (Grossman and Lai, 2001). However, since there is no one in charge of a global optimization, it is more likely that individual countries will argue for harmonizations that serve their own interests (Scotchmer, 2004a). The harmonization that arises in actuality will be a negotiated solution from these preferred outcomes, and there is no presumption that they will be efficient. These papers conclude that harmonization will generally increase protections and that countries that advocate stronger protection (either more subject matters or longer protection) are those that either have large markets or are more innovative.

In the actual TRIPS negotiation, it was mainly the large, innovative countries like the U.S. that were behind the expansion of protections. This was apparently due to their innovativeness, and not due to their size. In fact there are small, innovative countries like Switzerland that were equally behind the expansion.

Finally, it is worth noticing how the public sector fits into this analysis. Whether domestic R&D is funded by private inventors or public sponsors, domestic discoveries create externalities for foreign users. The externalities are greater with public sponsorship, since foreign users will pay competitive prices rather than proprietary prices (assuming that intellectual property rights are not asserted by the public sponsor abroad if not asserted at home).

However, in choosing their policies, domestic policy makers are presumably not influenced by the benefits they confer on foreign users. They are more likely to be influenced by the prospect of repatriating some of those benefits as profit. The prospect of profit can shift the political balance in favor of private funding mechanisms, and cause innovative

countries to argue for protecting innovations that might otherwise be judged suitable for the public sector (Scotchmer, 2004b).

The treaties that have evolved leave scope for national autonomy. The harmonizations generally specify minimum required protections, but do not prohibit stronger ones. But whether stronger domestic protections can survive the international trading arena, especially in the digital age, is unclear. Protected products can typically be stopped at a national border, so that a rightholder can control its distribution domestically, even if not in the international market. However, for other types of intellectual property, such as research tools that can be used abroad to create products patented at home, the absence of foreign protection may undermine domestic protection as well. See Samuelson (2004).

2. Protecting integrity of the market

The second principal branch of intellectual property protection—relating to trademarks and unfair competition—focuses upon the quality of information in the marketplace. Quite unlike patent, copyright, and trade secret law, trademark law does not protect innovation or creativity directly. Rather, it aims to protect the integrity of the marketplace by prohibiting the use of marks associated with particular manufacturers in ways that would cause confusion as to the sources of the goods. In so doing, trademark law reduces consumer confusion and enhances incentives for firms to invest in activities (including R&D) that improve brand reputation. This function, however, is part of a larger framework of laws and institutions that regulate the quality of information in the marketplace.

The fact that trademark law does not directly protect technology or works of authorship does not mean that trademarks do not have significant value. The market value of most companies lies predominantly in the goodwill of the brand (e.g., Coca-Cola, Microsoft, Google). Although such goodwill is intertwined with the physical and other intangible assets of the trademark owner, there is little question that trademarks play a critical role in the value of many companies and that licensing of trademarks has become a major business in and of itself.

2.1. The economic problem

The efficiency of the marketplace depends critically upon the quality of information available to consumers. In markets in which the quality of goods are uniform or easily inspected at the time of purchase, consumers can determine the attributes themselves and no information problem arises. In many markets, however—such as used automobiles, computers, watches, as well as designer handbags—an information asymmetry exists: sellers typically have better information about the products or services being offered than buyers can readily inspect (Economides, 1998;

Akerlof, 1970). Unscrupulous sellers will be tempted to make false or misleading product claims or copy the trademark of a rival producer known for superior quality. It is often easier to copy a trademark than to duplicate production techniques, quality assurance programs, and the like. For example, two watches that look the same on the outside may have very different mechanical features, manufacturing quality, and composition of materials used.

Proliferation of unreliable information in the marketplace increases consumers' costs of search and distorts the provision of goods. Consumers will have to spend more time and effort inspecting goods, researching the product market, and actually testing products. Manufacturers will have less incentive to produce quality goods as others will be able to free-ride on such reputations. In markets for products where quality is costly to observe, high quality manufacturers may not be viable in equilibrium without effective mechanisms for policing the source of products and the accuracy of claims regarding unobservable product characteristics.

Several mechanisms are available to provide and regulate market information: (1) deceit and fraud common law causes of action and privately enforced consumer protection statutes; (2) public regulation and public enforcement of unfair competition laws; (3) trademark, false advertising, and deceptive practices/unfair competition laws; (4) industry self-regulation and certification organizations; and (5) consumer information institutions. Since our focus is on intellectual property law, we begin with an overview and analysis of trademark and related private bodies of unfair competition law. In many markets, trademarks provide a simple, quick, and effective means of communicating valuable product information. We conclude by discussing the role of trademark and unfair competition laws within the broader range of mechanisms for protecting the informational integrity of the marketplace.

2.2. An overview of trademark law

Trademarks have been in existence for nearly as long as commerce itself. Once economies progressed to the point where a merchant class specialized in making goods for sale or barter, the people who made and sold clothing and pottery began to "mark" their wares with a word or symbol identifying the maker. These early marks served several functions, including advertising, proof of the sources of goods (of relevance to resolving ownership disputes), and as an indication of the quality of goods. Modern trademark law has retained these functions. Trademarks reduce information and transaction costs in the marketplace by allowing customers to gauge the nature and quality of goods before they purchase them. Consumers rely most on trademarks where it is difficult to inspect a product quickly and cheaply to determine its quality.

Trademark law facilitates and enhances consumer decisions and encourages firms to supply quality products and services by protecting means of designating the source of commercial products and services. Thus, a trademark does not "depend upon novelty, invention, discovery, or any work of the brain. It requires no fancy or imagination, no genius, no laborious thought" [Trade-Mark Cases, 100 U.S. 82, 94 (1879)]. Rather,

trademark protection is awarded merely to those who were the first to use a distinctive mark in commerce. In trademark parlance, the senior (that is, first) user of a mark may prevent junior (subsequent) users from employing the same or a similar mark in such a manner as to cause a "likelihood of confusion" among consumers as to the source of the goods or services in question.

Traditionally, there has been nothing in trademark law analogous to the desire to encourage invention or creation that underlies patent and copyright law. There is no explicit federal policy to encourage the creation of more trademarks. Rather, the fundamental principles of trademark law have developed from two tort-based causes of action: the tort of misappropriation of the goodwill of the trademark owner and the tort of deception of the consumer. In this sense, trademarks should not be thought of as "property" rights at all. Rather, they are rights which are acquired with use of a trade mark in commerce[6] and derive protection based on the likelihood of indirect harm to potential purchasers of the trademark owner's products.

More recent legislation and several lines of cases, however, have introduced more of a "property" dimension to trademark law. Under the Federal Trademark Dilution Act of 1995, owners of "famous marks" may now prevent others from using their marks even in contexts in which there is no likelihood of consumer confusion. (Several states had previously enacted anti-dilution legislation.) Congress sought to protect such marks from blurring—the erosion of the distinctive quality of a mark through its adoption and use across a variety of product markets unrelated to the one(s) in which it developed fame—and tarnishment—uses that reduce the mark's positive association. The Act exempts uses in comparative advertising, noncommercial settings, and news reporting so as to address First Amendment concerns.

With the rise of the Internet and the establishment of a first-come, first-served registration system for domain names, so-called "cybersquatters" began registering the trademarks of others and either seeking to extort payments in exchange for transfer of the domains, offering such marks to competitors of the trademark holders, or setting up their own websites at these locations as a means of attracting business. The Anticybersquatting Consumer Protection Act, passed in 1999, imposes liability for registering, trafficking in, or using a domain name that is identical or similar to or dilutive of a trademark with bad faith intent to profit.

Table 2 summarizes the principal attributes of contemporary trademark law.

Trademark law thus consists of two principal branches. The traditional and still most important form of trademark protection provides remedies against the use of trademarks in ways that cause confusion in the marketplace as to the source of goods and services. Passing off or counterfeit goods—the marketing of goods displaying another's

[6] The Trademark Law Revision Act of 1988 changed this general principle in an important respect. Under that Act, it is now possible to register and protect a trademark based on an *intention* to use that mark in commerce within the next three years [15 U.S.C. §1051(b) (1988)]. Filing an Intent to Use application enables the applicant to establish a constructive priority date prior to actual use so long as the applicant proceeds to use the mark in commerce within the prescribed window.

Table 2

	Trademark
Underlying theory	perpetual protection for distinctive nonfunctional indications of origin of goods and services in order to protect consumers against confusion in the marketplace
Source of law	Lanham Act (federal); state statutes; common law (unfair competition)
Subject matter	trademarks (any designation of origin—including words, slogans, symbols, sounds, color); service marks; certification marks (e.g., Good Housekeeping Deal of Approval); collective marks (e.g., Toy Manufacturers of America); trade dress (product configuration and packaging)
Limitations	*no protection for* functional features, descriptive terms or geographic names (unless they have acquired (secondary) meaning—consumer recognition and association with a specific source), misleading aspects of marks, or names that have become "generic"—become associated with a general product category unconnected with any particular source (e.g., thermos)
Reqts for protection	priority (first to use in commerce); distinctiveness; acquired (secondary) meaning (for descriptive and geographic marks); use in commerce (minimal)
Process for obtaining protection	Use in commerce (or filing of intent to use application (establishing priority date) and subsequent use). Registration is optional, but confers various benefits (establishes *prima facie* evidence of validity (i.e., shifts burden of proof to defendant), constructive knowledge of registration, federal jurisdiction, mark becomes incontestable after 5 years of continuous use (i.e., cannot be found to lack secondary meaning), authorizes treble damages and atty fees, and right to bar imports bearing infringing mark) Examination (prior art search, assessment of requirements for protection) conducted by Trademark Office examiners Full opposition process for federally registered marks Maintenance fees for registered marks
Scope of protection	protection against uses that create a "likelihood of confusion" among an appreciable number of reasonably prudent consumers; dilution of famous marks; registration, with bad faith intent to profit, of domains names that are confusingly similar to trademarks; false advertising
Duration	perpetual subject to abandonment or loss of distinctiveness (genericide)
Marking	Notice (® for federal registration, ™ SM otherwise) optional, but confers benefits (burden of proof, remedies)
Rights of others	truthful reflection of source of product; fair and collateral use (e.g., comment, news reporting, comparative advertising)
Costs of protection	registration search; marking product (optional); litigation costs

(*continued on next page*)

Table 2
(*Continued*)

	Trademark
Licensing and assignment	no naked licenses (owner must monitor licensee); no sales of trademark "in gross" (i.e., without accompanying goodwill of associated manufacturer or service provider); licenses cannot be assigned without licensor consent
Remedies	injunction; accounting for profits; damages (potentially treble); attorney fees (in exceptionable cases); seizure and destruction of infringing goods; criminal prosecution for trafficking in counterfeit goods or services

mark without authorization—represents the classic and most common example of trademark liability. Anti-dilution protection—a second and more recently developed branch of trademark protection—protects famous marks against some forms of non-confusing uses of trademarks. The economic rationales for these forms of trademark protection differ and hence we take them up separately.

2.3. Confusion-based protection

2.3.1. Basic economics

Economic analysis of seller-provided information (advertising and trademarks) grows out of several fields of economic research and has evolved significantly over the past century. Early industrial organization economists were critical of advertising (and hence marking) on the ground that such activities "unnaturally" stimulated demand, thereby fostering and perpetuating oligopoly through "artificial" product differentiation. Reflecting his interest in monopolistic competition, Chamberlin (1933) viewed trademarks as a means for reinforcing monopoly power by differentiating products and thereby excluding others from using the differentiating characteristic, even if only a mark. By generating a downward sloping demand curve for its brand, trademark owners could under this view generate monopoly rents (and resulting deadweight loss) (Robinson, 1933, p. 89; Comanor and Wilson, 1974; McClure, 1979, 1996; Lunney, 1999). Drawing upon this literature, Brown (1948) tied the analysis of trade symbols to the larger context of commercial advertising, which he considered to serve both useful (informative) and wasteful (persuasive—intended to suggest that one product is superior to a similar if not identical alternative) ends. This led him to approach trademark protection with ambivalence and caution.

The emergence of the modern information economics literature in the 1960s and 1970s offered a more productive view of the role of advertising in markets (Stigler, 1961; Nelson, 1970, 1974, 1975; Hirshleifer, 1973; Nagle, 1981). Trademarks, as a concise and unequivocal indicator of the source (e.g., Intel) and nature (e.g., Pentium) of particular goods, counteract the "market for lemons" problem (Akerlof, 1970) by

communicating to consumers the enterprise which is responsible for the goods and, in some cases, the specifications of the goods (Landes and Posner, 2003). The brand name Coca-Cola, for example, informs the consumer of the maker of the soft drink beverage as well as the taste that they can expect. If the product lives up to or exceeds expectations, then the trademark owner gains a loyal customer who will be willing to pay a premium in future transactions; if the product disappoints, then the trademark owner will have more difficulty making future sales to that consumer (or will have to offer a discount to attract their business). In this way, trademarks implicitly communicate unobservable characteristics about the quality of branded products, thereby fostering incentives for firms to invest in product quality, even when such attributes are not directly observable prior to a purchasing decision (Klein and Leffler, 1981; Hirshleifer, 1973; Shapiro, 1982, 1983; Milgrom and Roberts, 1986; Economides, 1998). Sellers who enter the high quality segment of the market must initially invest in building a strong reputation. Only after consumers become acquainted with the attributes of their brand can they recoup these costs. In equilibrium, therefore, high quality items sell for a premium above their costs of production to compensate for the initial investment in reputation (Shapiro, 1983). Trademarks also facilitate efficient new business models, such as franchising, which generate economies of scale and scope in marketing and facilitate rapid business diffusion across vast geographic areas (Wilkins, 1992; Williamson, 1986).

The marking of products also creates incentives for disreputable sellers to pass off their own wares as the goods of better respected manufacturers. Trademark law (as well as false advertising and unfair competition laws more generally) harnesses the incentives of sellers in the marketplace to police the use of marks and advertising claims of competitors. Sellers often have the best information about the quality of products in the marketplace; they also have a direct stake in preventing competitors from free riding on their brand, reputation, and consumer loyalty. By creating private causes of action, trademark and false advertising law take advantage of this informational base and incentive structure as well as the vast decentralized enforcement resources of trademark owners to regulate the informational marketplace, effectively in the name of consumers.

Under this now widely accepted view of consumer information economics, trademarks economize on consumer search costs (McClure, 1996; Kratzke, 1991; Economides, 1998; Landes and Posner, 1987). Consumers benefit from concise and effective designations of the source of products. For example, consumers can quickly assess the attributes of a computer made by Sony featuring an Intel Pentium Processor and Microsoft's XP Operating System. If such trademarks were not available or could not be relied upon, the consumer would have to incur substantial additional costs in shopping for a computer. The ability to establish and maintain reliable trademarks reinforces firms' desire to develop and maintain consistent quality standards. It also fosters competition among firms over a wide quality and variety spectrum (Economides, 1998).

In general, consumers distinguish among three types of product features: search attributes, such as color and price, which can be inspected prior to purchase; experience

attributes, such as taste, which can only be verified through use of the product (typically after purchase); and credence attributes, such as durability, which can only be verified over time (or through the use of surrogate sources of information—e.g., Consumer Reports) (Nelson, 1974; Darby and Karni, 1973; Bone, 2004). Brands signal experience and credence attributes. In an empirical study of branded and unbranded gasoline service stations, Png and Reitman (1995) found that branded dealers were more likely to carry products for which quality was more difficult to verify and to serve customers who placed a higher value on search.

Some trademarks also serve a more ambiguous function: signaling status or identity for some consumers. Some have referred to such commodities as "Veblen" goods, reflecting Thorstein Veblen's theory of conspicuous consumption. This theory posits that demand for status goods rise with increases in price (Leibenstein, 1950; Veblen, 1899). Purchasers of such goods may be interested in being associated with a particular brand—such as a Rolex watch, a t-shirt with the name and colors of a particular university, or a corporate brand—possibly apart from whether it is authentic or the quality associated with the authentic good (Kozinski, 1993; Dreyfuss, 1990; Higgins and Rubin, 1986). Some purchasers of such goods may well prefer a less expensive, counterfeit version. They presumably would not be confused when purchasing such goods (e.g., a Rolex watch sold on a street corner for $10).

The marketing of less expensive, lower quality imitations of status goods creates the possibility of separate harm to the sellers and purchasers of authentic goods. The availability of counterfeit articles could well divert some consumers who would otherwise purchase the authentic article, although this effect is likely to be relatively small due to the large price differential and the availability of the authentic goods for those who are interested. The lower quality of the counterfeit goods could, however, erode the goodwill associated with the authentic manufacturer through post-sale confusion—on-lookers who mistake the shoddier counterfeit good for the authentic good and are thereby less inclined to purchase the authentic version, thereby reducing sales by the trademark owner. In addition, due to the proliferation of non-easily recognized "fakes," prior and potential purchasers of the authentic "status" goods may be less interested in owning a much less rare commodity. The value of ownership may be sullied. In essence, status goods exhibit a negative network externality, whereby proliferation of such goods erodes the value to prior purchasers (Higgins and Rubin, 1986; Kozinski, 1993; Dogan and Lemley, 2004b). The significance of these harms is considered speculative. See Lunney (1999) (questioning the economic basis for protecting status values).

Notwithstanding the general benefits afforded by trademarks, such protection entails several types of costs. Protection of descriptive terms as trademark can increase search costs and impair competition by raising the marketing costs of competitors. For example, if a cookie manufacturer were to obtain a trademark on the word "cookie," then other companies interested in selling cookies would have a much more difficult time communicating the nature of their goods to consumers. If, however, the trademark was

to "Mrs. Fields Cookies" and any protection for "cookies" was disclaimed, then potential competitors would be able to describe their products in the most easily recognized manner and would be able to develop their own marks—such as "ACME Cookies." At a minimum, trademark protection for descriptive terms significantly reduces the effective range of terms that may be used by others.

A complicating factor in the protection of trademarks is the endogeneity of the usage and meaning of terms and symbols over time. Even a distinctive term can become "generic" (common) if consumers come to associate marks with a particular product (as opposed to its manufacturer). The evolution of the use of the term "thermos" illustrates this phenomenon. At the turn of the twentieth century, the original manufacturer of vacuum-insulated flasks selected the term "Thermos"—derived from the word "therme" meaning "heat"—to brand its product. At the time that it was selected (in effect, coined), Thermos was distinctive and not associated with any particular product. The American Thermos Bottle Company, which acquired the U.S. patent rights for this technology, undertook advertising and educational campaigns that tended to make "thermos" a generic term descriptive of vacuum-insulated flasks rather than of its origin. After the patents expired, other manufacturers began using this term to describe their own vacumm-insulated flask products. As we will discuss further below, use of the term became generic in the eyes of consumers, and hence the law, and the original manufacturer of the product (and developer of the mark) lost trademark protection [King-Seeley Thermos Co. v. Aladdin Industries, Inc., 321 F.2d 577 (2d Cir. 1963)].

More generally, trademark protection for descriptive terms can impede competition. Gaining control over the most effective term for describing a product raises the costs of potential competitors seeking to sell in that marketplace. By not being able to use a term or means of communication most easily understood by the consuming public, the entrant must bear higher marketing costs. Limitations on the use of trademarked terms for purposes of comparative advertising would also impede vigorous competition. Trademark law is least problematic at its traditional core: protecting inherently distinctive (i.e., non-descriptive) source identifying marks against directly competing uses that confuse consumers. The expansion of trademark protection to encompass non-competing products, dilution (non-confusing uses of famous marks), product configuration and packaging (trade dress), merchandising of trademarks (mere sponsorship), post-sale confusion, and more distant reputation zones have increased the tension between trademark protection on the one hand and competition and innovation policy on the other (Lemley, 1999; Lunney, 1999; Bone, 2004).

Trademark protection can also interfere with both communicative and creative expression. Broad exclusive trademark rights would limit the ability of others (including non-competitors) to comment on and poke fun at trademarks and their owners. As we will see below, various doctrines limit such adverse effects. But as trademark protection has expanded beyond the traditional core—for example, to encompass a broad conception of connection to, sponsorship, and affiliation with a trademark owner—it becomes more difficult to assess the boundaries, leading film and television production com-

panies, for example, to tread carefully (and increasingly incur the costs of licensing transactions) in the use of trademarks in their works.[7]

As with other modes of intellectual property, trademark protection also involves administrative and maintenance costs. Although the costs of acquiring trademark protection is relatively low, mark owners must police their marks to prevent use of the marks without authorization and supervise licensees to ensure that quality standards are maintained. As a mark enters common parlance and becomes associated in the minds of consumers with a general product category as opposed to the manufacturing source—as in the Thermos example—the owner must invest in advertising to clarify that the mark is associated with a particular supplier in order to prevent "genericide"—the death of a trademark due to its becoming generic. For many years, Xerox spent large sums on advertising to discourage generic usage of the term "xerox" as noun or verb for photocopying. Google faces a similar exposure today.

2.3.2. Policy levers

As summarized in Section 2.2, trademark law has evolved into a complex system of administrative rules and judicial doctrines. Such rules and doctrines can be seen as a series of policy levers that may be crafted by legislators and courts and administered by the Trademark Office (through a registration system) and the courts. The economic analysis of particular trademark doctrines (policy levers) focuses on balancing the salutary effects of trademarks with the various costs: constraints on the availability of language and symbols to economize on consumer search, protection costs (administrative, maintenance, and enforcement), anticompetitive effects, and limitations on the freedom of creative and communicative expression. We examine the administration of the trademark system in the following section.

Before turning to the analysis of specific rules governing trademark law, it is useful to re-emphasize one of the observations made at the outset: Although the "real property" metaphor provides some useful insights for understanding intellectual property law, simply extrapolating from economic analysis of real property overlooks important distinctions. Trademarks serve a different set of purposes than real property law and operate in a much more diffuse environment (control of words and symbols). Perhaps most significantly, trademark law standards use the public's perception of the meaning of words and symbols as the touchstone for determining the rights (validity, breadth, infringement, and duration) and limitations of trademark owners. As one jurist (Kozinski, 1993) has aptly noted:

[7] As a reflection of the growing importance of brand exposure and image creation, film and television production companies view product placements as an advertising revenue source. Nonetheless, to the extent that trademark law requires licensing of trademarks for such works, there is a cost (and potential distortion) imposed on the creative process. The fact that some trademark owners are willing to pay for exposure alleviates but does not eliminate this concern. Although some mark owners would compete for product placements even if the trademark licensing was not required, creators are hampered to the extent that they prefer to use marks that are not available for licensing.

> Words and images do not worm their way into our discourse by accident; they're generally thrust there by well-orchestrated campaigns intended to burn them into our collective consciousness. Having embarked on that endeavor, the originator of the symbol necessarily—and justly—must give up some measure of control. The originator must understand that the mark or symbol or image is no longer entirely its own, and that in some sense it also belongs to all those other minds who have received and integrated it. This does not imply a total loss of control, however, only that the public's right to make use of the word or image must be considered in the balance as we decide what rights the owner is entitled to assert.

2.3.2.1. Threshold requirements The three principal requirements for establishing trademark protection—distinctiveness, priority, and use in commerce—can all be understood to reflect economic considerations.

Distinctiveness. Trademark law affords protection to the first enterprise to use fanciful ("Kodak" (photographic products)), arbitrary ("Apple" (computers)), and suggestive (requiring a leap of imagination by the consumer, such "Chicken of the Sea" (tuna) or "cyclone" (braided wire fencing)) marks as soon as they are used in commerce, whether or not they are registered. (We discuss registration of trademarks in Section 2.5 below.) Descriptive terms (such as "Digital" for computers), surnames (McDonalds), and geographical designations (e.g., New York Times) are protectable only upon acquiring secondary meaning (denoting a single seller or source) in the minds of a substantial portion of the relevant consumer marketplace. Generic terms are ineligible for protection, reflecting the idea that search costs to consumers would be greater if new entrants could not use the common meanings to label and advertise their products.

Affording automatic protection to inherently distinctive marks (fanciful, arbitrary, and suggestive terms) rather than awaiting proof of secondary meaning can be justified on process, error cost, and predictability grounds (Bone, 2004; Denicola, 1999; Landes and Posner, 1987) (defending categorization of marks versus case-by-case balancing as saving administrative and dispute resolution costs). Proving secondary meaning requires relatively time consuming and costly consumer surveys. Moreover, inherently distinctive terms are plentiful in supply (an infinite number of fanciful and arbitrary terms are available) and hence potential entrants would not be constrained in any significant way by the removal of such terms from the universe of potential inherently distinctive marks. Providing automatic protection for such terms reduces the costs of entry and enables firms to make investments in developing brand equity secure in the knowledge that their mark will be valid (assuming priority of use, which can be assessed through a relatively quick and inexpensive trademark search).

By contrast, affording protection to descriptive terms (including geographic designations and surnames) before such terms became associated with a particular source would raise consumer search costs and impose undue barriers to entry by competitors. Effective descriptive terms are limited in supply (Carter, 1990) and therefore any restriction on the use of such terms by consumers and potential entrants raises search costs.

As Burge (1984, p. 126) notes, "suggestive and descriptive marks tend to be preferred by advertising people because these marks are thought to enhance initial product salability." But once a descriptive term becomes associated with a source—e.g., New York Times, "Bed & Bath" (home products store), "Chap-Stick" (lip balm), "McDonalds" (fast food)—allowing entrants to adopt identical or similar designations risks confusion in the marketplace. Trademark law balances these costs by delaying the time at which such marks can be protected until sufficient consumer recognition has been achieved.

In an interesting judicial use of the distinctiveness threshold as a policy lever, the Supreme Court has ruled that product configurations (as opposed to mere product packaging) can never be inherently distinctive—i.e., acquired meaning must always be established in order to obtain protection [Wal-Mart Stores, Inc. v. Samara Brothers, Inc., 529 U.S. 205 (2000)]. In so doing, the Court expressly used this tool to encourage competition in product markets by requiring express proof that product configurations, even if arbitrary, functioning as trademarks have acquired meaning in the minds of consumer before receiving protection. Note that this requirement is in addition to the separate rule, which we discuss below, that functional elements of products may not be protected as trademarks.

Priority and use in commerce. Although registration of trademarks is optional (see Section 2.5 relating to administration), trademark rights are accorded to the first user of a mark in commerce. Such a rule discourages rent seeking, such as the stockpiling of names for subsequent resale or the locking up of a large segment of the useful semiotic domain. Landes and Posner (1987); Carter (1990). Pure registration systems—such as the Japanese trademark system and the domain name registration for the Internet—have produced rent seeking behavior resulting in the warehousing of terms, making it more costly for others to enter markets (Landes and Posner, 2003, pp. 179–180). The use requirement also serves a notice function.

The use requirement can be criticized on economic grounds as being both too lax and too strict. Under current rules, even token use suffices to establish priority and with registration merely optional, the notice function may not be adequately served and banking of potential terms is still possible at relatively modest cost. On the other hand, requiring actual use exposes companies planning large product introductions to some risk that their mark could be preempted on the eve of the announcement. Such risk adds needless uncertainty. The introduction of the Intent to Use application process addressed this problem by enabling companies to obtain a certain priority date for a trade name in advance of use in commerce so long as use follows within a six month period (with extension possible for a total of up to three years). Carter (1990) has expressed concern that this system provides undue potential for anticompetitive warehousing behavior and calls for imposition of penalties where it appears that a registrant has filed numerous intent-to-use applications without a serious intention to use such marks in commerce.

2.3.2.2. Duration Given the primary purpose of trademark law of reducing consumer search costs, there is a strong justification for trademark protection lasting as long as

a mark represents a reliable designation of source of goods and services. Due to the infinite availability of arbitrary and fanciful marks, perpetual protection for trademarks does not prevent others from entering the market. And to the extent consumers connect a descriptive term to a particular source, confusion would result from expiration of that mark while the developer of the mark continues to operate under that name or logo. Unlike with copyrights or patents, there is no concern about perpetual duration hindering cumulative innovation by others because trademark protection does not extend to functionality or creativity per se. Limiting doctrines allow others to make some expressive usage of trademarks—e.g., for comparative advertising and social commentary.

Trademarks will lose protection, however, if they become generic. At the point at which an appreciable number of consumer associate a mark with a product category as opposed to a source, allowing one manufacturer exclusive rights to the mark raises consumer search costs and the marketing costs of competitors. We discuss genericide more fully in the section on rights of others (and defenses).

Trademarks will also lose protection through voluntary abandonment or dramatic changes in product quality. Abandonment occurs through a trademark owner exiting a business (without transferring the mark along with the associated goodwill to another firm). Whereas trademark owners may evolve their products and product quality, they may not so dramatically change the quality or nature of a product sold under the mark (without appropriate warnings) as to constitute fraud upon the consuming public. This prohibition discourages deceptive opportunism. Other consumer protection statutes—protecting against deception and fraud—potentially address this form of deception as well.

Once a mark is abandoned, it becomes fair game for new entrants or existing manufacturers. Such a doctrine could cause confusion in the marketplace to the extent that a new user of a recently abandoned trademark offers goods of different quality than the prior mark owner. For this reason, trademark law requires new users of abandoned marks to take reasonable precautions to prevent confusion until such time as the association with the prior supplier has faded from the public's lexicon. Other consumer protection statutes may also discourage deceptive practices that may occur following a change in trademark ownership.

2.3.2.3. Ownership and transfer rules The assignment and licensing of trademarks presents a problem for maintaining the integrity of quality standards, and hence the expectations of consumers. To allow free alienability of trademarks—as is permitted for conventional forms of property as well as patents and copyrights—could jeopardize the quality assurance implicit within the nature of a trademark. For this reason, U.S. trademark law prohibits marks to be assigned "in gross"—i.e., without the good will underlying the mark (including the right to produce the goods sold under the mark)—or licensed without ongoing supervision by the trademark owner. Such restrictions on alienability discourage "end game" opportunism—selling the mark at a premium upon exiting the trade (or entering bankruptcy) to a company that intends to reap a premium on the sale of shoddy goods.

The concern with this opportunism scenario appears to be overblown. Cf. McCarthy (2004, §18.10); McDade (1998). If a mark has value in the market, then the assignee/purchaser of the mark will jeopardize the long term value by lowering quality standards. Moreover, as noted above, radical changes in the quality of goods sold under a mark could result in abandonment. In fact, most other nations permit assignments of trademarks in gross,[8] suggesting that a rule regulating transfers may be unnecessary. Similarly in the licensing context, trademark licensors ultimately bear the costs of erosion in brand equity and therefore have strong incentives to put in place efficient supervisory systems to maintain or enhance brand equity without additional legal constraints—i.e., it is not at all clear that there is an externality. More generally, significant changes in products or services following assignment or under licensing agreements could result in liability for fraud or deceit or exposure under publicly enforced consumer protection statutes.

2.3.2.4. Breadth and infringement analysis As we saw earlier, patent and copyright law afford exclusive rights for purposes of promoting investment in the development of new works. Hence, infringement analysis focuses on a comparison of the elements of the protected work (the patent claims or the copyrighted book, musical composition, or other work) and the allegedly infringing work. By contrast, trademark law does not grant exclusive rights but rather limits protection to the purpose of protecting consumers against confusion as to the source goods or services. The touchstone for trademark protection is *consumer perception*—whether an appreciable number of reasonably prudent consumers perceive the defendant's product or services to be sponsored by, affiliated with, or otherwise connected to the trademark owner. This standard tailors the scope of trademark protection to the consumer search cost rationale, leaving freedom for competitors and others to use marks in ways that are not likely to cause consumer confusion.

The shift in focus from comparing protected and allegedly infringing works (irrespective of locus of use (so long as it is in the United States for patents) and product market) as in patent and copyright law to assessing consumer confusion requires a multidimensional framework. The scope of trademark protection can be thought of spatially along semiotic (linguistic and symbolic), product market, and geographic dimensions. To illustrate this framework, consider the trademark of the ACME Bread Company in Berkeley, California. Along the semiotic dimension, we can imagine a spectrum of marks from ACM to ECME to ACMF which bear some resemblance to ACME—all selling bread in Berkeley community. Along the product dimension, we can envision different companies in Berkeley operating under the ACME selling baked goods (pastries as well as bread), groceries, office furniture, as well as fishing supplies. Along the geographic dimension, we can imagine ACME Bread Companies (with different owners) in neighboring Oakland, California, St. Louis, Missouri, or Atlanta, Georgia.

[8] Article 21 of the Trade-Related Aspects of International Trade (TRIPS) agreement permits the owner of a registered trademark to assign the mark without transferring the "business" associated with the mark.

Which of these competing businesses, if any—from the ECME Bread Company in Berkeley to the ACME Fishing Supply Company in Berkeley to the ACME Bread Company in Atlanta Georgia—infringes the trademark owned by the ACME Bread Company in Berkeley[9]? Under early trademark law, protection was limited to goods of the same descriptive class—i.e., directly competing goods. Since 1946, however, protection has encompassed confusion as to origin, sponsorship, approval, and connection, whether or not goods are in direct competition. Thus, modern trademark law does not provide categorical answers to the scope of protection. Rather it determines liability and hence scope of protection on the basis of a comprehensive, fact-intensive examination of a wide range of factors bearing on the perceptions of reasonably prudent consumers in the relevant marketplace. Under modern tests, courts look to the following non-exhaustive list of factors:

- characteristics of the trademark (inherent distinctiveness, acquired meaning)
- characteristics of the allegedly infringing mark (similarity to the plaintiff's mark)
- marketplace considerations:
  - strength of the senior user's mark
  - nature of the product market (low cost versus high cost products; care exercised by consumers)
  - proximity of the goods
  - likelihood of expansion of either party into the other's product market
  - channels of trade and methods of distribution
  - advertising and promotion
  - nature and sophistication of consumers
- evidence of actual consumer confusion (e.g., misdirected service calls by consumers, testimonial evidence, surveys)
- evidence of bad faith (e.g., intentional copying of mark) by the junior user.

Over the past several decades, the effective scope of trademark protection has expanded to encompass promotional goods (enabling universities, sports teams, and corporate sponsors to enjoin clothing manufacturers from selling t-shirts emblazoned with trademarks without authorization), initial interest confusion (whereby consumers may be only initially confused as to source, but not at the time of purchase), post-sale confusion, and trade dress (product configuration and packaging). This has led some commentators to believe that trademark law has gone beyond the boundaries necessary to optimize consumer search costs and increasingly threatens competition [Lunney, 1999; Lemley, 1999; Bone, 2004 (suggesting that disclaimers ought to be more readily credited, especially in the case of promotional goods, and that trade dress protection be abolished); Dogan and Lemley, 2004a]. The application of the trademark law to Internet activities has continued this trend, with courts focusing on a rather limited range of

[9] We are assuming here that the ACME Bread Company in Berkeley has priority over these other enterprises—i.e., it was the first to use the ACME trademark in commerce. As an arbitrary mark, ACME would have received protection upon initial use.

factors for determining infringement (similarity of marks and relatedness of goods) and finding infringement on the basis of initial interest confusion readily, notwithstanding the rather modest costs of redirecting Internet searches (Dogan and Lemley, 2004b).

2.3.2.5. Breadth and the rights of others The other principal policy levers affecting the scope of trademark protection relate to exceptions and defenses to liability. Several doctrines limit the scope of protection in order to promote competition, innovation, and freedom of communicative and creative expression.

Functionality. The expansion of trademark law to encompass product configurations brought trademark's regime of perpetual protection on the basis of relatively low validity requirements potentially into conflict with patent law's exacting validity requirements and limited duration. Without appropriate limitations, trademark law could protect subpatentable technologies as well as extend protection for patented technologies beyond the expiration of the patents. To avoid upsetting the balances of the patent system, courts developed a rule that functional product features—defined as those elements that are essential to the use or purpose of a product that affect its cost or quality—cannot serve as a trademark. The aesthetic functionality doctrine applies a comparable channeling principle with regard to copyrightable product elements, such as pottery and silverware designs.

Parchomovsky and Siegelman (2002) show that the ability to protect distinctive functional features of patented technologies under trademark law beyond the expiration of a patent induces the patentee to moderate its pricing during the term of the patent in order to foster brand loyalty. This effect offsets to some extent the static deadweight loss of patent protection in anticompetitive effects of allowing perpetual trademark protection for functional product features. As they note, however, the optimal level of trademark leveraging will vary across patented technologies and policymakers will often lack the information needed to tailor the balance optimally.

Genericide. Consumer perception of the meaning of words and symbols can change over time, sometimes resulting in trademarks drifting from designating the source of a good to becoming a generic means of describing a category of products. Thermos, yo-yo, escalator, refrigerator, and aspirin have all made this transition. Once a substantial percentage of consumers come to treat a term as a generic product category rather than a brand designation, consumer search costs are raised (costs of having to communicate around a well-known, but protected, term) and undue market power conferred by allowing but one manufacturer to control the use of the term. For example, if the term "plexiglass" could not be freely used, competitors would have to resort to rather prolix expressions such as "unbreakable clear plastic sheets that function as glass" in order to describe their products (Merges, Menell, and Lemley, 2003, p. 685; Landes and Posner, 1987, p. 292). Such a mouth-full raises the costs of advertising and would likely engender significant consumer confusion as a result of consumers inferring that the purveyor must not mean "plexiglass" because that would obviously have been easier to convey.

In recognition of this phenomenon, the genericide doctrine strips trademark protection from terms whose primary significance in the minds of the consuming public signifies a general product category rather than a particular product sold by a manufacturer, even if the originator of the term put substantial effort into creating it and encouraging its use.. Although commentators differ over the appropriate standard for determining when a term has become generic—with some favoring expressly economic formulations [Folsom and Teply, 1980, 1988a, 1988b; Coverdale, 1984 (advocating use of antitrust-type cross-elasticities of demand approach for determining the degree of substitutability among terms); Landes and Posner, 1987 (proposing cost-benefit test)], and others favoring more conventional formulations on practical grounds (Swann, 1980; Swann and Palladino, 1988; Oddi, 1988)—there is general consensus that the genericide principle economizes consumer search costs.

Fair use and nominative use. Notwithstanding the protectability of descriptive marks, geographic designations, and personal names that acquire secondary meaning, trademark law balances the resulting constraint on the use of commonly understood terms by allowing competitors to make "fair use" of the protected terms to describe their own goods or services, their geographic origin, or the names of people involved in their own business. The nominative use doctrine allows others to use a protected mark to describe the mark owner's product, as, for example, in comparative advertising or in non-trademark usages. Allowing such uses reduces consumer search costs by making it easier to communicate relevant information to consumers, thereby promoting free competition and use of language.

Use in commerce and indirect liability. Trademark liability can only be imposed where a competitor uses a mark in advertising or commerce "causing the public to see the protected mark and associate the infringer's goods or services with those of the mark holder" [DaimlerChrysler AG v, Bloom, 315 F.3d 932, 939 (8th Cir. 2003)]. This doctrine has come under scrutiny with the emergence of Internet search technologies and business models. Website developers often insert hidden codes, such as metatags, that are used by web search engines to index web sites based on relevance of search queries. The question arises whether the placement of a competitor's trademark into a metatag constitutes a use in commerce. Similarly, search engine companies, such as Yahoo and Google, that deliver sponsored advertisements based on search queries derive a substantial portion of their revenue by selling keyword advertising placements. Does the sale of such keyword advertising placements constitute use of the terms in commerce? Such keyword advertising placements can be seen as a form of free-riding, seeking to divert web surfers looking for links to an established trademark; they can also be viewed as general inquiry into the commercial marketplace—a proxy for a range of relevant sites. Dogan and Lemley (2004b) advocate tying the liability for trademark infringement to the search cost rationale—only those who are using the mark to advertise their own wares or services have the motive and opportunity to interfere with the clarity of the mark's meaning in conveying production information to consumers. Hence, they would permit search engines to escape liability.

Freedom of expression. Courts recognize a First Amendment defense to trademark infringement where another seeks to use a mark to communicate ideas or express points of view. One court recently held that a song entitled "Barbie Girl," that poked fun at the Mattel Corporation's doll of the same name, did not infringe the trademark [Mattel, Inc. v. MCA Records, 296 F.3d 894 (9th Cir. 2002)].

2.3.2.6. Remedies Courts award injunctive relief as a matter of course upon a showing of likelihood of consumer confusion. There is no requirement of actual confusion [Bone, 2004 (suggesting that such rules may be justified by process cost considerations)]. Monetary relief (actual damages, lost profits, the defendant's profits attributable to the infringement, punitive damages in cases of willful infringement, and attorney fees) is also available, although damages are often difficult to quantify. In 1984, strong federal criminal sanctions along with public enforcement was put in place in order to stem a growing tide of international trademark counterfeiting. Due to the unique federal role in and resources for policing international borders, public enforcement of trademark counterfeiting offered significant economies of scale and scope over private enforcement by individual trademark owners.

2.4. Dilution-based protection

2.4.1. Basic economics

The economic rationales for dilution grow out of the same considerations applicable to confusion-based trademark liability—reducing consumer search costs and fostering investment in product quality and brand equity—although the concerns are somewhat more attenuated. The principal problem to which dilution protection is addressed concerns blurring (loss of distinctiveness) of brand identity (Schechter, 1927). As consumers develop their mental lexicon of brands, they associate both specific products and general attributes with particular trademarks. For example, Rolls Royce connotes both the source of a luxury automobile as well as a brand of uncompromising quality and ornate styling (as well as high cost). If another company were to introduce Rolls Royce candy bars, it is unlikely that many (if any) consumers would believe that the automobile manufacturer was the source. Whether intended or not, the candy company may benefit from the particular general attributes that the consuming public associates with the Rolls Royce brand. They may also gain some "status" equity to the extent that consumers value the signal associated with a mark. Thus, adopting the Rolls Royce name enables the newcomer some ability to free-ride on the general brand reputation of the famous trademark owner.

Such use, however, would impose some costs on consumers and the famous trademark owner. As this new use of the Rolls Royce term gained popularity, the association between the mark and a particular source would become blurred. Furthermore, as more companies in unrelated markets adopt this moniker—Rolls Royce tennis racquets, Rolls Royce landscaping, Rolls Royce tacos—the distinctive quality of the mark

would become further eroded. Over time, consumers would lose the non-product specific identity (i.e., Rolls Royce as a brand of uncompromising quality and ornate styling) that the original Rolls Royce mark once evoked. This raises consumers' search costs: consumers' mental lexicon has become more difficult to parse. A lack of protection against trademark dilution could weaken the incentives of suppliers to invest in and maintain their brand equity, although this effect is likely to be quite attenuated in most circumstances. Owners of famous marks have strong incentives to maintain and enhance their brand equity even without formal protection against dilution. Nonetheless, the full benefits of their investment are not internalized due to potential free-riding by others. Protection against blurring of famous marks has some parallels to the prospect and rent dissipation theories of intellectual property protection (Kitch, 1977; Grady and Alexander, 1992). Upon establishing a famous mark, the owner obtains broad scope for further developing the intellectual property right.

A second form of dilution relates to the tarnishment of a well-known brand. If the maker of pornographic films were to sell their movies under the brand "Disney," it is unlikely that consumers would believe that the Disney Corporation, famous for family oriented entertainment, was the manufacturer of such unwholesome products. Nonetheless, consumers' shopping lexicon would arguably be distorted because the Disney name would trigger associations with both family oriented content and smut. Such a negative association could well injure the Disney Corporation's brand equity. As with blurring, tarnishment interferes with established associations. Perhaps even more so than blurring, it undermines brand equity.

Anti-dilution protection prevents this erosion of the distinctive quality of a mark by prohibiting famous marks from being used by others—even in unrelated product markets and in non-confusing ways. The Rolls Royce automobile manufacturing company can preclude the marketing of Rolls Royce candies without its authorization. Disney can prevent pornographers from adopting the Disney name. This preserves distinctive brands and affords the owners exclusive rights to carry their brand names into wholly new markets (or not). We see examples of such brand migration in many markets. Sony Corporation, for example, which honed its reputation in the consumer electronics marketplace, has now developed products in the sound recording and motion picture marketplaces.[10] Cross-branding, such as the marketing of a Barbie doll adorned with Coca-Cola's logo and a distinctive red ensemble, is also increasingly common.

Expanding trademark law to protect against dilution of marks can impose several costs. Dilution law could operate to keep otherwise generic terms from being available to all. This marginally increases consumer search costs and raises the marketing costs of other companies. In effect, dilution law could conceivably constrain use of the language. Beyond this concern, to the extent that transaction costs discourage some

[10] The expansion of traditional trademark protection has, to some extent, afforded protection against diluting uses of trademarks. Courts consider the likelihood that a trademark owner would expand into a new market in determining infringement.

valuable licensing of "dilutive" uses, protection against dilution may well be inefficient. For example, there may be parodic uses of famous trademark that might well be valued by consumers but would not be licensed, notwithstanding minimal effects of brand equity. More generally, broad protection against dilution could chill news reporting (e.g., stories exposing negative information about companies with famous trademarks), comparative advertising, and expressive creativity. Constitutional safeguards of freedom of expression as well as exceptions to trademark dilution liability seek to balance these competing considerations. As suggested above, it is not at all clear that dilution poses significant harm. The economic benefits are attenuated in most circumstances and traditional trademark law addresses the most significant concerns—where likely consumer confusion can be demonstrated. Therefore, dilution protection may be of questionable net value (Port, 1994; Lunney, 1999).

2.4.2. *Policy levers*

Since trademark dilution is an outgrowth of traditional trademark protection, the same validity, duration, and ownership and transfer rules discussed above apply to protection against dilution. The principal critical levers for this cause of action are the additional thresholds (fame and possibly inherent distinctiveness), the standard for determining infringement, and limitations on liability.

Additional threshold requirements. Unlike confusion-based trademark protection, federal dilution protection is available only to famous marks. Thus, the threshold for determining fame serves as a policy lever for determining the availability of dilution protection. As noted earlier, some commentators have expressed greater concern about the need for and adverse effects of dilution protection. They would limit dilution protection to the best known national brands. The earliest and most ardent academic proponent of the dilution cause of action suggested that protection should be confined to inherently distinctive marks—e.g., Kodak—and not be available to descriptive marks (including geographical designations and surnames) that have acquired distinctiveness and fame, such as United Airlines and McDonalds (Schechter, 1927). Such a constraint on the range of marks eligible for protection allows descriptive terms to remain more available for use in other markets.

Infringement standard. As we saw in the context of general confusion-based trademark liability, a mark owner need only establish a "likelihood of confusion" in order to prove infringement. Congress adopted a "likelihood of dilution" standard in the Trademark Dilution Revision Act of 2006.

Limitations on liability. The scope of dilution protection is quite broad, affording the owner of a famous mark broad discretion to enter (or preclude others from entering) any market under the famous mark. The right also protects the owner against tarnishment. Due to this vast potential scope of protection, Congress included several exceptions for comparative advertising, noncommercial uses (e.g., product reviews), and news re-

porting so as to balance competitive considerations and to address First Amendment concerns.

Remedies. Dilution protection envisions injunctive relief as the principal remedy, although damages and profits are available upon a showing that the defendant "wilfully intended to trade on the owner's reputation or to cause dilution of the famous mark."

2.5. Administration

Trademarks may be secured under common law without the need for registration or through federal or state registration regimes. In either case, use in commerce is typically required, although the federal protection can be secured for inherently distinctive marks with minimal (token) use. As noted above, it is now possible to reserve a trademark (and establish a priority date) by filing an intent-to-use application. The benefit of using this process is that the initial application is considered "constructive use," entitling the registrant to nationwide priority from the date of the application.

Examination of trademarks falls somewhere between the patent and copyright extremes. Trademarks must overcome greater technical hurdles than copyright law—classification of marks along the distinctiveness spectrum, prior art search, evaluation of evidence bearing on secondary meaning (in the case of non-inherently distinctive marks), statutory bars (immoral, deceptive, scandalous, disparaging, or functional marks are not registrable)—but are typically more straightforward to assess than patents.

Unlike patent law, the trademark system provides for a full inter partes opposition system. Marks eligible for registration are published in the PTO's Official Gazette, after which third parties have a 30 day period during which they may oppose registration. If no opposition is filed or the applicant prevails, then the mark is registered. Five years after registration, the trademark owner may apply for incontestability status, which insulates marks from being invalidated on the grounds that they are merely descriptive (i.e., lack secondary meaning) or lack priority. They may, however, continue to be challenged on several grounds, including abandonment, fraud, functionality, and generic status. Such rules reduce the risks of improvident grants of rights, provide for greater security upon registration, and reduce the costs of litigation.

2.6. Comparative analysis

As in the context of promoting innovation, trademark law represents but a part of a broad and complex array of legal regimes and public and private institutions that address the problem of ensuring the informational integrity of markets (Best, 1985). Therefore, as with patent, copyright and trade secret law, economic analysis of trademark law should take into consideration the full landscape of governance institutions and instruments.

As trademark law has evolved from the common law action for passing off the goods of one manufacturer for another's into relatively broad set of rights, other legal

rights and institutions have also developed to police advertising and selling practices. In addition to common law causes of action for fraud and deceit, the federal and state governments have enacted consumer-protection statutes that both create private rights of action and empower public agencies (the Federal Trade Commission as well as state analogs) to investigate deceptive practices and enforce consumer protections (Sovern, 1991). Federal and state governmental agencies proactively develop advertising and trade guidelines, field consumer complaints about deceptive practices, and initiate enforcement actions. Private enforcement of such statutes has become a large legal practice area (Sheldon and Carter, 1997). Non-governmental consumer protection organizations have developed to conduct independent product review (such as Consumer's Union), advocate for consumer protection regulation (e.g., Public Citizen), and support private enforcement of consumer protection laws (such as the National Consumer Law Center). Both public and private organizations have developed to provide independent certification of advertising claims and product quality (e.g., Underwriters Laboratories). Industry self-regulation has also emerged, most notably the Better Business Bureau, which processes consumer complaints and provides an alternative dispute resolution process for resolving false advertising complaints among businesses.

Each of these institutions draw upon different enforcers—consumers (and their attorneys), public entities, sellers in the market place, advertising industry organizations, independent certification laboratories, and consumer consortiums—with different sources of information and motivation to provide information and police disreputable sellers. The emergence of this broad range of enforcement resources suggests that trademark law should not be viewed as the sole or even principal means of protecting consumers. Rather, it should be seen as part of the composite mix. As a result, it need not be greatly concerned with the more egregious problems of consumer deception as other institutions focus more directly on such concerns.

Some trademark doctrines—such as the rule prohibiting assignment of trademarks in gross or the licensing of trademarks without supervision—may no longer be particularly important and may in fact raise transaction costs needlessly. These doctrines can also be questioned directly on economic grounds. It is not at all clear why a company which acquires a valuable trademark "in gross" would be any more inclined to engage in opportunism than the original owner or an assignee which took over the underlying business. Similarly, trademark licensors strong incentives to develop efficient supervisory structures even without a rule prohibiting "naked" licenses. In any case, consumer protection laws may provide a better institutional means of confronting the problems that these trademark rules purport to address.

Acknowledgements

We would like to thank Mark Lemley and the editors for comments.

References

Abramowicz, M. (2003). "Perfecting patent prizes". Vanderbilt Law Review 56, 115–236.

Akerlof, G.A. (1970). "The market for 'Lemons': quality uncertainty and the market mechanism". Quarterly Journal of Economics 84, 488–500.

Akerlof, G.A. et al. (2003), as Amici Curiae in Support of Petitioners in Eldred v. Ashcroft, 537 U.S. 186.

Allison, J.R., Lemley, M.A. (1998). "Empirical evidence on the validity of litigated patents". AIPLA Quarterly Journal 26, 185–275.

Allison, J.R., Lemley, M.A. (2000). "How federal circuit judges vote in patent validity cases". Florida State University Law Review 27, 745–766.

Allison, J.R., Lemley, M.A. (2002). "The growing complexity of the United States patent system". Boston University Law Review 82, 77–145.

Allison, J.R., Lemley, M.A., Moore, K., Trunkey, R.D. (2004). "Valuable patents". Georgetown Law Journal 92, 435–479.

Allison, J.R., Tiller, E.H. (2003a). "Internet business method patents". In: Cohen, W., Merrill, S. (Eds.), Patents in the Knowledge-Based Economy. National Academy Press, Washington, D.C., pp. 259–285.

Allison, J.R., Tiller, E.H. (2003b). "The business method patent myth". Berkeley Technology Law Journal 18, 987–1084.

Alston, J. (2002). "Spillovers". Australian Journal of Agricultural and Resource Economics 46, 315–346.

Anton, J., Yao, D. (2007). "Finding 'lost' profits: an equilibrium analysis of patent infringement damages". Journal of Law, Economics, and Organization 23 (1), 186–207.

Aoki, R., Tauman, Y. (2001). "Patent licensing with spillovers". Economic Letters 73, 125–130.

Armond, M. (2003). "Introducing the defense of independent invention to motions for preliminary injunctions in patent infringement lawsuits". California Law Review 91, 117–162.

Arora, A., Fosfuri, A., Gambardella, A. (2001). Markets for Technology: The Economics of Innovation and Corporate Strategy. MIT Press, Cambridge.

Arrow, K. (1962). "Economic welfare and the allocation of resources for invention". In: Nelson, R.R. (Ed.), Universities-National Bureau of Economic Research Conference Series. The Rate and Direction of Economic Activities: Economic and Social Factors. Princeton University Press, New York.

Ayres, I., Klemperer, P. (1999). "Limiting patentees' market power with reducing innovation incentives: the perverse benefits of uncertainty and noninjunctive remedies". Michigan Law Review 97, 985–1033.

Bakos, Y., Brynjolfsson, E., Lichtman, D. (1999). "Shared information goods". Journal of Law and Economics 42, 117–155.

Bar-Gill, O., Parchomovsky, G. (2003). "The value of giving away secrets". Virginia Law Review 89, 1857–1895.

Barton, J.H. (1997). "Patents and antitrust: a rethinking in light of patent breadth and sequential innovation". Antitrust Law Journal 65, 446–449.

Barton, J.H. (2000). "Intellectual property rights: reforming the patent system". Science 287, 1933–1934.

Barton, J.H. (2002). "Antitrust treatment of oligopolies with mutually blocking patent portfolios". Antitrust Law Journal 69, 851–882.

Barton, J.H. (2003). "Nonobviousness". IDEA The Journal of Law and Technology 43, 475–508.

Barzel, Y. (1968). "Optimal timing of innovations". Review of Economics and Statistics 50, 348–355.

Benkler, Y. (2002). "Coase's Penguin, or Linux and the nature of the firm". Yale Law Journal 112, 369–446.

Benkler, Y. (2004). "Sharing nicely: on shareable goods and the emergence of sharing as a modality of economic production". Yale Law Journal 114, 273–358.

Besen, S.M., Kirby, S.N. (1989a). "Private copying, appropriability and optimal copyright royalties". Journal of Law and Economics 32, 255–280.

Besen, S.M., Kirby, S.N. (1989b). "Compensating creators of intellectual property". Rand Corporation, Santa Monica, CA, No. R-3751-MF.

Besen, S.M., Kirby, S.N., Salop, S.C. (1992). "An economic analysis of copyright collectives". Virginia Law Review 78, 383–411.

Besen, J., Maskin, E. (2002). "Sequential innovation, patents and imitation". MIT, Department of Economics, Cambridge, MA, Working Paper 00-01.

Best, A. (1985). "Controlling false advertising: a comparative study of public regulation, industry self-policing, and private litigation". Georgia Law Review 20, 1–72.

Bhattacharya, S., d'Aspremont, C., Gerard-Varet, J.-L. (2000). "Bargaining and sharing innovative knowledge". Review of Economic Studies 67, 255–271.

Bhattacharya, S., Glazer, J., Sappington, D.E.M. (1990). "Sharing productive knowledge in internally financed R&D contests". Journal of Industrial Economics 39, 187–208.

Bhattacharya, S., Glazer, J., Sappington, D.E.M. (1992). "Licensing and the sharing of knowledge in research joint ventures". Journal of Economic Theory 56, 43–69.

Blair, R.D., Cotter, T.F. (1998). "An economic analysis of damages rules in intellectual property law". William and Mary Law Review 39, 1585–1694.

Blair, R.D., Cotter, T.F. (2001). "Rethinking patent damages". Texas Intellectual Property Law Journal 10, 1–93.

Blair, R.D., Cotter, T.F. (2002). "Strict liability and its alternatives in patent law". Berkeley Technology Law Journal 17, 799–845.

Bone, R.G. (2004). "Enforcement costs and trademark puzzles". Virginia Law Review 90, 2099–2185.

Brown, R.S. Jr. (1948). "Advertising and the public interest: legal protection of trade symbols". Yale Law Journal 57, 1165.

Britt, B. (1990). "International marketing: Disney's global goals". Marketing, May 17, 22–26.

Brocas, I. (2004). "Optimal regulation of cooperative R&D under incomplete information". Journal of Industrial Organization 52, 81–120.

Burge, D. (1984). Patent and Trademark Tactics and Practice. John Wiley, New York.

Burk, D.L., Lemley, M.A. (2002). "Is patent law technology-specific?" Berkeley Technology Law Journal 17, 1155–1206.

Burk, D.L., Lemley, M.A. (2003). "Policy levers in patent law". Virginia Law Review 89, 1575–1696.

Burke, T.P. (1994). "Software patent protection: debugging the current system". Notre Dame Law Review 69, 1115–1166.

Cai, M. (2004). "Madey v. Duke university: shattering the myth of universities' experimental use defense". Berkeley Technology Law Journal 19, 175–192.

Calabresi, G., Melamed, A.D. (1972). "Property rules, liability rules, and inalienability: one view of the cathedral". Harvard Law Review 85, 1089–1126.

Carter, S.L. (1990). "The trouble with trademark". Yale Law Journal 99, 759–800.

Chamberlin, E. (1933). "The theory of monopolistic competition". Ph.D. Thesis, Harvard, pp. 56, 204.

Chang, H.F. (1995). "Patent scope, antitrust policy, and cumulative innovation". Rand Journal of Economics 26, 34–57.

Che, Y.-K., Gale, I. (2003). "Optimal design of research contests". American Economic Review 93, 646–671.

Chen, Y., P'ng, I. (2003). "Information goods, pricing, and copyright enforcement: welfare analysis". Information Systems Research 14, 107–123.

Choi, J.P. (1991). "Dynamic R&D competition under 'hazard rate' uncertainty". Journal of Industrial Economics 22, 596–610.

Chu, C.A. (2001). "Empirical analysis of the federal circuit's claim construction trends". Berkeley Technology Law Journal 16, 1075–1164.

Clarke, R. (2003). "U.S. continuity law and its impact on the comparative patenting rates of the U.S., Japan, and European Patent Offices". Journal of the Patent and Trademark Office Society 8, 335–349.

Cockburn, I., Henderson, R. (2003). "Survey results from the 2003 intellectual property owners association survey on strategic management of intellectual property". Boston University and MIT, Departments of Economics, mimeograph.

Cohen, J.E., Lemley, M.A. (2001). "Patent scope and innovation in the software industry". California Law Review 89, 1–57.

Cohen, W.M., Nelson, R.R., Walsh, J.P. (2000). "Protecting their intellectual assets: appropriability conditions and why U.S. manufacturing firms patent (or not)". National Bureau of Economic Research, Working Paper No. 7552, Cambridge, Massachusetts. (At: http://papers.nber.org/papers/w7552.pdf.)

Comanor, W.S., Wilson, T.A. (1974). Advertising and Market Power. Harvard University Press, Cambridge, MA.

Commission on Revision of the Federal Court Appellate System (1975). Structure and Internal Procedures: Recommendations for Change. Reprinted in Federal Register 67, 54878–54905.

Conner, K.R., Rumelt, R.P. (1991). "Software piracy: an analysis of protecting strategies". Management Science 37, 125–139.

Cooter, R. (1982). "Economic analysis of punitive damages". Southern California Law Review 56, 79–100.

Corcoran, R. (1999). "Quality review and control in the PTO: the historical evolution". Journal of the Patent and Trademark Office Society 81, 7–11.

Cornelli, F., Schankerman, M. (1999). "Patent renewals and R&D incentives". The Rand Journal of Economics 30, 197–213.

Cotter, T.F. (2004). "An economic analysis of enhanced damages and attorneys' fees for willful infringement". Federal Circuit Bar Journal 14, 291–331.

Coverdale, J.F. (1984). "Comment, trademarks and generic words: an effect-on-competition test". University of Chicago Law Review 51, 868–891.

Darby, M.R., Karni, E. (1973). "Free competition and the optimal amount of fraud". Journal of Law and Economics 16, 67–88.

Dasgupta, P., Stiglitz, J. (1980a). "Industrial structure and the nature of innovative activity". Economic Journal 90, 266–293.

Dasgupta, P., Stiglitz, J. (1980b). "Uncertainty, market structure and the speed of research". Bell Journal of Economics 11, 1–28.

d'Aspremont, C., Jacquemin, A. (1988). "Cooperative and noncooperative R&D in duopoly with spillovers". American Economic Review 5, 1133–1137.

David, P. (1985). New Technology, Diffusion, Public Policy, and Industrial Competitiveness. Stanford University, Center for Economic Policy Research.

David, P. (2003). "The economic logic of 'open science' and the balance between private property rights and the public domain in scientific data and information: a primer". In: National Research Council, The Role of the Public Domain in Scientific and Technical Data and Information. National Academies Press, Washington, D.C.

de Laat, E.A. (1996). "Patents or prizes: monopolistic R&D and asymmetric information". International Journal of Industrial Organization 15, 369–390.

Denicola, R.C. (1999). "Freedom to copy". Yale Law Journal 108, 1661–1686.

Denicolò, V. (1996). "Patent races and optimal patent breadth and length". Journal of Industrial Economics 44, 249–265.

Denicolò, V. (1997). "Patent policy with a finite sequence of patent races". Department of Economics, University of Bologna, mimeograph.

Denicolò, V. (2002a). "Two-stage patent races and patent policy". Journal of Industrial Economics 31, 488–501.

Denicolò, V. (2002b). "Sequential innovation and the patent-antitrust conflict". Oxford Economic Papers 54, 649–668.

Denison, E. (1985). Trends in American Economic Growth. Brookings Institute Press, Washington, D.C.

Desmond, R. (1993). "Nothing seems 'obvious' to the court of appeals for the federal circuit: the federal circuit, unchecked by the supreme court, transforms the standard of obviousness under patent law". Loyola of Los Angeles Law Review 26, 455–490.

Dogan, S.L., Lemley, M.A. (2004a). "Trademarks and consumer search costs on the Internet". Houston Law Review 41, 777–838.

Dogan, S.L., Lemley, M.A. (2004b). "The merchandising right: fragile theory or fait accompli?" Stanford Public Law, Working Paper No. 105.

Dratler, J. Jr. (2003). "Does Lord Darcy yet live? The case against software and business-method patents". Santa Clara Law Review 43, 823–899.
Dreyfuss, R.C. (1989). "The federal circuit: a case study in specialized courts". New York University Law Review 64, 1–77.
Dreyfuss, R.C. (1990). "Expressive genericity: trademarks as language in the pesi generation". Notre Dame Law Review 65, 397–424.
Dreyfuss, R.C. (2000). "Are business method patents bad for business?" Santa Clara Computer and High Technology Law Journal 16, 263–280.
Dreyfuss, R.C. (2003). "Varying the course". In: Scott Kieff, F. (Ed.), Perspectives on the Properties of the Human Genome Project. Elsevier, New York.
Duffy, J. (moderator) (1998). "Early patent publication: a boon or bane? A discussion on the legal and economic effects of publishing patent applications after eighteen months of filing". Cardozo Arts & Entertainment Law Journal 16, 601–633.
Dunner, D., Jakes, J., Karceski, J. (1995). "A statistical look at the federal circuit's patent decisions: 1982–1994". Federal Circuit Bar Journal 5, 151–180.
Dyson, J. (2001). A History of Great Inventions. Carroll & Graf, New York.
Economides, N. (1998). "Trademarks". In: Newman, P. (Ed.), The New Palgrave Dictionary of Economics and the Law. MacMillan, London, pp. 601–603.
Eisenberg, R.S. (1989). "Patents and the progress of science: exclusive rights and experimental use". University of Chicago Law Review 56, 1017–1086.
Eisenberg, R.S. (1996). "Public research and private development: patents and technology transfer in government-sponsored research". Virginia Law Review 82, 1663–1727.
Eisenberg, R.S. (2000). "The promise and perils of strategic prior art creation through publication: a response to professor Parchomovsky". Michigan Law Review 98, 2358–2370.
Epstein, R.A. (2003). "Steady the course: property rights in genetic material". In: Scott Kieff, F. (Ed.), Perspectives on Properties of the Human Genome Project. Elsevier, New York, pp. 153–156.
European Patent Office (2002a). "Trilateral Statistical Reports 2001". (At: http://www.european-patent-office.org/tws/tsr_2001/index.php.)
European Patent Office (2002b). "Facts and figures 2002". (At: www.european-patent-office.org/epo/facts_figures/facts2001/pdf/facts_figures_01.pdf.)
Farrell, J., Katz, M. (1998). "The effects of antitrust and intellectual property law on compatibility and innovation". Antitrust Bulletin 48, 609–650.
Farrell, J., Klemperer, P. (2004). "Coordination and lock-in: competition with switching costs and network effects". (At: www.paulklemperer.org.) Nuffield College, Oxford. In: Armstrong, M., Porter, R.H. (Eds.), Handbook of Industrial Organization, vol. 3. North Holland, Amsterdam. In preparation.
Feit, I.N. (1989). "Biotechnology research and the experimental use exception to patent infringement". Journal of the Patent and Trademark Office Society 71, 819–840.
Fisher, W. (2004). Promises to Keep: Technology, Law, and the Future of Entertainment. Stanford University Press, Palo Alto, CA.
Folsom, R.H., Teply, L.L. (1980). "Trademarked generic words". Yale Law Journal 89, 1323–1359.
Folsom, R.H., Teply, L.L. (1988a). "Surveying 'genericness' in trademark litigation". The Trademark Reporter 78, 1–31.
Folsom, R.H., Teply, L.L. (1988b). "A reply to Swann and Palladino's critique of Folsom and Teply's model survey". The Trademark Reporter 78, 197–208.
Foray, D., Hilaire-Perez, L. (2000). "The economics of open technology: collective organization and individual claims in the 'Fabrique Lyonnaise' during the old regime". (At: http://emlab.berkeley.edu/users/bhhall/ipconf01.html (conference paper).)
Friedman, D.D., Landes, W.M., Posner, R.A. (1991). "Some economics of trade secret law". Journal of Economic Perspectives 5, 61–72.
Galler, B.A. (1990). "A proposal for a software patent institute". June 26.
Gallini, N.T. (1984). "Deterrence by market sharing: a strategic incentive for licensing". American Economic Review 74, 931–941.

Gallini, N.T. (1992). "Patent policy and costly imitation". Journal of Industrial Economics 44, 52–63.
Gallini, N.T., Scotchmer, S. (2002). "Intellectual property: when is it the best incentive mechanism?" Innovation Policy and the Economy 2, 51–78.
Gallini, N.T., Winter, R. (1985). "Licensing in the theory of innovation". Journal of Industrial Economics 16, 237–252.
Gandal, N., Scotchmer, S. (1993). "Coordinating research through research joint ventures". Journal of Public Economics 51, 173–193.
Gervais, D. (2005). "The price of social norms: towards a licensing regime for file-sharing". Journal of Intellectual Property Law 12, 39–73.
Ghosh, S., Kesan, J. (2004). "What do patents purchase? In search of optimal ignorance in the patent office". Houston Law Review 40, 1219–1264.
Gilbert, R. (2002). "Patent pools: 100 years of law and economic solitude". (At: http://www.innovationlaw.org/pages/patent_pools.doc.)
Gilbert, R., Newbery, D. (1982). "Preemptive patenting and the persistence of monopoly". American Economic Review 72, 514–526.
Gilbert, R., Shapiro, C. (1990). "Optimal patent length and breadth". Journal of Industrial Economics 21, 106–112.
Gilbert, R., Shapiro, C. (1997). "Antitrust issues in the licensing of intellectual property: the nine no-no's meet the nineties". In: Winston, C. (Ed.), Brookings Papers on Microeconomic Activity. The Brookings Institution, Washington, D.C.
Gilbert, R., Sunshine, G.C. (1995a). "Incorporating dynamic efficiency concerns in merger analysis: the use of innovation markets". Antitrust Law Journal 63, 569–602.
Gilbert, R., Sunshine, G.C. (1995b). "The use of innovation markets: a reply to Hay, Rapp and Hoerner". Antitrust Law Journal 64, 75–82.
Goldstein, P. (1986). "Infringement of copyright in computer programs". University of Pittsburgh Law Review 47, 1119–1130.
Gordon, W. (1982). "Fair use as market failure: a structural and economic analysis of the betamax case and its predecessors". Columbia Law Review 82, 1600–1655.
Grady, M.F., Alexander, J.I. (1992). "Patent law and rent dissipation". Virginia Law Review 78, 305–350.
Green, J., Scotchmer, S. (1995). "On the division of profit in sequential innovation". Journal of Industrial Economics 26, 20–33.
Grossman, G., Lai, E. (2001). "International protection of intellectual property". Princeton University, Princeton, NJ, mimeograph.
Grushcow, J. (2004). "Measuring secrecy: a cost of the patent system revealed". Journal of Legal Studies 33, 59–84.
Hall, B.H. (2003). "Business method patents, innovation and policy". Working Paper 9717, National Bureau of Economic Research, Cambridge, MA.
Hall, B.H., Ziedonis, R.H. (2001). "The patent paradox revisited: an empirical study of patenting in the U.S. semiconductor industry 1979–1995". Journal of Industrial Economics 32, 101–128.
Heller, M.A., Eisenberg, R.S. (1998). "Can patents deter innovation? The anticommons in biomedical research". Science 280, 698–701.
Hicks, D. (1995). "Published papers, tacit competencies, and corporate management of the public/private character of knowledge". Industrial and Corporate Change 4, 401–424.
Hicks, J. (1932). The Theory of Wages. Macmillan, London.
Higgins, R.S., Rubin, P.H. (1986). "Counterfeit goods". Journal of Law and Economics 29, 211–230.
Hirshleifer, J. (1973). "Where are we in the theory of information?" American Economic Review. Papers and Proceedings 63, 31–39.
Hoerner, R. (1995). "Innovation markets: new wine in old bottles?" Antitrust Law Journal 64, 49–73.
Horstmann, I., MacDonald, G.M., Slivinski, A. (1985). "Patents as information transfer mechanisms: to patent or (maybe) not to patent". Journal of Political Economy 93, 837–858.
Hovenkamp, H., Janis, M., Lemley, M.A. (2004). IP and Antitrust: An Analysis of Antitrust Principles Applied to Intellectual Property Law. Aspen Publishers, New York.

Hunt, R. (1999). "Nonobviousness and the incentive to innovate: an economic analysis of intellectual property reform". Working Paper 99-3, The Federal Reserve Bank of Philadelphia, Philadelphia.
Hunt, R. (2004). "Patentability, industry structure, and innovation". Journal of Industrial Economics 52, 401–425.
Jaffe, A.B., Lerner, J. (2004). Innovation and Its Discontents: How Our Broken Patent System is Endangering Innovation and Progress, and What to Do About It. Princeton University Press, New Jersey.
Johnson, W.R. (1985). "The economics of copying". Journal of Political Economy 93, 158–174.
Kamien, M.I., Muller, E., Zang, I. (1992). "Research joint ventures and R&D cartels". American Economic Review 85, 1293–1306.
Kaplow, L. (1984). "The patent-antitrust intersection: a reappraisal". Harvard Law Review 97, 1813–1892.
Kaplow, L., Shavell, S. (1996). "Property rules versus liability rules: an economic analysis". Harvard Law Review 109, 715–789.
Karjala, D.S., Menell, P.S. (1995). "Applying fundamental copyright principles to Lotus Development Corp. v. Borland International, Inc." High Technology Law Journal 10, 177–192.
Katz, M.L., Shapiro, C. (1986). "Technology adoption in the presence of network externalities". Journal of Political Economy 94, 822–881.
Katz, M.L., Shapiro, C. (1987). "R&D; rivalry with licensing or imitation". American Economic Review 77, 402–420.
Kesan, J.P. (2002). "Carrots and sticks to create a better patent system". Berkeley Technology Law Journal 17, 763–797.
Kieff, F.S. (2001a). "Facilitating scientific research: intellectual property rights and the norms of science—a response to Rai and Eisenberg". Northwestern University Law Review 95, 691–705.
Kieff, F.S. (2001b). "Property rights and property rules for commercializing inventions". Minnesota Law Review 85, 697–754.
Kitch, E.W. (1977). "The nature and function of the patent system". The Journal of Law and Economics 20, 265–290.
Kitch, E.W. (1980). "The law and economics of rights in valuable information". Journal of Legal Studies 9, 683–723.
Klein, B., Leffler, K.B. (1981). "The role of market forces in assuring contractual performance". Journal of Political Economy 89, 615–641.
Klemperer, P. (1990). "How broad should the scope of patent protection be?" Journal of Industrial Economics 21, 113–130.
Kortum, S., Lerner, J. (1998). "Stronger protection or technological revolution: what is behind the recent surge in patenting?" Carnegie-Rochester Conference Series on Public Policy 48, 247–304.
Kozinski, A. (1993). "Trademarks unplugged". New York University Law Review 68, 960–978.
Kratzke, W.P. (1991). "Normative economic analysis of trademark law". Memphis State University Law Review 21, 199–286.
Kremer, M. (1998). "Patent buyouts: a mechanism for encouraging innovation". Quarterly Journal of Economics 113, 1137–1168.
La Manna, M., MacLewod, R., de Meza, D. (1989). "The case for permissive patents". European Economic Review 33, 1427–1443.
Landes, W.M., Posner, R.A. (1987). "Trademark law: an economic perspective". Journal of Law and Economics 30, 265–309.
Landes, W.M., Posner, R.A. (2003). The Economic Structure of Intellectual Property Law. Harvard University Press, Cambridge.
Landes, W.M., Posner, R.A. (2004). "Intellectual property, an empirical analysis of the patent court". University of Chicago Law Review 71, 111–128.
Lanjouw, J.O. (1998). "Patent protection in the shadow of infringement: simulation estimations of patent value". The Review of Economic Studies 65, 761–810.
Lanjouw, J.O., Lerner, J. (1997). "The enforcement of intellectual property rights: a survey of the empirical literature". National Bureau of Economic Research, Working Paper No. 6296.

Lanjouw, J.O., Lerner, J. (2001). "Tilting the table? The predatory use of preliminary injunctions". Journal of Law and Economics 44, 573–603.
Lanjouw, J.O., Schankerman, M. (2001). "Characteristics of patent litigation: a window on competition". The Rand Journal of Economics 32, 129–151.
Lanjouw, J.O., Schankerman, M. (2004). "Protecting intellectual property rights: are small firms handicapped?" Journal of Law and Economics 47, 45–74.
Lee, T., Wilde, L. (1980). "Market structure and innovation: a reformulation". Quarterly Journal of Economics 94, 429–436.
Leibenstein, H. (1950). "Bandwagon, snob, and veblen effects in the theory of consumers' demand". Quarterly Journal of Economics 64, 183–207.
Leibovitz, J.S. (2002). "Inventing a nonexclusive patent system". Yale Law Journal 111, 2251–2287.
Lemley, M.A. (1997). "The economics of improvement in intellectual property law". Texas Law Review 75, 989–1083.
Lemley, M.A. (1999). "The modern Lanham Act and the death of common sense". Yale Law Journal 108, 1687–1715.
Lemley, M.A. (2001). "Rational ignorance at the patent office". Northwestern University Law Review 95, 1495–1532.
Lemley, M.A. (2002). "Intellectual property rights and standard-setting organizations". California Law Review 90, 1889–1979.
Lemley, M.A. (2004). "Ex ante versus ex post justifications for intellectual property law". University of Chicago Law Review 71, 129–149.
Lemley, M.A., Chien, C.V. (2003). "Are the U.S. patent priority rules really necessary?" Hastings Law Journal 54, 1299–1333.
Lemley, M.A., Moore, K.A. (2004). "Ending abuse of patent continuations". Boston University Law Review 84, 63–126.
Lemley, M.A., O'Brien, D.W. (1997). "Encouraging software reuse". Stanford Law Review 49, 255–304.
Lemley, M.A., Reese, R. (2004). "Reducing digital copyright infringement without restricting innovation". Stanford Law Review 56, 1345–1434.
Lemley, M.A., Tangri, R.K. (2003). "Ending patent law's willfulness game". Berkeley Technology Law Journal 18, 1085–1125.
Lerner, J. (1994). "The importance of patent scope: an empirical analysis". Rand Journal of Economics 25, 319–333.
Lerner, J. (1995). "Patenting in the shadow of competitors". Journal of Law and Economics 38, 563–595.
Lerner, J., Tirole, J. (2000). "The simple economics of open source". Working Paper 7600. National Bureau of Economic Research, Cambridge.
Lerner, J., Tirole, J. (2004). "Efficient patent pools". American Economic Review 94, 691–711.
Lesser, W., Lybbert, T. (2004). "Do patents come too easy?" IDEA: The Journal of Law and Technology 44, 381–407.
Lessig, L. (2004). Free Culture: How Big Media Uses Technology and the Law to Lock Down Culture and Control Creativity. Penguin Press, New York.
Levin, J., Levin, R. (2003). "Benefits and costs of an opposition process". In: Cohen, W., Merrill, S. (Eds.), Patents in the Knowledge-Based Economy. National Academy Press, Washington, D.C.
Levin, R.C., Klevorick, A.K., Nelson, R.R., Winter, S.G. (1987). "Appropriating the returns from industrial R&D". Brookings Papers on Economic Activity: Microeconomics 3, 783–820.
Lichtman, D. (1997). "The economics of innovation: protecting unpatentable goods". Minnesota Law Review 81, 693–734.
Lichtman, D., Baker, S., Kraus, K. (2000). "Strategic disclosure in the patent system". Vanderbilt Law Review 53, 2175–2217.
Lichtman, D., Landes, W. (2003). "Indirect liability for copyright infringement: an economic perspective". Harvard Journal of Law & Technology 16, 395–410.
Liebowitz, S.J. (1982). "Durability, market structure, and new-used goods models". American Economic Review 72, 816–824.

Liebowitz, S.J. (1985). "Copying and indirect appropriability: photocopying of journals". Journal of Political Economy 93, 945–957.
Liebowitz, S.J. (2004). "File sharing: creative destruction or just plain destruction?". (At: http://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=59984.)
Liebowitz, S.J., Margolis, S.E. (1982). "Journals as shared goods: comment". American Economic Review 72, 597–602.
Long, C. (2002). "Patent signals". University of Chicago Law Review 69, 625–679.
Loury, G.C. (1979). "Market structure and innovation". Quarterly Journal of Economics 93, 395–410.
Lunney, G.S. Jr. (1999). "Trademark monopolies". Emory Law Journal 48, 367–487.
Lunney, G.S. Jr. (2000–2001). "E-obviousness". Michigan Telecommunication and Technology Law Review 7, 363–422.
Lunney, G.S. Jr. (2004). "Patent law, the federal circuit and the supreme court: a quiet revolution". Supreme Court Economic Review 11, 1–80.
Macedo, C. (1990). "First-to-file: is American adoption of the international standard in patent law worth the price?" AIPLA Quarterly Journal 18, 193–234.
Machlup, F., Penrose, E. (1950). "The patent controversy in the nineteenth century". Journal of Economic History 10, 1–29.
Mak, J., Walton, G.M. (1972). "Steamboats and the great productivity surge in river transportation". Journal of Economic History 32, 619–640.
Mansfield, E. (1986). "Patents and innovation: an empirical study". Management Science 32, 173–181.
Maskus, K. (2000a). Intellectual Property Rights in the Global Economy. Institute for International Economics, Washington, D.C.
Maskus, K. (2000b). "Lessons from studying the international economics of intellectual property rights". Vanderbilt Law Review 53, 2219–2240.
Matutes, C., Regibeau, P. (1992). "Compatibility and bundling of complementary goods in a duopoly". Journal of Industrial Economics 40, 37–54.
Maurer, S.M. (2003). "New institutions for doing science: from databases to open source biology". (At: http://www.merit.unimaas.nl/epip/papers/maurer_paper.pdf.)
Maurer, S.M., Scotchmer, S. (2002). "The independent invention defense in intellectual property". Economica 69, 535–547.
Maurer, S.M., Scotchmer, S. (2004a). "Profit neutrality in licensing: the boundary between antitrust law and patent law". Working Paper 10546. National Bureau of Economic Research, Cambridge.
Maurer, S.M., Scotchmer, S. (2004b). "Procuring knowledge". In: Libecap, G. (Ed.), Advances in the Study of Entrepreneurship, Innovation and Growth, vol. 15. JAI Press, Elsevier, Amsterdam, pp. 1–31.
McCalman, P. (2001). "Reaping what you sow: an empirical analysis of international patent harmonization". Journal of International Economics 55, 161–185.
McCarthy, J.T. (2004). McCarthy on Trademarks and Unfair Competition. West, St. Paul, MN.
McClure, D.M. (1979). "Trademarks and unfair competition: a critical history of legal thought". The Trademark Reporter 69, 305–356.
McClure, D.M. (1996). "Trademarks and competition: the recent history". Law and Contemporary Problems 59, 13–43.
McGowan, D. (2001). "The legal implications of open source software". University of Illinois Law Review 2001, 241–304.
McDade, A.S. (1998). "Trading in trademarks—why the anti-assignment in gross doctrine should be abolished when trademarks are used as collateral". Texas Law Review 77, 465–492.
Menell, P.S. (1987). "Tailoring legal protection for computer software". Stanford Law Review 39, 1329–1372.
Menell, P.S. (1989). "An analysis of the scope of copyright protection for application programs". Stanford Law Review 41, 1045–1104.
Menell, P.S. (1994). "The challenges of reforming intellectual property protection for computer software". Columbia Law Review 94, 2644–2654.
Menell, P.S. (1998a). "Intellectual property: general theories". In: Bouckaert, B., De Geest, G. (Eds.), Encyclopedia of Law and Economics, vol. II. Edward Elgar, Cheltenham.

Menell, P.S. (1998b). "An epitaph for traditional copyright protection of network features of computer software". Antitrust Bulletin 43, 651–713.
Menell, P.S. (2000). "Economic implications of state sovereign immunity from infringement of federal intellectual property rights". Loyola of Los Angeles Law Review 33, 1399–1466.
Menell, P.S. (2003). "Envisioning copyright law's digital future". New York Law Review 46, 63–193.
Menell, P.S. (2005). "Indirect copyright liability: a re-examination of Sony's staple article of commerce doctrine". UC Berkeley Public Law, Research Paper No. 682051.
Merges, R.P. (1988). "Commercial success and patent standards: economic perspectives on innovation". California Law Review 76, 803–876.
Merges, R.P. (1992). "Uncertainty and the standard of patentability". High Technology Law Journal 7, 1–70.
Merges, R.P. (1994). "Intellectual property rights and bargaining breakdown: the case of blocking patents". Tennessee Law Review 62, 75–76.
Merges, R.P. (1996). "Contracting into liability rules: intellectual property rights and collective rights organizations". California Law Review 84, 1293–1393.
Merges, R.P. (1999a). "As many as six impossible patents before breakfast: property rights for business concepts and patent system reform". Berkeley Technology Law Journal 14, 577–615.
Merges, R.P. (1999b). "Institutions for intellectual property transactions: the case of patent pools". In: Dreyfuss, R. (Ed.), Intellectual Products: Novel Claims to Protection and their Boundaries. Oxford University Press, Oxford.
Merges, R.P. (2004a). "Compulsory licensing vs. the three 'golden oldies' property, contract rights and markets". CATO Institute Policy Analysis No 508. (At: http://www.cato.org/pubs/pas/pa508.pdf.)
Merges, R.P. (2004b). "A new dynamism in the public domain". University of Chicago Law Review 71, 183–203.
Merges, R.P., Duffy, J. (2002). Patent Law and Policy. Matthew Bender, San Francisco.
Merges, R.P., Menell, P.S., Lemley, M.A. (2003). Intellectual Property in the New Technological Age. Aspen, New York.
Merges, R.P., Nelson, R. (1990). "On the complex economics of patent scope". Columbia Law Review 90, 839–916.
Meurer, M.J. (2002). "Business methods and patent floods". Washington University Journal of Law and Policy 8, 309–339.
Meurer, M.J. (2003). "Controlling opportunistic and anti-competitive intellectual property litigation". Boston College Law Review 44, 509–544.
Mikhail, P. (2000). "Hopkins v. CellPro: an illustration that patenting and exclusive licensing of fundamental science is not always in the public interest". Harvard Journal of Law & Technology 13, 375–394.
Milgrom, P., Roberts, J. (1986). "Price and advertising signals of product quality". Journal of Political Economy 94, 796–821.
Miller, J.S. (2004). "Building a better bounty: litigation-stage rewards for defeating patents". Berkeley Technology Law Journal 19, 667–739.
Minehart, D., Scotchmer, S. (1999). "Ex post regret and the decentralized sharing of information". Games and Economic Behavior 17, 114–131.
Mokyr, J. (1990). The Lever of Riches: Technological Creativity and Economic Progress. Oxford University Press, New York.
Moore, K.A. (2000). "Judges, juries, and patent cases—an empirical peek inside the black box". Michigan Law Review 99, 365–409.
Moore, K.A. (2001). "Are district court judges equipped to resolve patent cases?" Harvard Journal of Law & Technology 15, 1–39.
Mossinghoff, G. (2002). "The first-to-invent system has provided no advantage to small entities". Journal of the Patent and Trademark Office Society 84, 425–431.
Mowery, D., Nelson, R., Sampat, B.N., Ziedonis, A.A. (2001). "The growth of patenting and licensing by U.S. universities: an assessment of the effects of the Bayh-Dole act of 1980". Research Policy 30, 99–119.
Mowery, D., Rosenberg, N. (1998). Paths of Innovation: Technological Change in 20th-Century America. Cambridge Univ. Press, Cambridge, MA.

Mueller, J. (2001). "No 'dilettante affair': rethinking the experimental use exception to patent infringement for biomedical research tools". Washington Law Review 76, 1–66.
Nagle, T.T. (1981). "Do advertising-profitability studies really show that advertising creates a barrier to entry?" Journal of Law and Economics 24, 333–344.
National Research Council (2004). A Patent System for the 21st Century. Merrill, S.A., Levin, R.C., Myers, M.B. (Eds.). National Academies Press, Washington, D.C.
National Science Board (2002). Science and Engineering Indicators. National Science Foundation, Arlington, VA (NSB-02-1).
Nelson, P. (1970). "Information and consumer behavior". Journal of Political Economy 78, 311–329.
Nelson, P. (1974). "Advertising as information". Journal of Political Economy 82, 729–754.
Nelson, P. (1975). "The economic consequences of advertising". Journal of Business 48, 213–241.
Nelson, R.R. (1959). "The simple economics of basic scientific research". Journal of Political Economy 67, 297–306.
Nelson, R.R. (2003). "The market economy and the scientific commons". National Bureau of Economic Research, Summer Institute Working Paper. (At: http://www.nber.org/~confer/2003/si2003/papers/pripe/nelson.pdf.)
Nelson, R.R., Winter, S. (1982). An Evolutionary Theory of Economic Change. Harvard University Press, Cambridge.
Netanel, N. (2003). "Impose a noncommercial use levy to allow free peer-to-peer file sharing". Harvard Journal of Law & Technology 17, 1–84.
Newell, R.G., Jaffe, A.B., Stavins, R.N. (1999). "The induced innovation hypothesis and energy-saving technological change". The Quarterly Journal of Economics 114, 941–975.
Nordhaus, W. (1969). Invention, Growth and Welfare. MIT Press, Cambridge.
Novos, I., Waldman, M. (1984). "The effects of increasing copyright protection: an analytic approach". Journal of Political Economy 92, 236–246.
Oberholzer, F., Strumpf, K. (2004). "The effects of filesharing on record sales: an empirical analysis". (At: http://www.unc.edu/~cigar/papers/FileSharing_March2004.pdf.)
Oddi, A.S. (1988). "Assessing 'genericness': another view". The Trademark Reporter 78, 560–578.
Oddi, A.S. (1989). "Beyond obviousness: invention protection in the twenty-first century". American University Law Review 38, 1097–1148.
O'Donoghue, T. (1998). "A patentability requirement for sequential innovation". Journal of Industrial Economics 29, 654–679.
O'Donoghue, T., Scotchmer, S., Thisse, J.F. (1998). "Patent breadth, patent life and the pace of technological progress". Journal of Economics and Management Strategy 7, 1–32.
Ordover, J.A., Willig, R.D. (1978). "On the optimal provisions of journals qua sometimes shared goods". American Economic Review 68, 324–339.
O'Rourke, M. (2000). "Toward a doctrine of fair use in patent law". Columbia Law Review 100, 1177–1250.
Ottoz, E., Cugno, F. (2004). "The independent invention defence in a Cournot-Duopoly model". Economics Bulletin 12, 1–7.
Pakes, A. (1985). "On patents, R&D and the stock market rates of return". Journal of Political Economy 93, 390–409.
Pakes, A. (1986). "Patents as options: some estimates of the value of holding European patent stocks". Econometrica 54, 755–784.
Pakes, A., Schankerman, M. (1984). "The rate of obsolescence of patents, research gestation lags, and the private rate of return to research resources". In: Griliches, Z. (Ed.), R&D, Patents, and Productivity. University of Chicago Press, Chicago.
Parchomovsky, G. (2000). "Publish or perish". Michigan Law Review 98, 926–952.
Parchomovsky, G., Siegelman, P. (2002). "Toward an integrated theory of intellectual property". Virginia Law Review 88, 1455–1458.
Parchomovsky, G., Wagner, R.P. (2005). "Patent portfolios". University of Pennsylvania Law School, Public Law Working Paper 56. (At http://ssrn.com/abstract=582201.)

Park, Y., Scotchmer, S. (2004). Digital Rights Management and the Pricing of Digital Content. NET Institute, New York University.

Png, I.P.L., Reitman, D. (1995). "What are some products branded and others not?" The Journal of Law & Economics 38, 207–224.

Polinsky, A.M. (1980). "Resolving nuisance disputes: the simple economics of injunctive and damage remedies". Stanford Law Review 32, 1075.

Polinsky, A.M., Shavell, S. (1997). "Punitive damages: an economic analysis". Harvard Law Review 111, 869–962.

Port, K.L. (1994). "The 'unnatural' expansion of trademark rights: is a federal dilution statute necessary?" Seton Hall Legislative Journal 18, 433–488.

Posner, R. (1985). The Federal Courts. Harvard University Press, Cambridge.

Quillen, C. Jr., Webster, O. (2001–2002). "Continuing patent applications and performance of the U.S. Patent and Trademark Office". Federal Circuit Bar Journal 1, 1–21.

Rai, A.K. (2000). "Addressing the patent gold rush: the role of deference to PTO patent denials". Washington University Journal of Law & Policy 2, 199–227.

Rai, A.K. (2002). "Specialized trial courts: concentrating expertise on fact". Berkeley Technology Law Journal 17, 877–897.

Rai, A.K. (2003). "Engaging facts and policy: a multi-institutional approach to patent system reform". Columbia Law Review 103, 1035–1135.

Raymond, E. (2001). The Cathedral and the Bazaar. O'Riley Press, New York.

Reinganum, J. (1981). "Dynamic games of innovation". Journal of Economic Theory 25, 21–41.

Reinganum, J. (1982). "A dynamic game of R&D: patent protection and competitive behaviour". Econometrica 50, 671–688.

Reinganum, J. (1985). "Innovation and industry evolution". Quarterly Journal of Economics 100, 81–99.

Reinganum, J. (1989). "The timing of innovation: research, development and diffusion". In: Schmalensee, R., Willig, R.D. (Eds.), Handbook of Industrial Organization. Elsevier, Amsterdam, pp. 849–908.

Robinson, J. (1933). The Economics of Imperfect Competition. Macmillan, London.

Rockett, K.E. (1990). "Choosing the competition and patent licensing". The Rand Journal of Economics 21, 161–171.

Rosenberg, N. (1972). "Factors affecting the diffusion of technology". Explorations in Economic History 10, 3–33.

Ruttan, V.W. (2001). Technology, Growth and Development: An Induced Innovation Perspective. Oxford University Press, Oxford.

Ryan, M.P. (1998). Knowledge Diplomacy. Brookings Institution Press, Washington, D.C.

Samuelson, P. (1990). "Benson revisited: the case against patent protection for algorithms and other computer program-related inventions". Emory Law Journal 39, 1025–1154.

Samuelson, P. (2004). "Intellectual property arbitrage: how foreign rules can affect domestic protections". University of Chicago Law Review 71, 223–239.

Samuelson, P., David, R., Kapor, M., Reichman, J.H. (1994). "A manifesto concerning the legal protection of computer programs". Columbia Law Review 94, 2308–2431.

Samuelson, P., Scotchmer, S. (2002). "The law and economics of reverse engineering". Yale Law Journal 111, 1575–1663.

Sandburg, B. (2001). "Battling the patent trolls". The Recorder, July 30.

Saxenian, A. (1994). Regional Advantage Culture and Competition in Silicon Valley and Route 128. Harvard University Press, Cambridge.

Schankerman, M. (1998). "How valuable is patent protection? Estimates by technology field". The Rand Journal of Economics 29, 77–107.

Schankerman, M., Pakes, A. (1986). "Estimates of the value of patent rights in European countries during the post-1950 period". Economic Journal 97, 1052–1076.

Schankerman, M., Scotchmer, S. (2001). "Damages and injunctions in protecting intellectual property". Journal of Industrial Economics 32, 199–220.

Schechter, F.I. (1927). "The rational basis of trademark protection". Harvard Law Review 40, 813.
Scherer, F.M., Ross, D. (1990). Industrial Market Structure and Economic Performance, 3rd edn. Houghton Mifflin, Boston.
Schumpeter, J.A. (1942). Capitalism, Socialism and Democracy. Harper & Row, New York.
Scotchmer, S. (1991). "Standing on the shoulders of giants: cumulative research and the patent law". Journal of Economic Perspectives 5, 29–41.
Scotchmer, S. (1996). "Protecting early innovators: should second-generation products be patentable?" Journal of Industrial Economics 27, 322–331.
Scotchmer, S. (1999). "On the optimality of the patent system". The Rand Journal of Economics 30, 181–196.
Scotchmer, S. (2003). "Intellectual property: when is it the best incentive mechanism for S&T data?" In: National Research Council, The Role of the Public Domain in Scientific and Technical Data and Information. National Academies Press, Washington, D.C.
Scotchmer, S. (2004a). "The political economy of intellectual property treaties". Journal of Law, Economics and Organizations 20, 415–437.
Scotchmer, S. (2004b). Innovation and Incentives. MIT Press, Cambridge.
Scotchmer, S. (2004c). "Patent design, patent quality, and patent politics". Remarks before the European Patent Office, Munich, Germany.
Scotchmer, S. (2005a). "Consumption externalities, rental markets and purchase clubs". Economic Theory 25, 235–253.
Scotchmer, S. (2005b). Innovation and Incentives. MIT Press, Cambridge.
Scotchmer, S., Green, J. (1990). "Novelty and disclosure in patent law". Journal of Industrial Economics 21, 131–146.
Shapiro, C. (1982). "Consumer information, product quality, and seller reputation". Bell Journal of Economics 13, 20–35.
Shapiro, C. (1983). "Premiums for high-quality products as returns to reputations". Quarterly Journal of Economics 98, 659–679.
Shapiro, C. (1985). "Patent licensing and R&D rivalry". American Economic Review 75, 25–30.
Shapiro, C. (2001). "Navigating the patent thicket: cross licenses, patent pools, and standard-setting". Innovation Policy and the Economy 1, 119–150.
Shapiro, C. (2003). "Antitrust limits to patent settlements". Journal of Industrial Economics 34, 391–411.
Shapiro, C. (2004). "Patent system reform: economic analysis and critique". Berkeley Technology Law Journal 19, 1017–1047.
Shavell, S., van Ypersele, T. (2001). "Rewards versus intellectual property rights". Journal of Law and Economics 44, 525–547.
Sheldon, J., Carter, C.L. (1997). "Unfair and deceptive acts and practices". National Consumer Law Center.
Shy, O., Thisse, J.-F. (1999). "A strategic approach to software protection". Journal of Economics and Management Strategy 8, 163–190.
Sidak, J. (2004). "Trade secrets and the option value of involuntary exchange". (At: http://ssrn.com/abstract=577244.)
Sobel, D. (1995). Longitude: The True Story of a Lone Genius Who Solved the Greatest Scientific Problem of His Time. Walker & Co., New York, NY.
Solow, R.M. (1957). "Technical change and the aggregate production function". Review of Economics and Statistics 39, 312–320.
Soobert, A.M. (1998). "Breaking new grounds in administrative revocation of U.S. patents: a proposition for opposition—and beyond". Santa Clara Computer and High Technology Law Journal 14, 63–187.
Sovern, J. (1991). "Private actions under the deceptive trade practices acts: reconsidering the FTC act as rule model". Ohio State Law Journal 52, 437–467.
Sprigman, C. (2004). "Reform(aliz)ing copyright". Stanford Law Review 57, 485–568.
Stigler, G.J. (1961). "The economics of information". Journal of Political Economy 69, 213–225.
Strandburg, K. (2004). "What does the public get? Experimental use and the patent Bargain". Wisconsin Law Review 2004, 81–153.

Suzumura, K. (1992). "Cooperative and noncooperative R&D in an oligopoly with spillovers". American Economic Review 82, 1307–1320.
Swann, J.B. (1980). "The economic approach to genericism: a reply to Folsom and Teply". The Trademark Reporter 70, 243–252.
Swann, J.B., Palladino, V. (1988). "Surveying 'genericness': a critique of Folsom and Teply". The Trademark Reporter 78, 179–196.
Swiss Reinsurance Company (2000). "The significance of intellectual property assets, risks and insurance". (At: http://www.swissre.com.)
Tandon, P. (1982). "Optimal patents with compulsory licensing". Journal of Political Economy 90, 470–486.
Tandon, P. (1983). "Rivalry and the excessive allocation of resources to research". Bell Journal of Economics 14, 152–165.
Taylor, C., Silberston, Z.A. (1973). The Economic Impact of the Patent System. Cambridge University Press, London.
Thomas, J.R. (1999). "The patenting of the liberal professions". Boston College Law Review 40, 1139–1185.
Thomas, J.R. (2001). "Collusion and collective action in the patent system: a proposal for patent bounties". University of Illinois Law Review 2001, 305–353.
Thomas, J.R. (2002). "The responsibility of the rulemaker: comparative approaches to patent administration reform". Berkeley Technology Law Journal 17, 727–761.
United States Department of Justice and Federal Trade Commission (1995). Antitrust Guidelines for Licensing of Intellectual Property.
United States Federal Trade Commission (2003). To Promote Innovation: The Proper Balance of Competition and Patent Law and Policy.
United States Patent and Trademark Office (2001). Guidelines for Written Description and Utility Requirements.
Varian, H. (2000). "Buying, renting, sharing information goods". Journal of Industrial Economics 48, 473–488.
Veblen, T. (1899). The Theory of the Leisure Class: An Economic Study of Institutions. Macmillan, New York.
von Hippel, E. (1988). The Sources of Innovation. Oxford University Press, Oxford.
von Hippel, E. (2001). "Innovation by user communities: learning from open source software". Sloan Management Review 42, 82–86.
Wagner, R.P. (2003). "Of patents and path dependency: a comment on Burk and Lemley". Berkeley Technology Law Journal 18, 1341–1360.
Wagner, R.P., Petherbridge, L. (2004). "Is the federal circuit succeeding? An empirical assessment of judicial performance". University of Pennsylvania Law Review 152, 1105–1180.
Walsh, J.P., Arora, A., Cohen, W.M. (2003). "Effects of research tool patents and licensing on biomedical innovation". In: Cohen, W.M., Merrill, S.A. (Eds.), Patents in the Knowledge-Based Economy. National Academy Press, Washington, D.C., pp. 285–340.
Wilkins, M. (1992). "The neglected intangible asset: the influence of the trade mark on the rise of the modern corporation". Business History 34, 66–95.
Wiley, J.S. Jr. (2003). "Taming patent: six steps for surviving scary patent cases". UCLA Law Review 50, 1413–1483.
Williamson, O. (1986). Economic Organization: Firms, Markets, and Policy Control. New York University Press, New York.
Wright, B. (1983). "The economics of invention incentives: patents, prizes and research contracts". American Economic Review 73, 691–707.
Ziedonis, R.M. (2003). "Patent litigation in the U.S. semiconductor industry". In: Cohen, W.M., Merrill, S.A. (Eds.), Patents in the Knowledge-Based Economy. National Academy Press, Washington, D.C., pp. 180–212.

Cases

Cuno Engineering Corp. v. Automatic Devices Corp., 314 U.S. 84 (1941).
DaimlerChrysler AG v, Bloom, 315 F.3d 932 (8th Cir. 2003).
Diamond v. Chakrabarty, 447 U.S. 303 (1980).
Fonar Corp. v. General Electric, 107 F.3d 1543 (Fed. Cir. 1997).
Gottschalk v. Benson, 409 U.S. 63 (1972).
Graver Tank & Mfg. Co. v. Linde Air Products Co., 339 U.S. 605 (1950).
King-Seeley Thermos Co. v. Aladdin Industries, Inc., 321 F.2d 577 (2d Cir. 1963).
Madey v. Duke Univ., 307 F.3d 1351 (Fed. Cir. 2002), *cert. denied* 539 U.S. 958 (2003).
Mattel, Inc. v. MCA Records, 296 F.3d 894 (9th Cir. 2002), *cert. denied* 537 U.S. 1171 (2003).
O'Reilly v. Morse, 56 U.S. 62 (1854).
Roche Prods., Inc. v. Bolar Pharmaceutical Co., 733 F.2d 858 (Fed. Cir. 1984).
Trade-Mark Cases, 100 U.S. 82 (1879).
Wal-Mart Stores, Inc. v. Samara Brothers, Inc., 529 U.S. 205 (2000).
Warner-Jenkinson Co. v. Hilton Davis Chem. Co., 520 U.S. 17 (1997).
Whittemore v. Cutter, 29 F. Cas. 1120 (C.C.D. Mass. 1813).
W.L. Gore & Assocs. v. Garlock, Inc., 721 F.2d 1540 (Fed. Cir. 1983).

Chapter 20

NORMS AND THE LAW

RICHARD H. MCADAMS[*]
School of Law, University of Chicago

ERIC B. RASMUSEN[†,‡]
Kelley School of Business, Indiana University

Contents

[*] *McAdams*: Professor of Law, University of Chicago Law School, 1111 60th Street, Chicago, Illinois 60637. Office: 773-834-2520. Fax: 773-834-4409.

[†] *Rasmusen*: Dan R. and Catherine M. Dalton Professor, Department of Business Economics and Public Policy, Kelley School of Business, Indiana University, BU 456, 1309 E. 10th Street, Bloomington, Indiana, 47405-1701. Office: 812-855-9219. Fax: 812-855-3354. http://www.rasmusen.org.

[‡] We thank Jonathan Baron, Lisa Bernstein, F.H. Buckley, Robert Ellickson, Timur Kuran, Thomas Miceli, Geoffrey Miller, Peter Ordeshook, Richard Posner, J. Mark Ramseyer, Barak Richman, and Thomas Ulen for their comments, and Sean Mead for research assistance.

Handbook of Law and Economics, Volume 2
Edited by A. Mitchell Polinsky and Steven Shavell

DOI: 10.1016/S1574-0730(07)02020-8

Abstract

Everyone realizes the importance of social norms as guides to behavior and substitutes for law, but coming up with a paradigm for analyzing norms has been surprisingly difficult, as has systematic empirical study. In this chapter we survey the topic.

JEL classification: A12, A13, D63, K0, Z13

1. Introduction

Law seeks to regulate behavior when self-interest does not produce the correct results as measured by efficiency or fairness. If people behave well without regulation, law is superfluous and just creates extra costs. If law is not what actually determines human behavior, scholars debating it are wasting their time. For this reason, law matters primarily to the "bad man" of Oliver Wendell Holmes, Jr. (1897). The "bad man" is, in effect, "economic man," caring only about the material consequences of his actions:

> You can see very plainly that a bad man has as much reason as a good one for wishing to avoid an encounter with the public force, and therefore you can see the practical importance of the distinction between morality and law. A man who cares nothing for an ethical rule which is believed and practised by his neighbors is likely nevertheless to care a good deal to avoid being made to pay money, and will want to keep out of jail if he can.

The man who is not "bad" in this sense, however, *is* influenced by the ethical rule, either because he cares directly about it or because he cares about other people who do. Since the perfect "bad man" is atypical, we should revise our first sentence above to say that law becomes relevant only when neither self-interest *nor social norms* provide the right incentives for behavior.

Since the early 1990s, considerable scholarship in law and economics has turned its attention to norms, as Ellickson (1998) details. Numerous articles and at least six law review symposium issues have addressed the power of social norms and their relevance to law (see "Symposium, 1996, 1998, 1999, 2000a, 2000b, 2001" in the References section). Holmes also said: "For the rational study of the law the blackletter man may be the man of the present, but the man of the future is the man of statistics and the master of economics." And indeed, the same economic methods useful for analyzing law are useful for analyzing norms, a tradition going back as far as Adam Smith (1776) [e.g., his explanation in *The Wealth of Nations* (1776, Book V, Chapter 1) of how religious sects flourish in the anonymity of cities to provide indicators of good morals]. Economics is eminently suitable for addressing questions of the various incentives mediated neither by the explicit price of some good nor by the threats of government, incentives such as guilt, pride, esteem and disapproval, which we contend underlie norms.

We will proceed as follows. Section 2 addresses the definition of "norms" and contrasts it with "conventions." Section 3 discusses the sources and workings of conventions and norms, paying particular attention to the normative incentives of guilt and esteem. Section 4 provides a general overview of the norms literature in law and economics, separately discussing how such regularities matter to the positive and normative analysis. Section 5 reviews applications of this literature to particular areas of law—torts, criminal law, constitutional law, and so forth. Section 6 concludes.

2. Defining "norms"

Ellickson's seminal work, *Order Without Law* (1992, p. 126) notes a fundamental ambiguity in the word *norm*, that it denotes "both behavior that *is* normal, and behavior that people *should* mimic to avoid being punished." Confusion arises because law and economics scholars use the term in both senses. All contributors to the literature seem to agree that a norm at least *includes* the element of a behavioral regularity in a group—what is typical or "normal"—but they do not agree on whether a norm also requires that the behavior be normatively required. "Norm" means merely equilibrium behavior in Picker (1997); Mahoney and Sanchirico (2001, 2003); and E. Posner (2000a, 2000b). Others, however, restrict the term to the combination of an attitudinal regularity and a behavioral regularity—i.e., the situation where people believe that the behavior is normatively appropriate (Cooter, 1996; Ellickson, 1991; Kaplow and Shavell, 2001a, 2002a, 2002b; McAdams, 1997, 2001a, 2001b).[1] The attendant attitude may be as strong as a perceived moral obligation—that most people believe that everyone should conform to the regularity and that it is wrong to do otherwise (Cooter, 1996; Kaplow and Shavell, 2001a)—or as weak as a simple sense of approval or disapproval (McAdams, 1997; Pettit, 1990). Normative attitudes not only add a distinct element to a behavioral regularity, they also contribute to stability by creating the *normative incentives*—guilt, esteem, shame—that we discuss below.[2]

Here we will define "norms" as behavioral regularities supported at least in part by normative attitudes. We will refer to behavioral regularities that lack such normative attitudes as "conventions." This is because we think it useful to have one term—"convention"—for a mere equilibrium that plays out without anyone holding beliefs about the morality of the behavior, and another term—"norm"—for a behavioral regularity associated with a feeling of obligation. This usage also aligns with that in other social sciences. By contrast, if norms are nothing but behavioral regularities without support from attitudes, norms are not really a subject distinct from game theory. Indeed, the concept of "norms" under the broad definition has been justly criticized by such scholars as Kahan (2001) and Scott (2000) as too broad to be useful.

In excluding conventions, we clearly exclude some of what the law-and-economics literature has discussed as "norms"—for example, the equilibria that emerge from the evolutionary models of Picker (1997) and Mahoney and Sanchirico (2001, 2003), and the signalling model of Eric Posner (2000a, 2000b). Similarly, we exclude what

[1] We include Ellickson (1991, p. 124), whom we read as implicitly referring to normative attitudes when he describes norms as a form of "social control," where "social control" means enforced rules of "normatively appropriate behavior."

[2] Ellickson notes (1991, p. 128) that "the best, and always sufficient, evidence that a rule is operative is the routine . . . administration of sanctions . . . upon people detected breaking the rule." Although we agree that third-party sanctions commonly reflect the existence of an attitudinal pattern—that the third parties believe the sanctioned behavior violates an obligation or at least that they disapprove of it—game theory shows that such an attitudinal pattern is not strictly necessary. See Mahoney and Sanchirico (2003). Third party enforcement can, in theory, exist merely as a matter of convention.

Hetcher (2004) calls "epistemic norms," regularities that arise when individuals faced with information scarcity follow the crowd as in the cascades of Banerjee (1992) and Bikhchandani, Hirshleifer, and Welch (1992). All these contributions are useful, but we see their point as explaining what *seem* to be norms, motivated by feelings of right and wrong, as really being something else—conventions motivated by simple self-interest.

Even using the narrow definition of norm, conventions remain relevant. First, conventions are invaluable for testing whether a norm-based explanation is strictly necessary. As a first step, ask of each behavioral regularity whether it is really due to a convention. Often it will be, and there is no need to employ the special tools of this chapter. Second, conventions sometimes explain the origin of norms. Human beings quickly come to hold normative attitudes about an existing state of affairs, believing that other people *should* do what they are expected to do, especially when unexpected behavior causes harm (Sugden, 1998). Once everyone expects motorists to drive on the right side of the road, we come to believe that someone who drives on the left is not just foolish, but immoral. What is at first merely a convention becomes a norm. In such cases, an understanding of what maintains the end state requires the idea of norms, but the best tools for understanding norm origin come from game theory. We will therefore discuss conventions in some detail below.

Aside from definitions, there remain other sources of confusion that we hope to avoid. First, although sociologists and anthropologists refer to "legal norms," we will, following the convention of the legal literature, discuss norms as distinct from law. Although we comment below on the two important meta-law norms of legal obedience and the rule of law, we view law and norms as distinct incentives for behavior. Second, some theorists use "norms" to refer only to decentralized and informally created regularities, while others use the term to refer to rules of private institutions or organizations—rules that are often highly centralized and formal. We consider norms to encompass both types of regularities, though we recommend the term "organizational norms" to refer to centralized norms. Third, scholars such as Miller (2003) and Strahilevitz (2000, 2003) refer to norms that arise between strangers in large populations, whereas others, such as Bernstein (1992) and Ellickson (1991), discuss the norms of small and close-knit subpopulations. Norms in the sense we study here arise in both settings, though we will use the term "group norm" to refer to norms limited to a particular group. Finally, some theorists implicitly reserve the term "norms" to refer only to general regularities, such as the norms of reciprocity or individualism, while others use the term for specific regularities, such as giving gifts on Secretary's Day or shutting off cell phones in church. Norms under our definition encompass regularities at all levels of generality.

It is also important to distinguish norms from the rules of thumb and psychological heuristics studied by behavioral economics. Books such as Kahneman, Slovic, and Tversky's 1982 *Judgment Under Uncertainty: Heuristics and Biases* and Dawes's 1988 *Rational Choice in an Uncertain World* document and discuss many cognitive biases and compensating heuristics, but it is quite possible for a decisionmaker to be perfectly rational yet driven by norms, or radically irrational yet indifferent to norms. If most individuals in a social group eat spinach ice cream, a conventional economist might

rest content with the explanation that they like the flavor, while a behavioral economist might attribute that odd behavior to a bias or heuristic. A norms scholar, in contrast, would look for whether there was a desire to conform to what others expect and approve and would check to see if people in the group believed eating spinach ice cream was morally obligatory. Heuristics and rules of thumb do have important implications for laws and lawmaking [see, e.g., Baron (2001)], and they have been called norms (e.g., Epstein, 2001), but they really are a different subject. Psychology does, however, have application in the experimental study of what people mean by such things as "fairness," as may be seen in Thibaut and Walker's 1975 *Procedural Justice: A Psychological Analysis* and the literature that followed it [e.g., the criticism in Hayden and Anderson (1979) and Rabin (1993)]. Much may be discovered by experiments such as those of Cox and Deck (2005) that do not investigate only whether people behave as the simplest economic models of rational and selfish decisionmaking predict, but also carefully distinguish between different possible motives for deviating from the simple model.

3. How norms work

In this section, we will first discuss what we mean by normative incentives, and then contrast that with the numerous ways in which conventions can imitate, generate, or sustain norms.

3.1. Types of normative incentives

People feel obligations in a variety of ways, some internal and some external. Normative incentives are frequently negative—costs imposed on those who fail to conform to a behavioral regularity (such as guilt from not protecting a child from drowning)—but can also be positive—benefits conferred on those who exceed the normative requirement (as a person who incurs great danger saving a stranger's life). A significant literature documents and discusses negative sanctions, usually imposed by third parties but sometimes by the victim of a norm violation. Examples include gossip (Ellickson, 1991, pp. 214–215; McAdams, 1996); admonishment and insult (Miller, 2003; Buckley, 2003); social ostracism and shunning (E. Posner, 1996a); economic boycott and exclusion (Bernstein, 1992, 1996, 1999, 2001; Skeel, 2001); property destruction (Ellickson, 1991, pp. 215–219; Miller, 2003, p. 931); and violence (McAdams, 1995; Milhaupt and West, 2000, p. 68). Positive sanctions have received less attention, but see Ellickson's (1991, pp. 236–239) discussion of rewards for third party norm enforcers.

Whether they are positive or negative, by "normative incentives," we do not mean merely these external sanctions, which are the proximate but not ultimate influence on behavior. Instead, we must ask why third parties ever bother to incur the costs of sanctioning norm violators. Often it is possible to explain the third party behavior as itself part of a convention, not dependent on normative beliefs—see Mahoney and Sanchirico (2003), Hirshleifer and Rasmusen (1989) or West (1997). Underlying a norm in the

strict sense of the word, however, is a non-material motivation, either for the primary behavior of the person who follows the norm or for the secondary behavior of the people who reward his conformity or punish his violation.

The place to look for norms as opposed to conventions, therefore, is in the utility function. Normative attitudes are beliefs about the appropriateness of behavior, and the starting point for analysis is how these beliefs influence utility. Consider three possibilities.

(i) Guilt and pride An internalized normative incentive means that an individual sanctions himself. Guilt is disutility that arises when a person behaves in ways he thinks morally wrong. The converse, pride, is utility that arises when he behaves in ways he thinks virtuous. That someone can feel guilt and pride is equivalent to saying that he has a taste for behaving in conformity with his moral beliefs. Moral philosophers have for a considerable time emphasized the role of guilt and pride in moral behavior, as in Hume (1751, p. 150). These incentives do not require that anyone else know how the person acted. Nor do they preclude the individual from acting contrary to his moral beliefs—sometimes the payoffs for doing so are greater than the anticipated guilt costs. As elsewhere in economics, in this style of analysis the individual calculates what maximizes his utility and acts accordingly. As elsewhere, the empirical prediction is that when prices change, so will behavior: if the material benefit of norm violation rises while the guilt penalty stays constant, we will observe more violations. This is a crucial point that economics brings to the study of norms. [For a discussion of how sociology and economics interact, see Ramseyer (1994).]

As with other tastes in the utility function, if a person's taste for pride and distaste for guilt varies widely from day to day, the rational-actor approach will not yield useful predictions. By the end of childhood, however, the moral beliefs that underlie guilt and pride are fixed enough to be difficult to change. Some psychologists claim that there is a genetic basis for guilt. This idea has been picked up in law-and-economics by Rubin (1982), Richard Posner (1997) and Kaplow and Shavell's (2002a) *Fairness and Welfare* because evolutionary theory can explain moral tastes in the same way that it explains the taste for leisure or sweets. Part of the evolutionary explanation is the insight that potential feelings of guilt can be useful as a means of self-control, especially if this potential is visible to others.

(ii) Esteem and disapproval Esteem is a normative incentive that exists if a person cares intrinsically (in addition to instrumentally) what others believe about his behavior. Someone might gain utility directly from believing that others esteem him and lose utility from believing that others disapprove of him, regardless of whether these outsiders take actions that materially affect him. This effect of others' beliefs on one's utility is equivalent to saying that a person has a taste for others' esteem. The idea is older than Adam Smith, but he put it well when he said in *The Theory of Moral Sentiments* that

> Nature, when she formed man for society, endowed him with an original desire to please, and an original aversion to offend his brethren. She taught him to feel pleasure in their favourable, and pain in their unfavourable regard. She rendered their approbation most flattering and most agreeable to him for its own sake; and their disapprobation most mortifying and most offensive (Smith, 1790, p. 116).

Unlike utility from pride or guilt, utility from esteem or disapproval arises only when one believes other people have formed beliefs about one's behavior. Disapproval can therefore be avoided by misbehaving secretly. As with guilt, the benefits of acting contrary to what others approve may outweigh the expected disapproval, especially when the disapproval is contingent on the offending behavior being detected. On the other hand, proper behavior is no guarantee of esteem, because esteem depends on one's perception of other people's beliefs, not on one's own behavior. Not just good conduct, but other's knowing about it—and knowing that they know about it—is necessary for esteem. And one may gain esteem without good behavior by fooling others into thinking one has behaved well.

Esteem and disapproval differ too from *praise* and *censure*, which are merely the *expression* of esteem and disapproval. Esteem and approval are subjective, based on beliefs about others' opinions rather than on the actual opinions or their public declaration. Praise and censure are evidence of what others believe but expression is not necessary for an individual to believe that others have formed judgments of esteem or disapproval. The fact that actual expression is not required reduces the transaction costs of esteem as an incentive—though it also can lead to misincentives because of misperceptions. See Kuran (1995). Note, too, that praise and censure might also be valued for their own sake; one may value the expression even if it is already common knowledge that the speaker holds the expressed view, or even if it is common knowledge that the speaker is being hypocritical. The sweetest congratulation might be from a disappointed rival.

Brennan and Pettit (2004, Chapter 1) have traced the history of the idea of esteem, and Fershtman and Weiss (1998) identify conditions under which a preference for esteem (or what they call "status") is evolutionarily stable. Various theorists, including Pettit (1990), McAdams (1997), Brennan and Pettit (2000), Cowen (2002), and Brennan and Pettit (2004) use esteem as the key explanation of norms.

(iii) Shame There is a third possibility. Some scholars, e.g. R. Posner and Rasmusen (1999), distinguish shame from guilt. Often, shame is used to mean what we have termed disapproval, though with an emphasis on particularly intense and widespread disapproval. Shame might, however, mean something else: a negative emotion that arises from believing one has failed to meet standards set by the normative beliefs of others. On this account, shame falls between guilt and disapproval. Like guilt, it is an internalized sanction that occurs even if no one observes a norm being broken. Unlike guilt, the person feeling shame has failed to live up to the normative beliefs of others, which may be the case even if he has lived up to his own principles. As with

disapproval, the standards of behavior are external, but unlike disapproval, the shamed person suffers disutility regardless of what others think. Suppose someone privately engages in sexual behavior X without feeling guilt (because he has not violated his own moral principles) or disesteem (because nobody else knows he has done it). Later he discovers that a friend strongly disapproves of X. The loss of utility that occurs only at this discovery would clearly not be guilt—by his own principles, he has done nothing wrong—nor disapproval—since the friend does not know that he has done X. We need a new category: shame. Likewise, there is a positive incentive analogous to pride or esteem if someone gains utility from successfully living up to the standards set by the normative beliefs of others, regardless of where whether he holds those same normative beliefs or whether others know he has succeeded (cf. McAdams, 1997, pp. 382–386). Shame and guilt are, of course subjects long studied in psychology. For entry into the literature, see Cosmides, Tooby, and Barkow (1992), Harder (1995), and Tangney (1995).

3.2. Conventions

As noted above, many behavioral regularities that seem normative may in part or whole be motivated by non-moral concerns, even when not driven by common tastes or fear of government penalties. These are conventions. Scholars in law and economics were analyzing social behavior driven by what we call conventions well before the word "norms" became popular, e.g., Brinig (1990) on wedding rings and Schwartz, Baxter, and Ryan (1984) on dueling. A number of simple ideas from game theory can explain seemingly normatized behavior as driven by the usual incentives studied by economists, with no need to appeal to tastes.

One of the most important settings for conventions is the coordination game, in which the payoffs of the players are highest if they coordinate with each other. The problem is not conflicting desires but the need to avoid a discoordination that hurts everyone. This game leads the establishment of standards, whose importance is explained in Kindleberger (1983). A simple example is driving on the right side of the road.

Conventions also are important in repeated games, in particular when reputations can arise. Klein and Leffler's seminal (1981) article on reputation essentially models it as an equilibrium of a repeated game in which a player is willing to forgo present profits in exchange for a good reputation that will yield him future profits. It may look as if a seller is providing high quality out of pride of workmanship or fear of disapproval, but he is actually motivated purely by material gain. Hirshleifer and Rasmusen (1989) use the idea of repeated games to explain ostracism—the expulsion of rule-breakers from groups, and Axelrod and Hamilton (1981) show the power of reciprocal altruism in "tit-for-tat."

Signalling equilibria create still another form of convention. Someone may take a costly action to signal his inclinations or ability. This occurs if someone with baser inclinations or lower abilities would not be willing to bear the cost of the signal, whether it be the provision of advertising or restraint in taking advantage of the uninformed,

a requirement known as the "single-crossing property" because it can be formalized as requiring that the indifference curves in money-signal space of different types of agents cross only once (see Rasmusen, 2006, Chapter 12). For example, E. Posner (1998, 2000b) has explained a wide variety of behaviors as signals of one's discount rate, which is important to revealing one's suitability as a partner in repeated games (though see McAdams (2001b) for a critique), and Fremling and R. Posner (1999) apply signalling models to sexual harassment law. Often, however, it is hard to tell which convention is at work—signalling information or reciprocating in a repeated game—as Kahan (2002) observes.

Sometimes conventions are formalized in the shape of institutions, as demonstrated by Ostrom (1990, 1991) in general, Cooter (1991) in the land system in New Guinea, and Milhaupt and West (2000) in organized crime. Institutions are rule-setting bodies that unlike government lack the power to coerce through the use of legal force but that can use conventions—involving ostracism, reputation, or information transmission—to enforce their rules.

Since these convention models so often obviate the need to use norms to explain behavior, we will lay them out in slightly greater detail before proceeding to analysis of norms proper.

Coordination games In a coordination game, two or more players make choices that will help them both if they match. Two drivers, Row and Column, may each need to decide whether to drive on the right side of the road or the left as they approach each other. The most important thing for each is that they make the same choice (which will mean that they avoid hitting each other). Assume it is also better if both choose to drive on the right, since they are driving cars with steering wheels on the left side. Table 1 shows the payoffs.

This game has two Nash equilibria if the choices are made simultaneously—(*Right, Right*) and (*Left, Left*). These equilibria can be Pareto ranked, but each is an equilibrium. If each expects the other to drive on the *Left*, that is a set of self-fulfilling expectations in a simultaneous-move game. If the game were sequential, the only equilibrium would be for Row to choose *Right* and for Column to follow a strategy of imitating Row.

Table 1
Ranked coordination

		Column	
		Drive on Right	Drive on Left
Row	Drive on Right	7,7	0,0
	Drive on Left	0,0	6,6

Payoffs to: (Row, Column).

Many behavioral regularities are coordination games. Such behavioral regularities are often called norms, but not in our terminology because they are driven by simple self-interest rather than normative beliefs. Normative rules are not necessary to persuade people to avoid self-destruction in car crashes.

The repeated prisoner's dilemma A second major category of convention model is the repeated prisoner's dilemma. Unlike coordination games, prisoner's dilemmas have complete conflict between the objectives of the players. In the classic story, two prisoners, Row and Column, are being questioned separately. If both confess, each is sentenced to eight years in prison. If both deny their involvement, each is sentenced to one year. If just one confesses, he is released but the other prisoner is sentenced to ten years, as shown in Table 2.

The equilibrium of Table 2's game is (*Confess, Confess*), with equilibrium payoffs of $(-8, -8)$, worse for both players than $(-1, -1)$. Sixteen, in fact, is the greatest possible combined total of years in prison.

So far, no useful convention has emerged. But what if the game is repeated? Would the players arrive at a convention of choosing *Deny* in the early repetitions, knowing that they will be in the same situation in the future, with the possibility of revenge? Not if this is all there is to the game. Using an argument known as the Chainstore Paradox after its application to store pricing (where the *Deny/Confess* actions become *Price-High/Undercut-Price*), Selten (1978) explains that in the last repetition, the players will choose *Deny* because future revenge will be impossible, so in the second-to-last repetition the players will not have any hope for future cooperation, so in the third-to-last they will have no hope, and so on to the first repetition.

If the game is infinitely repeated, the Chainstore Paradox does not apply, and there exists an equilibrium in which the players choose *Deny* each time. Real-world interactions do not last forever, but Kreps et al. (1982) show that with incomplete information, the addition of a small possibility of emotional behavior by a player such that he will choose *Deny* until the other player chooses *Confess*, can make *(Deny, Deny)* an equilibrium until near the last repetition. This is true even if the game does have a definite end, because if the other player does not know whether his opponent is emotional in this way

Table 2
The prisoner's dilemma

		Column	
		Deny	Confess
Row	Deny	$-1, -1$	$-10, 0$
	Confess	$0, -10$	$-8, -8$

Payoffs to: (Row, Column).

or not, his best strategy turns out to be to treat him gently until late in the game. The infinitely repeated game with complete information is often used as a simpler model that comes to conclusions similar to those of the more realistic but more complicated finitely repeated game with incomplete information.

Signalling The last type of convention model that we will describe here is the signalling game—one which is especially prominent in the norms literature because it is a central idea in Eric Posner's work, including in his 2000 *Law and Social Norms*. We will use a particular example from R. Posner and Rasmusen (1999), a model of employers preferring married over single workers. Suppose that 90 percent of workers are "steady," with productivity $p = x$, and 10 percent are "wild," with productivity $p = x - y$. Each worker decides whether to marry or not. Marriage creates utility $u = m$ for a steady worker and utility $u = -z$ for a wild worker. Employers, observing whether workers are married but not whether they are wild, offer wages w_m or w_u in competition with other employers, depending on whether a worker is married or not. We observe that $w_m > w_u$.

We do not need norms to explain the higher wage for married workers. Employers have incentive to use marital status as a signal of productivity and to discriminate against single workers even if nobody thinks that marriage per se makes someone better or worse. The employer has no intrinsic reason to care whether the worker is married or not, since wild workers are less productive whether they are married or not. The only significance of marriage for the employer is its informational value as a signal of steadiness.

Unlike many signalling models, here there is only a single equilibrium. If z is large enough (greater than y), the employer will pay wages of $w_u = x - y$ and $w_m = x$, the steady worker will get married, and the wild worker will stay single. Steady workers will marry regardless of the effect on their wage, and wild workers will stay single even though they know that if they married an employer could be fooled into believing them to be steady—an example of the "single-crossing property" mentioned above.

The employers in this example might be unthinkingly obeying a rule of thumb of paying married workers more. Businessmen, like private individuals, follow many rules of behavior without inquiring into their rationality. Following the rule is efficient and profit-maximizing even if no businessman understands its origin or rationale. When asked, an employer might say he pays married workers more because they deserve the higher wage, or need the higher wage, even though that is not the true reason. Thus, the convention of signalling is easily confused with a norm.

Signalling has implications for how laws should be designed. In this model, subsidizing marriage not only would be useless for raising productivity, but would lower it by depriving employers of useful information about the marginal product of their workers. Similar loss of information would occur if government forbade employers to use an applicant's marital status in making a hiring decision. Thus, in this model, it would be wrong for the government to start with the true premise that married workers are more productive and arrive at the conclusion that if more workers were married, productivity

would rise; but it would also be wrong for the government to start with the equally true premise that a worker's getting married has no effect on his productivity and arrive at the conclusion that it would make no economic difference if firms were forbidden to discriminate by marital status.

Signalling models must be treated with care. They are "all-purpose" models that can "explain" practically any pattern of observed behavior give the right assumptions. The model above, for example, could as easily have been made a model in which steady workers derive *less* direct utility from marriage, in which case *singleness* would be the signal of ability, not marriage. This flexibility is both a strength and a weakness of signalling models.

Bayesian learning in cascade and bandit models What seems to be norm-based behavior can also be entirely non-strategic, so neither norms nor conventions are needed to explain group behavior. One example is Rasmusen (1996), which explains stigma against the employment of criminals as arising from employer calculations of average ability based on population averages that can "tip" the level of criminality even if no single worker or employer thinks his own behavior will affect which equilibrium is played out. Another is the single decisionmaker "Two-Armed Bandit" model of Rothschild (1974), which shows how seemingly irrational, mistaken behavior can arise as the result of a rational policy of first investigating various possible behavior rules and then settling down to what seems best and never again experimenting.

A model of this type which has attracted considerable attention is the theory of cascades, originating with Banerjee (1992) and Bikhchandani, Hirshleifer, and Welch (1992) and summarized in Hirshleifer (1995). It shows how fashions and fads may be explained as simple Bayesian updating under incomplete information, without any strategic behavior. Consider a simplified version of the first example of a cascade in Bikhchandani, Hirshleifer, and Welch (1992). A sequence of people must decide whether to *Adopt* at cost 0.5 or *Reject* a project worth either 0 or 1 with equal prior probabilities, having observed the decisions of people ahead of them in the sequence plus a private signal. Each person's private signal is independent. A person's signal takes the value *High* with probability $p > 0.5$ if the project's value is 1 and with probability $(1 - p)$ if the project's value is 0, and otherwise takes the value *Low*.

The first person will simply follow his signal, choosing *Adopt* if the signal is *High* and *Reject* if it is *Low*. The second person uses the information of the first person's decision plus his own signal. One Nash equilibrium is for the second person to always imitate the first person. It is easy to see that he should imitate the first person if the first person chose *Adopt* and the second signal is *High*. What if the first person chose *Adopt* and the second signal is *Low*? Then the second person can deduce that the first signal was *High*, and choosing on the basis of a prior of 0.5 and two contradictory signals of equal accuracy, he is indifferent—and so will not deviate from an equilibrium in which his assigned strategy is to imitate the first person when indifferent. The third person, having seen the first two choose *Adopt*, will also deduce that the first person's signal was *High*. He will ignore the second person's decision, knowing that in equilibrium that person

just imitates, but he, too will imitate. Thus, the first person's decision has started a cascade, and even if the sequence of signals is (*High, Low, Low, Low, Low*...), everyone will choose *Adopt*. A "cascade" has begun, in which players later in the sequence—starting with the second one in this example!—ignore their own information and rely on previous players completely. We have chosen an extreme example, in which the cascade starts immediately with probability one, but the intuition is robust, and more complicated models yield interesting implications for how signal quality and correlation affect the probability of a cascade starting.

Learning models such a these are useful for modeling apparently irrational behavioral regularities. Suppose we observe a culture that tries to cure malaria by bleeding the patient. This does not have to be the result of norms. Rather, it may be that after trying other methods and failing—perhaps even trying quinine bark without consistent success—the tribe has rationally if mistakenly settled down to bleeding as the best method based upon available evidence. But this is neither a norm in our sense nor a convention, since it is the result of neither obligations nor strategic interactions.

Conventions interact with normatized incentives. Kreps et al. (1982) show how just a few people with normatized incentives can lead many others to imitate them in their behavior. Kuran's (1995) *Private Truths, Public Lies: The Social Consequences of Preference Falsification* shows how such deception about one's true preferences can lead to sudden reversals of public opinion. One of Kuran's examples is the collapse of communist regimes when most citizens suddenly discovered that the support for the regime was not genuine but based on a complex "web of lies." Similarly, Kuran and Sunstein (1999) propose that informational and reputational cascades sometimes combine to cause stampedes toward ill considered regulations, and Kuran (1998) analyzes the interaction between cascades and norms in the context of ethnic identification. Kuran and Sunstein distinguish "informational" cascades of the sort just described and "reputational" cascades that occur only because individuals expect to gain by being known to conform. In our terminology, informational conformity is one way to produce a convention; reputational conformity is one way to produce a norm.

Thus, after discussing such diverse convention models as signalling, repeated prisoner's dilemmas, and cascades, we see that much of human behavior that seems to be driven by moral beliefs is actually driven by utility maximization in the narrow sense of Holmes's bad man, though by a bad man sophisticated enough to know how important strategic behavior is to his success. La Rochefoucauld said, "Hypocrisy is the homage vice pays to virtue." In the present context, "Convention is the homage *homo economicus* pays to norms."

3.3. The origin of norms

Although we have shown how a variety of apparent norms could actually be conventions, the study of conventions is important to norms more than just for explaining them away. While we have distinguished between conventions that work by appealing

to standard, non-normative tastes and norms that work only when supported by feelings of obligation, there remains the question of where those feelings come from. Conventions are an important part of the answer because people easily come to believe that others should do what they are expected to do, especially when unexpected behavior causes harm (Sugden, 1998).

How people come to have any of the three normative drives just discussed—guilt, esteem, or shame—is a subject given considerable attention by biologists ever since Darwin's (1874) *Descent of Man*. For the biologist, any kind of tastes, standard, or norm is the result of an equilibrium, an evolved outcome of a process similar to maximization, although less calculated and with results harder to call "optimal" in a meaningful way. Biologists have also studied what would be conventions in humans (because motivated by calculation) but are commonly norms in animals (because motivated by preferences—the inborn preferences we call instinct). Though genes are "selfish," E. Wilson (1980) shows that there are conditions under which helping behavior is necessary to survive—e.g., hunting large prey, warding off predators, etc.—and motives such as guilt, shame, or esteem may induce such helping behavior. See Trivers (1971); Jack Hirshleifer (1978, 1987); Fershtman and Weiss (1998). The approach has been picked up in ethics (Peter Singer's 1981 *The Expanding Circle: Ethics and Sociobiology*), anthropology (Boyd and Richerson's 1988 *Culture and the Evolutionary Process*, which applies evolution to the "functionalism" of, e.g., Marvin Harris's 1974 *Cows, Pigs, Wars and Witches: The Riddles of Culture*), political science [Ostrom (1991) and her (1990) *Governing the Commons*], and economics [Jack Hirshleifer (1978), Bergstrom (2002) and Sartorius (2002) generally; Cameron (2001) on sexual behavior].

The biological approach is really an extension of the idea that humans are born with certain norms instilled in them (e.g. *Romans* 1, Aristotle's *Nichomachean Ethics*, Aquinas's *Summa Theologica*), an idea that under the name of "natural law" is the subject of a quite different branch of scholarship [e.g., James Q. Wilson's *The Moral Sense* (1993); Budziszewski's *Written on the Heart: The Case for Natural Law* (1997); the essays in George (1995)]. The biological approach, however, with its analytic framework of evolution as a source of utility functions, has proven more useful than natural law as a source of explanations in law and economics.

Biological evolution brings to mind the literature in law and economics on the evolution of the common law, for which Rubin (1977) and Priest (1997) provide seminal articles and Zywicki (2003) summarizes and gives historical detail. The common law is a special example of customary law, and in primitive societies, even more than modern ones, it can be difficult to distinguish between norms and laws—between what is enforced by guilty, esteem, and shame, and what is enforced by the power of the state. In medieval Europe the function of the government was not to make law, but to discover it, as Hayek (1973) discusses in Volume 1, Chapter 4 of his *Law, Legislation and Liberty*. It was natural for Hayek to precede his discussion of laws in that work with discussions of the evolution of norms. What his verbal discussion means and whether it is correct is controversial, as detailed in Whitman (1998), but the basic project is a sound one:

to examine how norms evolve and the extent to which group selection favors desirable norms.

We should emphasize, however, that norms need not always be preceded by conventions. For example, Pettit (1990) and McAdams (1997) claim that a new pattern of approval and disapproval can create a new behavioral regularity, given a desire for esteem. A norm arises when individuals desire esteem and these three conditions hold: there is a strong pattern of approval or disapproval for a given activity, there is a risk that others will detect one's engaging in the activity, and there is something approaching common knowledge of the approval pattern and risk of detection. Geisinger (2002) and McAdams (2000) claim that law can facilitate the process of norm emergence by publicizing the existence of a new consensus.

4. The importance of norms to legal analysis

In this section we describe how the existence and operation of norms affect the positive and normative economic analysis of law and legal institutions.

4.1. Positive analysis: how norms affect behavior

(i) Generally Norms matter to the positive economic analysis of law in two respects: in predicting how a change in legal rules affects behavior, and in explaining how law is made.

One cannot accurately predict behavior without knowing something about all the incentives that influence behavior—which includes normative incentives—as well as the way that legal change interacts with them. Economic analysis of law needs to consider carefully how norms may govern behavior in the absence of law and how a new legal rule may intentionally or unintentionally change (or fail to change) a norm.

Norms are, of course, highly diverse—as diverse in application as laws. Ellickson (1991, p. 132) has usefully categorized rules of any kind, including norms, into five groups. *Substantive* norms concern the conduct that is to be regulated in the first place, and the other four categories are ancillary. *Remedial* norms prescribe penalties or rewards for norm violation, *procedural* norms determine how information about violation is to be collected and used, *constitutive* norms govern how norms are created, and *controller-selecting* norms divide the labor of social control among different people.

Consider the norm of property in snowy-weather parking spots in Chicago described by Epstein (2002). The substantive norm says that only the person who dug the snow out of a spot is entitled to park there, whereas others who park will suffer guilt, disesteem, shame, or more concrete sanctions (Mayor Daley said, "I tell people, if someone spends all that time digging their car out, do not drive in that spot. This is Chicago. Fair warning."). Epstein does not describe the ancillary norms, but let us imagine what they might be. The remedial norm might be that someone who parks in the wrong spot will have his car window broken. The procedural norm might require that the enforcer make

some attempt to find and warn the violator before he resorts to violence. The constitutive norm might be that the norm can be changed only by explicit agreement of the residents of a street, and the controller-selecting norm might be that only the "owner" of a space is allowed to punish the violator.

Ignoring norms (or conventions) can cause one to overstate the significance of law, as suggested by the comments of Mayor Daley, the official ultimately in charge of enforcing both parking and vandalism laws for the City of Chicago. Norms matter in several ways. First, economists sometimes assume that a legal rule influences behavior, when an empirical investigation would show that the legal rule has no influence because group norms exclusively govern the behavior. Ellickson (1986, 1991) famously found that ranchers in Shasta County, California ignored legal rules concerning animal trespass and resolved disputes over cattle trespass damages according to "neighborly" norms, even though they had the legal right to go to court. Indeed, often one group norm is that members should never make use of their legal rights. For similar results concerning workplace norms (or conventions) and law, see Kim (1999) and Rock and Wachter (2002). Either can be strong enough to trump laws. Second, one might think a legal rule is a necessary condition of some observed behavioral regularity when a norm would maintain the same (or nearly the same) regularity without the law. For example, a norm that promises must be kept might, in identifiable circumstances, produce as much promise-keeping as legal liability, or at least enough so as to make the costs of legal enforcement no longer worthwhile (Macaulay, 1963; Scott, 2003). Third, one might overestimate the ability of legal change to produce a behavioral change by underestimating the degree to which the existing behavior is driven by norms (Kahan, 2000).

On the other hand, ignoring norms can also cause one to *under*state the significance of law. Economists sometimes assume that a legal rule is *not* necessary to change behavior when on closer analysis they would find that without new laws, norms will freeze the behavior in place. For example, market competition might not eliminate race discrimination if social norms require such discrimination [McAdams (1995)]. Moreover, changing a law might have a greater effect if legal sanctions work not just directly, by raising the price of a behavior, but indirectly, by changing norms. A new law might change perceptions of what incurs disapproval (McAdams, 2000), create a new basis for shame, or even change a person's own preferences and create guilt as Dau-Schmidt (1990) discusses in the context of criminal law. Kahan (1999, 2003) writes of the pervasive norm of "reciprocity," which he believes underlies much mutually productive cooperation in both small groups and society generally, but notes many ways that law can unintentionally undermine or intentionally facilitate such reciprocity. The extent to which the law actually does affect norms—and the ease with which such claims for new laws can be made—is an interesting question discussed in number of articles, e.g., Posner and Rasmusen (1999), Picker (1997), Hetcher (1999), Dharmapala and McAdams (2003); McAdams (2000); and Kahan (2000). Ellickson (2001) addresses the issue by comparing the ability of government and private "norm entrepreneurs" to change norms. Empirical work is harder to come by, but see Massell (1968) on law and change in Soviet Central Asia.

Positive analysis of law also seeks to explain a second point: why particular lawmaking institutions—the legislature, courts, or administrative agencies—create particular laws. Often one cannot fully explain the existence of a law without understanding the norms that give rise to it, or the absence of norms that would block it. Where public choice theory emphasizes the material interests citizens have in enacting or defeating legislation, attention to norms reveals that many people are highly motivated to create rules that do not affect their material interests. A person who believes that certain behaviors are immoral—e.g., pornography, abortion, flag burning, animal testing, or environmental exploitation—often favors laws forbidding or restricting such behavior. In turn, in a democratic system, such people's votes give the legislature incentive to enact laws supporting the norm. Or, if the political system gives him enough slack, the legislator, judge, or administrator may use his power to enforce the behavior he views as morally required.

Why would voters or lawmakers believe a law is necessary if the behavior is already enforced by guilt, disapproval or other normative incentives? An obvious reason is that the existence of a norm does not imply perfect compliance. Many people will occasionally face situations where the expected benefits of norm violation exceed the expected costs, and certain people may never obey the norm because they feel no guilt from violating it and can avoid detection. Another reason, more in keeping with public choice theory, is that even the norm, much less compliance, might not be universal. Different lawmakers will push to enforce the norms of the groups that support them, norms which come into conflict just as much as budget priorities, and often with more bitterness because of the normatized preferences of each group and the difficulty of compromise. It is hard to "split the difference" on abortion.

Still another motivation for laws as a supplement for norms is that the lawmaker may gain from purely symbolic endorsement of a norm, even if that endorsement is not expected to change behavior. There may be no observed flag-burning in a jurisdiction with strong patriotic norms, but voters may want to go further and express their disapproval by a symbolic declaration. Indeed, it is all the easier to pass such a law if nobody in the jurisdiction actually does want to burn a flag so no resources would have to be devoted to enforcement. Closely related is the function of laws as helping to create and perpetuate norms—one of the "expressive" functions of law discussed in Dharmapala and McAdams (2003), Geisinger (2002) and McAdams (2000). By saying what people *should* do, even if there is no penalty, the law tries to shift or maintain tastes, and to educate a society's newcomers—children and immigrants—in its norms. Law may serve the same function as the "rituals" that Cappel (2003) discusses, reinforcing attitudes by aiding communication of what is esteemed or by actually changing tastes by changing habits (on which see the experiments in Wells and Petty (1980): subjects instructed to nod their heads "Yes" repeatedly while listening to someone speak came to agree more with what the speaker said).

(ii) Specific norms regarding law Besides these general points, some specific law-related norms have particular relevance for positive legal analysis. The most important are the norms of "legal obedience"—that people should obey the law—and "the rule of law"—that laws should be knowable in advance rather than be the purely discretionary decision of some authority.

People often feel obliged to obey laws, or at least laws they perceive to be "legitimate," from the very fact that they are laws, rather than from any other motivation. These people suffer guilt, shame, or disapproval from breaking the law. The norm of legal obedience provides an incentive to obey the law that is independent of material sanctions (though if it is based on esteem and disapproval it still depends on violations being detected). This effect is particularly important for offenses that are *malum in prohibitum*—wrong only because illegal—because the prohibited act is not itself governed by a norm and the only relevant norm is legal obedience. One should not bring more than $10,000 in currency into the United States without declaring it on the customs form, but only because it is a legal wrong. By contrast, the norm against *malum in se* offenses such as murder is independent of its illegality. One should not kill unjustifiably, because that is a moral wrong—which also happens to be a legal wrong. [Other acts may be *malum in se* but not *malum in prohibitum*, e.g., adultery in the contemporary United States, as discussed in Rasmusen (2002) and Shavell (2002).]

Related to the norm of legal obedience is the ideal of "the rule of law." Defining this norm is difficult, but a central element is the idea that, as Fallon (1997, p. 3) puts it, "the law—and its meaning—must be fixed and publicly known in advance of application, so that those applying the law, as much as those to whom it is applied, can be bound by it." This norm constrains government officials who wield official power, a non-legal sanction against their illegal use of discretion or violation of rules. Thus, the rule of law is contrasted with "the rule of men." The norm of the rule of law is of great significance to how well laws work, since the alternative is costly, perhaps prohibitively costly, monitoring of executive and judicial officials. Development economics is by now quite conscious that it is not enough to establish good laws; one must, in addition, eliminate corruption and enforce laws fairly. See Rose-Ackerman (1999); Brooks (2003). We will not go into the large topic of jurisprudence, but merely note that it is an area in which the law-and-economics of norms might be usefully applied.

Other norms govern specific legal actors. To understand how a legal institution works, one must understand the norms governing that institution. For example, given how central the jury is to the legal system, it is odd how little attention economists have paid to the fact that jurors are paid by the day (and frequently less than their forgone wage) rather than based on the quality of their understanding or resolution of the case. Pettit (1990) notes that without normative motivations the successes of juries are puzzling, but that with such motivations we may explain why jurors (to some degree) pay attention to evidence, deliberate, and vote according to their evaluation of the evidence.

Similarly, other group norms besides the norm of the rule of law appear to be important—if not entirely effective—in constraining the behavior of judges, legislators, prosecutors, police, and other executive branch officials. A particularly interesting

set of law-related norms are those governing lawyers. Many of the ethical rules governing lawyers lack genuine sanctions and may be understood as efforts to strengthen or create professional norms [Painter, 2001; Wendel, 2001—though see Fischel (1998) for a more skeptical view]. There is some empirical evidence that norms even constrain the fees lawyers seek (Baker, 2001), though here there may be difficulty separating norms from convention or private rules. An interesting example is *Goldfarb v. Virginia State Bar*, 421 U.S. 773 (1975), concerning whether a bar association may expel as unethical members who charge a fee below a posted minimum. *Goldfarb* examines an industry group that seeks to regulate itself with professional norms, which shade into organizational rules and then into law.

(iii) Specific laws regarding norms Often one cannot understand the meaning of a specific legal rule without understanding the norms or conventions to which the rule explicitly or implicitly refers. At one extreme, law simply incorporates certain customs *in toto*. Cooter and Fikentscher (1998, p. 315) gives an example:

> When making Indian common law, tribal judges confront a central problem in legal anthropology: How to distinguish customary obligations that are enforceable at law (which can be called "common law") from customary obligations that are not enforceable at law (which can be called "mere customs")? Put succinctly, the problem is to distinguish "law from custom." If a custom is law, then legal officials are obligated to enforce it, whereas if custom is not law, then legal officials require an independent justification for enforcing it.

Obviously, in this case, one cannot know the content of the law without knowing the content of the custom (convention or norm) it enforces.

Norms are also used more narrowly, to flesh out statutes or judge-made law rather than to create laws out of whole cloth. Rather than fully specifying a substantive standard, many legal rules and doctrines "incorporate by reference" existing customs or practices, which in some contexts means norms and in other contexts means conventions. Legal definitions of obscenity explicitly incorporate local "community standards" [see *Jenkins v. Georgia* 418 US 153 (1974)]. Given the strong normative attitudes about the depiction of sex acts, the "standards" that the law incorporates are norms. Also, various torts—e.g., battery, invasion of privacy, and intentional infliction of emotional distress—include open-ended elements such as outrageousness or the absence of a customary privilege that implicitly incorporate norms. Both the crime and the tort of battery refer in part to the "offensive" touching of a person, which refers to norms. Other rules incorporate norms only indirectly or implicitly. Defamation law determines that certain statements are defamatory *per se* because they presumptively hurt the individual's reputation. What is defamatory *per se* is often the accusation of a norm violation—e.g., accusing a person of committing adultery. In recent years, changing attitudes towards homosexuality have made norms a subject of interest to courts trying to determine what is defamatory—see, for example, *Donovan v. Fiumara*, 114 N.C. App. 524 (1994). In many statutes, the crime of extortion, coercion, or blackmail includes

the threat to reveal any secret that will tend to expose the victim to "hatred, contempt or ridicule," which often includes the threat to reveal a norm violation. See Model Penal Code §223.4 (1985) (Theft by Extortion) and §212.5 (Criminal Coercion). A full understanding of the content of the law in these cases (and others) must correctly understand the content of a norm. A positive analysis of the consequences of the legal rule must also consider possible dynamic effects of incorporating the norm into the rule.

4.2. Normative analysis: how norms affect welfare

How does the normative analysis of law need to account for norms? Broadly speaking, there are two issues. First, how should welfare analysis incorporate the existence of norms and normative incentives? Second, when are norms efficient—or, more to the point, when are they preferable to law as a way to regulate behavior?

(i) Welfare analysis Norms change the welfare calculus in several ways. First, we must incorporate guilt, esteem, shame and pride into welfare via their direct effects on utility. Kaplow and Shavell (2001a, 2002b) examine how the normative incentives of guilt and pride (which they term "virtue") affect the welfare analysis of legal and moral rules. Ideally, there is a set of guilt and pride inclinations that ensures optimal behavior by each individual. Individuals sufficiently motivated would act optimally and therefore never have to incur guilt, which otherwise decreases welfare. But Kaplow and Shavell introduce reasonable constraints that complicate the analysis: that the process of inculcating guilt and pride is costly, that there is some psychological limit to the degree of guilt or pride individuals can feel, and that guilt or pride can be inculcated only for broad "natural" groupings of acts—such as lying—rather than for each particular act depending on its welfare effect—such as an inefficient lie. The result is a series of interesting tradeoffs between the use of guilt and pride and the optimal groupings of acts. Shavell (2002) then examines the optimal tradeoff between the use of these moral motivations and the legal system. The advantage of morality is that, compared to law, it is cheap and its internal incentives work without the external detection of anti-social acts. But the legal system can impose rules involving finer gradations in conduct than guilt-enforced morality, can change the rules more quickly than morality in response to changed circumstances, and can usually impose higher sanctions for the most tempting suboptimal acts.

Second, people who can feel guilt or pride in their own behavior will likely feel similar emotions as a consequence of observing the behavior of others, including that of government agents acting on their behalf. Thus, individuals may believe that certain legal outcomes are "fair" or "unfair," and thereby gain or lose utility from observing the outcomes. If so, the welfare analysis of legal rules must account for these effects on utility. For example, several theorists, such as Polinsky and Shavell (2000), Sunstein, Schkade, and Kahneman (2000), and Kaplow and Shavell (2002), consider the significance of the popular view that punishment should be proportionate to the crime. Those

holding this view may suffer disutility if maximal sanctions are imposed for non-serious offenses. Consequently, even if maximal sanctions (and minimal levels of detection) would otherwise be socially optimal, they might be suboptimal once we include the disutility of disproportionality. This argument is open to abuse, since it can be called in to defend any policy that some people favor, but that does not so much diminish its validity as call for empirical validation of claims that the utility effect is large enough to matter. Moreover, it provides one way to interpret the "retributive" function of punishment: observers feel utility when they observe misbehavior punished proportionately, and feel disutility when misbehavior receives disproportionately low punishment, including the extreme case of escaping punishment altogether.

Third, normative analysis must address the question of whether the social objective function should incorporate norms merely via their effect on the utilities of individuals or in addition to those utilities. If the objective function maximizes utility, it will take into account, for example, the distress that people feel at what they consider to be "unfair" outcomes, but the social planner might wish to reduce "unfairness" even beyond the effect on utilities. This double-counting might be legitimate, but the analyst should be aware of what he is doing, and double-counting necessarily means abandonment of Pareto optimality, an important argument against it. See Kaplow and Shavell (2001a, 2001b, 2002a). The logic of conventional welfare economics, with its criteria of efficiency or wealth maximization, requires instead that norms enter via their effects on utilities. Richard Zerbe's 2001 book, *Efficiency in Law and Economics*, is useful in clarifying this, and in showing how norms in utility functions can be operationalized by looking at a person's willingness to pay to have a norm obeyed.

All three of these welfare considerations treat norms as exogenous. We previously noted, however, that the law might in some cases influence norms, which can involve (when guilt inclinations are changed) a change in tastes. The welfare analyst must therefore decide how to deal with the possibility of preference change. Economists since Strotz (1955) have studied the problem of how to do welfare analysis when tastes are variable or when people are poorly informed. For example, legal rules against race discrimination might initially generate direct utility for people who regard such rules as fair and disutility for people who regard them as unfair. But if the rules produce a change in preferences over time, diminishing the internalized norm of discrimination, static analysis will be misleading [Kaplow and Shavell (2002a, 2002b)]. If the analyst knows that an anti-discrimination law will lead an individual to change his preferences and actually prefer the law after five years, should the individual's present preferences trump his future preferences? Should this be the case even if the individual knows how he will change? Such questions have been much discussed in various contexts; see Dau-Schmidt (1990), Kuran (1995); and Ng (1999).

Cooter (1998) links norms to the concept of a "Pareto self-improvement": an individual who perceives the advantage of having different preferences, even from the vantage point of his existing preferences, may work to change his preferences. If people are poorly informed, however, there can be a conflict between maximizing their utility ex ante—making the choices that they, with their poor information, would make—

and ex post—the choices they would have made if well informed. Richard Posner (1992) applies this idea in his *Sex and Reason*, in which he suggests that norms against sexual practices such as homosexuality would disappear if their holders had better information. The big practical problem, of course, is determining whose information is wrong, since each side may well believe that the other's beliefs are sincere but misguided.

Welfare analysis of preference change is particularly complex in the case of interdependent utility functions. The norm of retribution, for example, may be supported by preferences in which one derives utility from the disutility of another—the offender [see Kahan (1998)]. By contrast, altruism may underlie norm of gift-giving, which, as Kaplow (1995) explains, increases the utility both of the individual holding the norm and of others on whom he acts. John Stuart Mill (1859) is hostile to what he calls "other-regarding" preferences in *On Liberty*, though as has been pointed out in Kaplow and Shavell (2002a, 2002b) and James Fitzjames Stephen's *Liberty Equality Fraternity* (1873), his own tone is highly moralistic, and sufficiently obscure that it is hard to make sense of how he decides which nonmaterialistic preferences are legitimate and which are not.

(ii) Norms versus law The second broad issue is whether norms, or certain identifiable classes of norms, are *generally* efficient or inefficient. This matters to whether the coverage of law should be expanded or contracted.

Norms have the obvious advantage of low transactions costs compared to law. They do not require police, courts, collection agencies, or prisons. If they are fully internal, they do not even require detection. Thus, they seem particularly appropriate for regulating externalities too small to justify appeal to the courts, or for those whose detection and proof are particularly difficult. On the other hand, norms are trickier to create than laws, and are not typically the subject of policy discussion. Rather, the usual question is whether society should create laws to supplant norms.

Legal authorities will often wish to defer generally (rather than case by case) to norms in domains where norms are more efficient regulators of human conduct than legal rules. Thus, one needs to compare the efficiency of legal rules to decentralized norms. Eric Posner (1996a, 1996b) examines the case for deferring to norms in groups governed by them. Legal regulations intended to protect individuals may have the unintended consequence of lowering the value of group membership, thus weakening the power of groups to enforce their norms. Thus, the efficient legal rule might be one of noninterference. Shavell (2002) makes a general comparison of the comparative advantages of law and morality, where morality includes both internalized and non-internalized norms.

Whether norms are generally efficient, or even efficient in identifiable circumstances, is contested. Everyone acknowledge the existence of dysfunctional norms, but Ellickson and Cooter, to take two major figures in the literature, are optimistic about the efficiency of group norms that affect only the members of the group, viewing norms as

mechanisms for deterring behavior with negative externalities and encouraging behavior with positive externalities. Both are pessimistic about norms between groups, where there is no incentive to account for the external effects. See also McAdams (1995) on the stability and inefficiency of norms of racial discrimination.

Others are more pessimistic about norms, even when they apply only to the group in which they arise. Several theorists make the general point that law is often superior to norms or conventions. Kaplow and Shavell (2002a, 2002b) argue that norms of fairness usually enhance social welfare by curbing self-interested behavior with negative externalities, but nonetheless conclude that norms are in many particular cases inferior to regulation by the optimal legal rule. They emphasize two disadvantages of norms that arise because norms are frequently inculcated in children and supported by feelings of guilt. First, the norm is often simpler than the optimal rule (e.g., never break a promise, rather than never inefficiently break a promise). Second, the norm is hard to change when new conditions make a different rule optimal. Kahan (2000) also emphasizes the stickiness of obsolete norms. McAdams (1997) raises the possibility that groups will enforce "nosy" norms that regulate behavior with only mild externalities. Norms may demand conformity to the other-regarding tastes of the majority even when the minority loses much more by frustration of its self-regarding preferences than the majority gains (e.g., regarding mate-selection criteria).

Similar points apply to conventions. Eric Posner (1996b, 2000a, 2000b) identifies various problems arising from poor information or strategic behavior that can make conventions and norms inefficient, justifying a corrective or supplementary legal rule. Mahoney and Sanchirico (2001) use evolutionary game theory to explain how the fittest convention in a given environment often deviates from the efficient one. See also Horne (2001) and Kübler (2001). In the case of either norms or conventions, of course, we must keep in mind that just because norms are inefficient does not mean laws would be efficient, any more than market failure in standard economic markets means that government regulation would be optimal rather giving rise to government failure instead. The issue is which is the greater danger, the purposeless inefficiency of norms or the purposeful inefficiency of law.

A second possibility is that efficient norms are "fragile" and therefore require not just non-interference but affirmative legal protection. This might justify otherwise puzzling rules of market-inalienability. Why does the law constrain the sale of parental rights and child labor? One possibility is that parental norms are a more efficient regulator of parenting practices than is law, but that normative incentives are weak compared to market incentives and that parenting norms would unravel if parents were fully subject to market incentives. An unregulated market would, on this view, leave parents to make individually maximizing but socially inefficient decisions about their children, such as the choice to curtail their education in order to exploit their short-run potential in the labor market. Rather than overcome this problem by directly regulating the precise boundaries of parental conduct, however, one might instead enact rules of market-inalienability, constraining the operation of market incentives on parents, and leaving them subject to *relatively* more powerful normative incentives. This idea is distinct from but related

to the idea that monetary incentives might "crowd out" non-monetary incentives (Frey, 1994).

5. Specific applications

5.1. Tort law

Kaplow and Shavell (2002a, pp. 134–143) note that there is a strong norm to avoid injuring others and to compensate them for injuries one does cause. See also Smith (1790, p. 104). Consistent with their general thesis, they claim that these norms generally improve social welfare but that law can make additional gains. Tort laws and litigation processes have certain advantages over norms in collecting the relevant information, imposing more optimal rule complexity, changing rules more quickly in response to changed conditions, and imposing greater sanctions. Compensation norms, however, also encourage law-makers to compensate via tort law, and may therefore hinder reliance on an insurance regime when it is a better means of compensating accident victims.

Custom has long played a key role in tort law because it helps to decide whether an injurer was negligent. Since negligence is closely allied to failure to fulfill an obligation, negligence is, in the terminology of this chapter, the violation of a norm. As Hetcher (1999) explains, two possible uses of custom are as a *per se* rule, under which adherence to custom is a complete defense [exemplified by the leading case of *Titus v. Bradford*, 20 A. 517 (Pa. 1890)], and as an "evidentiary rule," under which adherence to custom is only evidence about what is non-negligent. Justice Holmes preferred this second use, and said in *Texas and Pacific Railway Company v. Behymer*, 189 U.S. 468 (1903), "What usually is done may be evidence of what ought to be done, but what ought to be done is fixed by a standard of reasonable prudence, whether it usually is complied with or not." Judge Learned Hand's opinion in *The T.J. Hooper*, 60 F2d 737 (2d Cir. 1932), is similar, adding the idea that norms may become outdated because of technological advances—the invention of the radio in that case [on which see Epstein (1992b)]. Hetcher, however, argues against the modern preference for the evidentiary rule using the idea of the coordination game.

Norms are also important to what damages are awarded in tort. Cooter and Porat (2001) address the question of whether legal damages ought to be adjusted for the normative penalties that an injurer has paid for his misdeed. Cooter (1997) argues that norms are central to whether punitive damages are awarded, though not useful to deciding their magnitude.

5.2. Contracts and commercial law

Kaplow and Shavell (2002a, pp. 203–213) review some evidence for the existence of a strong norm of promise-keeping, supported by guilt. Macaulay (1963) first documented

that businesses do not rely on exclusively or even primarily on law to enforce agreements. They use norms, which may most simply be reduced to two: "(1) Commitments are to be honored in almost all situations; one does not welsh on a deal; (2) One ought to produce a good product and stand behind it" (p. 63). Cooter and Landa (1984) find that ethnically homogenous minority groups often dominate certain "middleman" positions in the markets of nations in which judicial enforcement of contracts is weak or nonexistent. They claim that ethnic ties create informal enforcement mechanisms (norms or conventions of promise-keeping) that substitute for state enforcement of contracts. See also Landa (1981, 1994) and Davis, Trebilcock, and Heys (2001). In a series of meticulous studies, Bernstein (1992, 1996, 2001) finds that many merchants groups prefer to enforce contracts, when disputes arise, through private trade association mechanisms, rather than rely on the state; see also Richman (2004) on Jewish diamond merchants in New York.

In cases of informal enforcement, norms do nicely in handling clear cases, but problems do arise because it is hard for a norm to sharply define when one "honors a deal" or "produces a good product." Kaplow and Shallow (2002a, 2002b) suggest that although promise-keeping norms improve social welfare by making certain trades uniquely possible, often the optimal legal rule would do even better. In particular, norms are often too simple and their penalties are too weak to deter misbehavior when the stakes become high.

Much of what appears to be norms in business may be driven by the convention of the infinitely repeated prisoner's dilemma. A particular important variant of this convention is reputation, as laid out in the classic article of Klein and Leffler (1981). We will recast their idea here as the formal model used in Chapter 5 of Rasmusen (2006). Suppose we have sellers who can produce either a high-quality good at cost c or a low-quality good at cost 0. Buyers all value the high-quality good at some amount much greater than c, and the low-quality good at 0, but cannot tell quality until after they have bought the product. The players make choices each period of what quality to choose, what price, p, to charge, and whether to buy or not, with a small discount rate of r between periods and no end period.

If there were only one repetition, the unique equilibrium would be for the sellers to produce low quality and for the buyers to decide not to buy. In the repeated game, low quality will remain the equilibrium outcome if expectations are pessimistic, but if the buyers believe that a given seller is reputable and will produce high quality, that too can be a self-fulfilling expectation.

In the high-quality equilibrium, a seller is willing to produce high quality because it can then sell at a high price for many periods, but if it produces low quality, the one-time savings in production costs is offset by the loss of future returns. Thus, an essential part of the model is that the equilibrium price be well above marginal cost. For the seller to produce high quality, its one-time gain from cheating and producing low quality—the revenue of p^*—must be no greater than the present discounted value of the alternative long-term profit of $(p^* - c)$ each year, a value equal to $(p^* - c)/r$. This requires that $p^* = (1+r)c$. Any seller selling at a price higher than p^* would be undercut by a seller

that sold at p^*, but any seller selling at a price less than p^* would get no customers, since buyers would realize that such a low price does not provide enough incentive to keep quality high. Buyers rightly do not trust a seller who is not charging a high enough price. The reputable sellers then make positive profits because undercutting each other drives away customers.

Thus, a convention—that quality be high and any deviant firm be punished by future boycott—results in high quality despite the lack of immediate observability or enforcement by laws. The model can be applied to many situations of good behavior seemingly enforced by norms. In particular, under the name of "efficiency wages" it can explain high wages and honest behavior of employees in industries where trust is important [see Akerlof and Yellen (1986)].

Custom has played an important role in contract law at least since the time of Lord Mansfield. The Uniform Commercial Code says in Section 1-201(3) that an agreement is to be interpreted "by implication from other circumstances including course of dealing or usage of trade or course of performance" and numerous other sections of the UCC, listed in Bernstein (1999, note 1), follow this "incorporation strategy." Custom here is more important as a convention than as a norm, and often its role really is just as an aid in interpreting a contract term; the use of custom as evidence of meaning is uncontroversial. Customs may become normatized, however, in a relatively simple way: someone who violates a norm is considered to be cheating in the same way as someone who lies or breaks a law, and the victim of the violation feels a visceral response.

Bernstein (1996) makes a useful distinction between "relationship-preserving" norms, which apply to continuing good relationships between businesses, and "end-game" norms, which apply to distressed relationships that are winding down. She notes that courts should not fall into the mistake of using relationship-preserving norms as a guide to how businesses would wish their affairs to be conducted once the relationship is ending, and uses arbitrations of the National Grain and Feed Association as an illustration. Those arbitrations are heavily formalistic, she shows, in contrast to the UCC approach, and do not consider elements such as "good faith" in making decisions.

Naturally, norms vary from industry to industry, making this fertile ground for empirical study. Drahozal (2000), for example, argues that international commercial arbitration is more like the UCC than the arbitrations studied by Bernstein. A leading case in international commercial arbitration, *Pabalk Ticaret Limited Sirketi v. Norsolor S.A., ICC Award* of Oct. 26, 1979, No. 3131, 9 Y.B. Commercial Arb. 109 (1984), applied "the international lex mercatoria," which the court said included a principle of "good faith which must preside over the formation and performance of contracts," so that a party would be liable because of its faithless conduct. A requirement of "good faith" sounds like and may reflect moral obligations—norms—rather than mere convention. This example illustrates a problem running through the literatures both of norms and conventions: these rules are no more likely than laws to be universal, so case studies, requiring considerable study to yield single scholarly papers, are often more useful than either statistical or theoretical articles.

5.3. Corporate law

Corporate law has been the object of a surprising amount of scholarship on norms, considering that the corporation itself has no shame, guilt, or appreciation of esteem. One of the best articles is Skeel (2001), which focuses on three examples: the California state pension fund's list of firms with poor governance, the *Business Week* and *Fortune* lists of America's worst boards of directors, and Robert Monks's battle to shame the board of Sears into changing the company's policies. Skeel's emphasis in all of these is how norms affect the individuals who govern a corporation, and he makes a persuasive case that norms do change their behavior. This is an effective counter to the skepticism of Kahan (2001), who wonders whether the idea of norms has much to add to corporate law unless "norms" is defined to include ideas already used from game theory and other sources. Norms may not be able to explain why corporate law takes the form that it does, or how corporate law should be shaped, but it may be very helpful in explaining how corporations behave within a given framework.

5.4. Property and intellectual property law

Norms are sometimes considered to be the origin of property, since property can exist in the absence of government, but in our terminology the origin is a convention. Some property rules can be modeled as simple coordination games like that in Table 1 above, for example, the decision of people not to fruitlessly try to use the same land or airwaves at once. In other cases they are in the category of coordination games variously known as Hawk-Dove, Chicken, or the Battle of the Sexes, in which there are two or more equilibrium conventions, but players differ in which convention they prefer. In the Hawk-Dove game, two identical players, Row and Column, contend over a valuable good. Each player chooses whether to be an aggressive "Hawk" and fight for the resource if necessary, or a timid "Dove" who retreats rather than fights for the good. As Table 3 shows, the best outcome for a player is if he plays Hawk and the other player selects Dove, and the worst is if both players choose to play Hawk, which results in a destructive fight.

The two pure-strategy equilibria are (*Hawk, Dove*) and (*Dove, Hawk*). These are asymmetric equilibria, and the question naturally arises of how the players know that the convention is, for example, that Row gets to be the *Hawk* and Column plays *Dove*. Without further information, the model cannot answer that question. There does exist a symmetric equilibrium, but it is in mixed strategies. Each player chooses *Hawk* with probability $2/3$ (yielding an expected payoff of $(2/3)(-1)+(1/3)(3) = (1/3)$ and *Dove* with probability $1/3$ (yielding an expected payoff of $(2/3)(0) + (1/3)(1) = (1/3)$. The payoffs in the mixed strategy equilibrium are $(1/3, 1/3)$, below the average payoff of $2/3$ that arises if conventions are used.

Maynard-Smith and Parker (1976) noted that both players would prefer a pure-strategy equilibrium if somehow they could take turns playing *Hawk* and getting the high payoff. One way to take turns is to expand the model and establish a convention

Table 3
Hawk-Dove

		Column	
		Hawk-Aggressive	Dove-Timid
Row	Hawk-Aggressive	−1, −1	3,0
	Dove-Timid	0, 3	1,1

Payoffs to: (Row, Column).

that the player who arrives first and claims the valuable good plays *Hawk* and the second player to arrive plays *Dove*, which is known as "The Bourgeois Strategy." This is a symmetric equilibrium in the expanded game, and has higher expected payoff than the mixed strategy. Such behavior looks very much like a norm of ownership rights for the first-possessor, but in our terminology it is a convention, being based solely on shared expectations of who will fight and who will flee rather than on any notions of right and wrong independent of material consequences.

Once the convention emerges, it is easy to imagine how it becomes a norm. Among other mechanisms, parents instructing their children to respect the convention—do not take goods first possessed by others and do not let others take goods you first possessed—would intentionally or unintentionally instill a sense that such unconventional takings were unfair and wrong. In any event, today law and norms of property are largely co-extensive. Norms against theft and "misuse" support legal property rights, and vice versa. Miller (2003), for example, examines the internalization of the legal norm against parking in "handicapped spaces."

There are, however, exceptions: norms that constrain the exercise of property rights and norms of "property" that are unsupported by law. First, consider how norms sometimes oppose the exercise of legal property rights. A scholar who sues another scholar for infringing his copyright by photocopying his book chapter may face social penalties. Though orchard owners have the legal right to grow apple trees without keeping bees on their property to pollinate them, and to free-ride off the bees kept on neighboring orchards, Cheung (1973) showed that Washington state orchard owners followed a norm of keeping bees proportionate to their orchard size. Ellickson (1991) famously showed how local Shasta County, California norms governed relations between neighboring ranchers to the exclusion of law. Thus, even when legal rules governing animal trespass damages or the maintenance of boundary fences create certain legal rights, an individual would forgo those rights and follow the norm.

Norms constrain the use of public as well as private property. Ellickson (1996, p. 1172) thinks of urban problems using the paradigm of "a public space as an open-access territory where users are prone to create negative externalities." These problems have traditionally been regulated in large part by unwritten rules—either unwritten (or vaguely written) rules enforced entirely according to the discretion of local officials—

vagrancy laws, for example—or norms. There is a norm, for example, that a person should not make excessive use of a nonpriced good such as a park bench. Someone who spends the entire day on the park bench with the best view of the White House, whether a vagrant or a journalist, is violating that norm, which establishes a temporal limit to the right to use public property. Ellickson also discusses how norms establish informal zoning for behavior. A typical city dweller drastically changes what he considers bad behavior, worthy of reprimand, depending on where the behavior occurs. He may heap abuse—or at least raise his eyebrows noticeably—on someone inebriated in a residential neighborhood while tolerating much worse inebriation in Skid Row or the Red Light District.

Second, consider how norms sometimes create quasi-property rights, not recognized by law. A scholar who uses someone else's idea without attribution may avoid violating copyright but may be punished by norms against plagiarism. See Green (2002). We have already mentioned that Epstein (2002) discusses the importance of norms in establishing informal property rights in parking spaces. Sometimes, it is because conventions or norms recognize a quasi-property right that a court will give the right legal recognition outside of statutory law, as when the U.S. Supreme Court recognized a property right in news in *International News Service v. Associated Press*, 248 U.S. 215 (1918), a case analyzed in Epstein (1992a).

Intellectual property presents its own interesting set of issues. Software has proven to be an interesting industry for the study of norms, perhaps because the Internet is new and important enough to have stimulated the creation of new norms. Strahilevitz (2003) focuses on the role of optimistic lies ("charismatic code") in establishing norms. Gnutella was a network that allowed members to share computer files. It told them, "Almost everyone on Gnutella.Net *shares* their stuff," which is false—only one-third of users shared. Also, networks like Gnutella "by no means cede the moral high ground," despite the dubiousness of their interpretation of the copyright laws. Rather, they try to create norms of cooperation, starting from the general norm of reciprocity. They call people who download files without making their own files available for upload "freeloaders," even though the record companies might use the same term against Gnutella.

A number of scholars, including McGowan (2001); Benkler (2002) and Lerner and Tirole (2002), have examined the puzzle of why people create open-source software—software that is given away for free, yet is costly to produce. Benkler discusses enterprises such as the Linux operating system, Napster music distribution, and Project Gutenberg—electronic texts that rely on thousands of volunteers to contribute their effort towards a public good. He makes the important point that by dividing the effort into small parts, these enterprises can make do with a small amount of norms; no one person must make a large sacrifice, and each participant can feel satisfaction at having aided the common good.

5.5. Criminal law

Criminal law is intimately linked with norms, as one might expect from the fundamental idea of *malum in se* versus *malum prohibitum*. That a crime is *malum in se*—wrong in itself—means that the law prohibits something that also violates a norm, such as theft and unjustified killing. Thus, norms and criminal law may reinforce each other by sanctioning the same conduct. Even if the crime is *malum prohibitum*—wrong only because illegal—the norm of obeying the law may generate some compliance above that predicted by the expected sanction. The level of this effect, however, may depend on the law's or law-maker's perceived "legitimacy" [see Thibaut and Walker (1975); Robinson and Darley (1997); Kaplow and Shavell (2002a, p. 370)] both because people are more likely to obey a legitimate rule and more likely to cooperate with police in apprehending those who violate legitimate rules. Norms may also help us understand particular crimes. McAdams (1996), for example, claims that norms of privacy help to make the prohibition on blackmail efficient.

Criminal law is the most common outlet for "expressive law": laws that are meant to express disapproval more than to actually punish it [see Dharmapala and McAdams (2003), Kahan (1998), McAdams (2000), Sunstein (1996)]. It might be a crime to commit adultery or to disrespect one's parents, but the law's purpose and effect may be more to express disapproval as to detect and punish these hard-to-prove norm violations. Prosecutorial discretion results in the laws remaining purely expressive; if private law were used, the courts would have to deal with messy civil lawsuits induced by the financial incentives.

The normative economic analysis of criminal law focuses on optimal punishment. Norms are relevant here as well. There is a strong norm of retribution (Kaplow and Shavell (2002a, pp. 352–359), which means that there is a taste for punishing wrongdoers and also a taste for the punishment being proportionate to the crime. The presence of this taste has direct and indirect effects on optimal punishment. The direct effect is that punishment does not only deter and incapacitate, but satisfies or dissatisfies tastes for proportionate retribution (which may also in turn affect perceived legitimacy). Thus, where optimal punishment might otherwise be low (as where the risk of detection is high), the norm of retribution may require that it be higher. Conversely, where optimal punishment might otherwise be high (as where the risk of detection is low), the norm of proportionality—that the punishment "fit" the crime—might require that it be lower. See Polinsky and Shavell (2000).

The indirect effect is that the norm of retribution means that some individuals will punish a wrongdoer privately, which affects the optimal level of legal punishment. Optimal deterrence, for example, depends on total sanctions for wrongdoing, not just governmental sanctions. Cooter and Porat (2001) argue that when a tort or contract breach also constitutes a norm violation, it is generally advisable to deduct from tort or civil liability the amount of private sanctions the wrongdoer incurs (and even, in theory, to deduct certain external benefits norms violations create, as where business is diverted to one's competitors). The same general point may be made about criminal liability.

Stigma is an important punishment for wrongdoing, one that combines both public and private punishers. As Rasmusen (1996) explains, the court's official declaration that someone has committed criminal actions can be important even if there is no material punishment by the state, because the information thereby transmitted makes private actors behave differently towards the criminal. A controversial current application of this idea detailed by Teichman (2005) is in "Megan's Law" statutes, which publicize the identity and living location of sex offenders. Kahan (1996) uses the differing ability of public sanctions to condemn and stigmatize to explain political support for or opposition to alternative sanctions.

Another effect of non-legal sanctions arises on the issue of optimal sanctions for repeat offenders. In addition to other reasons, Dana (2001) justifies higher legal sanctions for repeat offenders on the ground that the probability of incurring non-legal sanctions declines with the number of violations (because non-legal sanctions such as boycotting are often exhausted after the first or second violation or because the repetition of offenses signals the fact that the offender is not a member of a community that will subject him to non-legal sanctions for such violations). To maintain the same total sanction, it would then be necessary to raise the legal sanction.

5.6. Discrimination and equality law

Discrimination law is similar to morals law in being closely entangled with norms specific to a time and place. Norms and conventions often govern behavior according to the social groups to which someone belongs, particularly sex, race, ethnicity, or religion. In various times and places, norms or conventions defining sex roles have allocated some jobs exclusively to men and others exclusively to women (Hadfield, 1999); compelled women to take their husband's surname upon marriage and stay at home to rear children; and differentiated the sexes by dress. Racial, ethnic, and religious group norms often require that members of a group adhere to distinctive codes of dress or food consumption that publicly identify group membership or loyalty. See Kuran (1998). Other norms or conventions compel group members to "discriminate" against non-members, as by prohibitions or limitations on economic transactions or marriages outside the group, the refusal to accord non-group members customary signs of respect, or even the use of violence to suppress non-group members in competition for scarce resources or governmental control. See McAdams (1995) and E. Posner (1996b).

A number of scholars have discussed such norms. Akerlof (1980, 1985) claims that "customs" of caste may survive market competition because third parties punish those who violate the custom, though he doesn't explain why third parties willingly incur such enforcement costs. McAdams (1995) offers to explain third party enforcement by the "payment" of esteem or status. Discrimination arises as groups compete for social status and individual members are rewarded with intra-group status for contributing their group's societal status by discriminating against others or by punishing non-discriminators. E. Posner (2000b) instead describes race discrimination as a convention (in our terms) that emerges from a signalling game. In his model, individuals

incur costs to conform to the convention to signal their low discount rates. These authors view discriminatory norms as socially costly (e.g., E. Posner, 2000a, pp. 1722–1723). Even those who are relatively optimistic about group norms have predicted efficiency only where norms primarily affect group members, and have expressed pessimism about the external effects of norms on non-group members—see Ellickson (1991, p. 169) and Cooter (1996, pp. 1684–1685). Indeed, Kuran (1998) raises the concern of sudden "cascades" in the level of ethnic identification—which he calls "ethnification"—a process that can lead to violence.

In American history, and in other societies today, law has been used to reinforce such norms. In the Jim Crow era of the American South, state and local laws mandated segregation of certain types of public transportation, barred racial minorities from attending certain schools, and prohibited racial intermarriage. More recently, laws seek to suppress and undermine discriminatory norms. One step has been to interpret federal constitutional law to invalidate state law requiring discrimination, as in the *Brown v. Board of Education*, 347 U.S. 483 (1954) invalidation of formal racial segregation in public schools, and to prohibit other official action based on discriminatory motives. Similarly, McAdams (2000) argues that the First Amendment's constitutional prohibition on the state "establish[ing]" a religion might be understood as an effort to prevent cascades toward extreme religious conformity. As a second step, federal law now prohibits private discrimination on grounds of sex, race, ethnicity, religion and other such factors, in employment, housing, lending, public accommodations, and other such domains. A third step has been to use law to permit private individuals to favor previously disfavored minorities through "affirmative action," or to require government agents to do so. A major part of the debate over affirmative action is whether it works ultimately to promote or undermine discriminatory norms.

5.7. Family law

Family law is saturated with the influence of norms (see, e.g. E. Posner, 1999). Indeed, it is separate from contract law largely because of a longstanding belief that social norms are crucial to how families will be allowed to make use of the courts—that, unlike in commercial contracts, courts ought not to enforce all marital agreements. Instead, the courts should allow social norms to regulate behavior within the family, even behavior that between strangers might be grounds for suit. The motivation was not only to keep courts out of a sphere in which they could not make well-informed decisions, but also to prevent government from aiding agreements in violation of social norms or from intervening in ways that, as Stephen (1873) argued, would weaken marriage norms. He claimed that law could not govern families as well as norms but could have the unintended consequence of damaging norms. As Rasmusen and Stake (1998) note, the difficulty of customizing legally enforceable marriage agreements has remained, however, even as social norms have weakened and the default definition of marriage has departed radically from the traditional idea of dissolution only for fault.

While there has been attention to economic models of the family in law-and-economics, there has been less attention to norms. One exception is the article by

Elizabeth Scott and Robert Scott (1995), "Parents as fiduciaries," which analyzes the legal role of parents as closer to that of fiduciaries such as trustees who act for beneficiaries than of agents who act for principals. A fiduciary incurs legal liability as well as any norm-based penalty for violations of his duty, but norms enter in defining that duty, an example of the "incorporation by reference" that we discussed earlier. Robert Ellickson's 2005 "Norms of the household" takes a different approach, focusing not on the family, but on the related situation where more than one person lives in the same residence with the possibility of exit. A "household" is different from a family not only by including mere roommates, but also by excluding traditional marriages (from which exit was difficult) and single-parent families (because children cannot exit). Ellickson argues that consensus is a desirable method to make decisions in such an organization.

Another example, which shows the possibilities for empirical work in this area, is Margaret Brinig and Steven Nock's 2003 article, " 'I only want trust': norms, trust, and autonomy." Brinig and Nock examine data on the mental health and other characteristics of divorced couples. They find that marriages break up after a collapse in trust, which might be a failure of either a convention or a norm. Also, divorced men who fail to gain any custody of their children have a significant increase in depression, although remarriage reduces the amount of depression. Brinig and Nock suggest that the depression might arise as a result of punitive social norms triggered by the disgrace of losing custody, but their data does not permit them to test this against the simpler theory that the men miss the company of their children.

5.8. Other public law

We have discussed criminal law, discrimination law, and family law, but these are only three of many areas in which the government regulates behavior. Here we briefly discuss the relevance of norms to tax compliance, environmental compliance, driving behavior, and voting.

Norms are important to understanding tax compliance. E. Posner (2000a) began the discussion of whether strict enforcement of tax laws is a substitute or a complement for norms of tax-paying. Lederman (2003) argues that stricter enforcement of tax laws is actually complementary to norms of legal obedience. Enforcement will increase the number of people who obey the laws for prudential reasons and creating this "critical mass" of taxpayers will create disutility to others if they violate the law. Evidence for her argument is the experiment conducted by the State of Minnesota, which sent a sample of potential taxpayers a letter telling them, truthfully, that most citizens do pay their state income tax. People who received the letter paid more taxes than those who did not. See also Kahan (2002); Murphy (2004). Kirsch (2004) critiques the use of shaming sanctions and "norm management" as an alternative to traditional penalties for tax avoidance, concluding that the problems of such an approach justify only a narrow use of such sanctions.

There is also some literature on how norms matter to compliance with environmental regulations. Vandenbergh (2003) identifies norms that influence corporate environmen-

tal compliance. He discusses the empirical evidence for the existence of several relevant norms, including the substantive norms of law compliance, human health protection, environmental protection, and autonomy. He explores the implications of these norms, concluding that future environmental enforcement policies should strive to harness them or at least to avoid undermining them. Carlson (2001) looks at the effort by local governments in the United States to influence individual behavior by inventing new norms of recycling. She claims that the most important policies are those that reduce the cost of recycling rather than those that try to change people's preferences, make signalling more effective, or direct esteem towards those who recycle. The question remains, however, why anyone bothers to incur positive costs to recycle. Carlson concludes that signalling and the desire for esteem are important, even though the cost of recycling can easily swamp their effect.

Traffic laws are another fertile area. Sugden (1986; 1998) frequently uses a simple traffic conflict—the "Crossroads Game"—to illustrate his theories of the evolution of conventions and norms. Strahilevitz (2000) provides a case study of traffic compliance that explores the effect of "commodification" on norms, a matter that has worried some theorists. He studies San Diego's FasTrak carpool lane program. Under FasTrak, drivers could either carpool to be allowed to drive legally in special fast lanes for free or pay a price to drive in the fast lane without carpooling. Establishing a price for the fast lanes for non-carpoolers actually increased the amount of carpooling. In addition, the price was paid by many non-carpoolers who had before been violating the rules by driving alone in the fast lane. Strahilevitz suggests that this is because "[t]he commodification of the road makes other drivers less sympathetic to cheaters. The Express Lanes violator is transformed from a rebel into a scofflaw" (p. 1231).

Finally, voting has posed a challenge for rational choice theories because the expected benefits from influencing an election are so small compared to the costs of voting. Hasen (1996) offers norms as a possible explanation: as with other types of socially beneficial behavior, in some communities individuals receive small social rewards for voting or small sanctions for not voting. In this sense, voting rates may reflect the degree to which a community more generally succeeds in encouraging privately costly but socially beneficial behavior among its members. Because American society has relatively low levels of voting, Hasen explores the possibility of creating a legal duty to vote and supplementing the informal incentives with legal sanctions, as a few European nations do.

5.9. Constitutional law

A constitution cannot itself be based on law, since law is only established by the constitution, a meta-law. Thus, compliance with the constitution must be based on norms or convention. A variety of scholars, including Hardin (1991) and Buckley and Rasmusen (2000), discuss constitutions—written and unwritten—as particular equilibria of coordination games and consider how norms may help support such constitutions. Writing down the constitution, as in the United States, helps to establish the equilibrium, but

the equilibrium does not require writing, as the British Constitution shows. Even where there is a writing, it may be irrelevant to how things actually work [see also Ordeshook (2002)]. In addition, there is a normative element to constitutions. Certain government action is thought to be wrong and called "unconstitutional," a pejorative term used not just in the United States, but in Britain, where there is no Supreme Court to officially label behavior as unconstitutional. At least one written constitution—the 1793 Constitution of France—enshrines the norm of rebellion: "When the government violates the rights of the people, insurrection is for the people, and for each part of the people, the most sacred of rights and the most indispensable of duties" ("Declaration of the Rights of Man and Citizen," Article 35—a provision prudently omitted from the Constitution of 1795).

5.10. International law

International law is a natural setting in which to expect norms or conventions to be important, because there is no authority above nations to enforce the rules by which they behave. Whether norms actually arise is an open question. Goodman and Jinks (2004) argue that international law can influence states by "acculturation" of state actors. Goldsmith and E. Posner (1999) offer a more skeptical view. They argue (at p. 1132) that nations generally "do not act in accordance with a norm that they feel obligated to follow; they act because it is in their interest to do so." Frequently, nations will view it in their interest to comply with their international treaties, which essentially define the parameters of cooperation and defection in an iterated prisoners' dilemma model. See also Guzman (2002). By contrast, Goldsmith and Posner claim that the non-treaty obligations international lawyers term "customary law" are really no more than the description of what states have found it in their interest to do in the past, which does not even state a convention governing future behavior. Ginsburg and McAdams (2004) also emphasize a state's self-interest but envision a slightly larger role for international law. They contend that international adjudication generates compliance not because of a norm of complying with legitimate authority but because the adjudication signals disputed facts and clarifies disputed conventions, and in each case it is then often in the parties' interest to comply.

Under any of the latter three accounts—Goldsmith and E. Posner (1999), Guzman (2002), and Ginsburg and McAdams (2004)—international law works to the extent it does because it is a convention—under the terminology of this chapter—and not a norm. This is a useful conclusion, if true. It suggests that the existing successes of international law (e.g., rules on diplomatic immunity and on the treatment of neutral shipping during war) have been achieved without internalization of incentives and raises the question of whether international law can achieve further success without first creating a genuine norm.

6. Conclusion: the state of research on norms

For its first two decades, law and economics largely ignored social norms and conventions. In this period, law and economics scholars implicitly embraced "legal centralism"—the idea that government provides the only source of order, and law the only set of enforced rules. In the 1990's however, law and economics "discovered" social norms and (in the terminology we use) conventions as sources of what Ellickson (1991) calls "Order Without Law." Incorporating informal order into the analysis substantially changes both normative analysis and positive predictions of behavior with and without law.

The effects are so varied and pervasive that they are difficult to summarize, but we note a few major points. Where there was once a presumption that a given legal rule influenced behavior, there is now greater appreciation of the need for empirical research to verify a law's influence, given the possibility that a norm produced the required behavior prior to (and independent of) the law, or that a norm causes people to ignore the law. Where it once appeared that law offered the only solution to the market failure of externalities, we now see that norms often work to punish those who create negative externalities and reward those who create positive ones. At the same time, where theorists once emphasized the need for law to overcome collective action problems caused by individuals maximizing their narrow self-interest, theorists now recognize an important role for law to correct inefficient conventions and norms. Where welfare economics once gave little attention to the fact that the rules it identified as optimal might be perceived as unfair, it is now more accepted that such perceptions are common and their effect on utility—whether or not it fits the taste of the analyst—must be incorporated into social welfare. Where it once seemed that legal compliance was simply a function of deterrence and incapacitation, we can now explain why the norms of legal obedience and the rule of law matter too, and how more specific norms and conventions can either reinforce or undermine legal sanctions. Indeed, the very operation of some core legal institutions—the jury, the police, the bar—may depend significantly upon the norms that regulate them, and we cannot say whether an institution is efficient or inefficient without knowing which norms are interacting with it.

The breadth of the norms literature can also be understood by the variety of issues and legal topics that it has addressed—property, torts, contracts, criminal law, tax, etc. Here is a sign that the literature is maturing. The initial wave of norms scholarship in the early to mid 1990s (e.g. Katz, 1996) tended to discuss the general topic of norms and to justify its importance by considering various puzzles or anomalies that could be explained only by norms. Some papers did focus on a narrow legal topic or problem and used norms as one part of the economic toolkit, but it is only in the past few years that this form of scholarship has dominated. We take this change as a sign of progress. Norms are no longer the concern of only "norms scholars" but of a large set of law and economics scholars—indeed, of rational-choice scholars generally—who see norms as one useful concept among many for understanding behavior.

In this regard, the literature retains huge potential, as there are many areas of law in which there has been little or no attention to norms, and since so much of the work will require detailed empirical examination. Indeed, Scott (2000, p. 1644) turns the tables on the heterodoxy nicely, saying that scholars in "behavioral" and "norms" law and economics alike have fallen into the "Fundamental Attribution Error,"—"the experimentally observed tendency of humans to make the mistake of overestimating the importance of fundamental human traits and underestimating the importance of situation and context." By this he means that we scholars like broad theories and dislike the hard work of learning the facts of particular situations. This is a valid point, and, indeed, economists fall into this trap when they reject "taste-based" explanations of behavior—which require empirical validation to be useful—in favor of more general explanations based on price changes.

Norms scholarship is much like public choice theory. Both give new insight by asking new questions. Public choice theory asks questions such as, "Are the costs of a law concentrated and the benefits diffused?" and "Is this law difficult to understand for those who would be hurt by it?" Norms scholarship asks questions such as, "Is there a reason why this form of disutility would benefit a society in which it exists?" and "Was there a reason for this norm in the past, even if it is pernicious now?" The two ideas are complementary, as any two big good ideas would be (Geoffrey Miller's (2004) article, "Norms and Rents" is a good example of how they can be combined smoothly). Public choice helps explain why an inefficient norm might exist—it might have been to the advantage of certain concentrated interests to create such a norm. And norms help explain why lobbying groups exist—citizens may feel badly if they fail to aid a lobby that helps them, even though the lobby would probably succeed without any one person's contribution. Yet both norms and public choice are subdisciplines that claim much, and whose reach sometimes exceeds their grasp.

Norms have explanatory power. They explain why so much behavior seems to be efficient, internalizing externalities even when laws and material self-interest do not constrain behavior. We must beware, however, of simply saying "It's a norm that's doing it!" whenever behavior seems puzzling. And we must avoid attributing too much influence to norms even when they do exist. It is clear that people act on their principles, but it is also clear that people will sacrifice one principle for another on occasion, or sacrifice a principle for a taste. Any economist knows full well that if the price of a good rises, the quantity demanded will fall, and a principle is in this respect like any other good. Recall the conclusion of Carlson (2001) that convenience was crucial in determining the amount of recycling, and note the psychology experiments recounted in Jeffrey Rachlinski's 2000 warning that mundane considerations of instructions and inertia often trump even norms as religious beliefs (in the Darley-Batson "good Samaritan" experiment) or opposition to torture (in the Milgram "electric shock" experiment). We, like most of those who have thought hard about norms, believe that they are important and useful in explaining behavior. But it is important also not to forget magnitudes of incentives, or the need to carefully consider how hard it is to change those magnitudes. Every one of us has principles—but how many of us are principled?

References

Articles and Books

Akerlof, G.A. (1985). "Discriminatory, status-based wages among tradition-oriented, stochastically trading coconut producers". Journal of Political Economy 93 (2), 265–276.

Akerlof, G.A. (1980). "A theory of social custom, of which unemployment may be one consequence". Quarterly Journal of Economics 94 (4), 749–775.

Akerlof, G.A., Yellen, J. (1986). Efficiency Wage Models of the Labor Market. Cambridge University Press, Cambridge.

Aquinas, T. (1274). Summa Theologica. Ave Maria Press. (http://www.newadvent.org/summa/.)

Aristotle, Nichomachean Ethics, tr. D.P. Chase, Dover Publications (1998). (http://www.gutenberg.net/browse/bibrec/br8438.htm.)

Axelrod, R., Hamilton, W. (1981). "The evolution of cooperation". Science 211, 1390–1396.

Baker, T. (2001). "Blood money, new money, and the moral economy of tort law in action". Law and Society Review 35, 275–319.

Banerjee, A.V. (1992). "A simple model of herd behavior". Quarterly Journal of Economics 107, 797–817.

Baron, J. (2001). Thinking and Deciding, 3d edn. Cambridge University Press, Cambridge.

Benkler, Y. (2002). "Coase's penguin, or, Linux and the nature of the firm". Yale Law Journal 112, 369–446.

Bergstrom, T.C. (2002). "Evolution of social behavior: Individual and group selection". Journal of Economic Perspectives 16 (2), 67–88.

Bernstein, L. (1992). "Opting out of the legal system: Extralegal contractual relations in the diamond industry". Journal of Legal Studies 21, 115–157.

Bernstein, L. (1996). "Merchant law in a merchant court: Rethinking the code's search for immanent business norms". University of Pennsylvania Law Review 144, 1765–1821.

Bernstein, L. (1999). "The questionable basis of Article 2's incorporation strategy: A preliminary strategy". The University of Chicago Law Review 66, 710–780.

Bernstein, L. (2001). "Private commercial law in the cotton industry: creating cooperation through rules, norms, and institutions". Michigan Law Review 99 (7), 1724–1788.

Bikhchandani, S., Hirshleifer, D., Welch, I. (1992). "A theory of fads, fashion, custom, and cultural change as informational cascades". Journal of Political Economy 100, 992–1026.

Boyd, R., Richerson, P.J. (1988). Culture and Evolutionary Process. University of Chicago Press, Chicago.

Brennan, G., Pettit, P. (2004). The Economy of Esteem: An Essay on Civil and Political Society. Oxford University Press, Oxford.

Brennan, G., Pettit, P. (2000). "The hidden economy of esteem". Economics and Philosophy 16 (1), 77–98.

Brinig, M. (1990). "Rings and promises". Journal of Law, Economics and Organization 6 (1), 203–216.

Brinig, M., Nock, S. (2003). "'I only want trust': Norms, trust, and autonomy". Journal of Socio-Economics 32, 471–487.

Brooks, R.E. (2003). "The new imperialism: Violence, norms, and the 'rule of law'". Michigan Law Review 101 (7), 2275–2340.

Buckley, F.H. (2003). The Morality of Laughter. University of Michigan, Ann Arbor.

Buckley, F.H., Rasmusen, E. (2000). "The uneasy case for the flat tax". Constitutional Political Economy 11, 295–318.

Budziszewski, J. (1997). Written on the Heart: The Case for Natural Law. Inter-Varsity Press.

Cameron, S. (2001). "The Economic analysis of social customs: The case of premarital sex". Journal of Evoluationary Economics 11 (4), 457–473.

Cappel, A.J. (2003). "Bringing cultural practice into law: Ritual and social norms jurisprudence". Santa Clara Law Review 43 (2), 389–494.

Carlson, A. (2001). "Recycling norms". California Law Review 89, 1231–1300.

Cheung, S.N. (1973). "The fable of the bees: An economic investigation". Journal of Law and Economics 16, 11–33.
Cooter, R.D. (1991). "Inventing market property: The land courts of Papua New Guinea". Law and Society Review 25 (4), 759–801.
Cooter, R.D. (1996). "Decentralized law for a complex economy: the structural approach to adjudicating the new law merchant". University of Pennsylvania Law Review 144, 1643–1696.
Cooter, R.D. (1997). "Punitive damages, social norms, and economic analysis". Law and Contemporary Problems 60, 73–91.
Cooter, R.D. (1998). "Models of morality in law and economics: self-control and self-improvement for the 'bad man' of Holmes". Boston University Law Review 78, 903–930.
Cooter, R.D., Landa, J.T. (1984). "Personal versus impersonal trade: the size of trading groups and contract law". International Review of Law and Economics 4 (1), 15–22.
Cooter, R.D., Porat, A. (2001). "Should courts deduct nonlegal sanctions from damages?" Journal of Legal Studies 30 (2), 401–422.
Cooter, R.D., Fikentscher, W. (1998). "Indian common law: The role of custom in Indian tribal courts". American Journal of Comparative Law 46, 287–337.
Cosmides, L., Tooby, J., Barkow, J.H. (1992). "Introduction: Evolutionary psychology and conceptual integration". In: Barkow, J.H., Cosmides, L., Tooby, J. (Eds.), The Adapted Mind: Evolutionary Psychology and the Generation of Culture. Oxford University Press, Oxford, pp. 3–15.
Cowen, T. (2002). "The esteem theory of norms". Public Choice 113 (1–2), 211–224.
Cox, J.C., Deck, C.A. (2005). "On the nature of reciprocal motives". Economic Inquiry 43 (3), 623–635.
Darwin, C. (1874). The Descent of Man, Reprint of 2d edn. Prometheus Books, Amerherst, N.Y.
Dana, D. (2001). "Rethinking the puzzle of escalating penalties for repeat offenders". Yale Law Journal 110, 733–783.
Davis, K., Trebilcock, M.J., Heys, B. (2001). "Ethnically homogeneous commercial elites in developing countries". Law and Policy in International Business 32, 331–361.
Dau-Schmidt, K.G. (1990). "An economic analysis of the criminal law as a preference-shaping policy". Duke Law Journal 1990, 1–38.
Dawes, R. (1988). Rational Choice in an Uncertain World. Harcourt Brace, Fort Worth, Texas.
Dharmapala, D., McAdams, R.H. (2003). "The Condorcet jury theorem and the expressive function of law: A theory of informative law". American Law and Economics Review 5, 1–31.
Drahozal, C.R. (2000). "Commercial norms, commercial codes, and international commercial arbitration". Vanderbilt Journal of Transnational Law 33, 79–146.
Ellickson, R.C. (1986). "Of Coase and cattle: Dispute resolution among neighbors in Shasta County". Stanford Law Review 38, 623–688.
Ellickson, R.C. (1991). Order without Law: How Neighbors Settle Disputes. Harvard University Press, Cambridge, Mass.
Ellickson, R.C. (1996). "Controlling chronic misconduct in city spaces: of panhandlers, skid rows, and public-space zoning". Yale Law Journal 105, 1165–1248.
Ellickson, R.C. (1998). "Law and economics discovers social norms". Journal of Legal Studies 27, 537–552.
Ellickson, R.C. (2001). "The market for social norms". American Law and Economics Review 3 (1), 1–49.
Ellickson, R.C. (2005). "Norms of the household". In: Drobak, J. (Ed.), Norms and the Law. Cambridge University Press, Cambridge.
Epstein, J.M. (2001). "Learning to be thoughtless: Social norms and individual computation". Computational Economics 18 (1), 9–24.
Epstein, R.A. (1992a). "International News Service v. Associated Press: custom and law as sources of property rights in news". Virginia Law Review 78, 85–128.
Epstein, R.A. (1992b). "The path to The T.J. Hooper: the theory and history of custom in the law of tort". Journal of Legal Studies 21 (1), 1–38.
Epstein, R.A. (2002). "The allocation of the commons: Parking on public roads". The Journal of Legal Studies 31 (2), S515–S544.

Fallon, R.H. Jr. (1997). " 'The rule of law' as a concept in constitutional discourse". Columbia Law Review 97 (1), 1–56.
Fershtman, C., Weiss, Y. (1998). "Why do we care what others think about us?" In: Ben-Ner, A., Putterman, L. (Eds.), Economics, Values, and Organization. Cambridge University Press, Cambridge, pp. 133–150.
Fischel, D. (1998). "Lawyers and confidentiality". University of Chicago Law Review 65 (1), 1–33.
Fremling, G.M., Posner, R.A. (1999). "Status signaling and the law, with particular application to sexual harassment". University of Pennsylvania Law Review 147, 1069–1109.
Frey, B.S. (1994). "How intrinsic motivation is crowded out and in". Rationality and Society 6 (3), 334–352.
Geisinger, A. (2002). "A belief change theory of expressive law". Iowa Law Review 88, 35–73.
George, R.P. (Ed.) (1995). Natural Law Theory: Contemporary Essays. Oxford University Press, Oxford.
Ginsburg, T., McAdams, R.H. (2004). "Adjudicating in anarchy: An expressive theory of international dispute resolution". William and Mary Law Review 45, 1229–1339.
Goldsmith, J., Posner, E.A. (1999). "A theory of international customary law". University of Chicago Law Review 66, 1113–1177.
Goodman, R., Jinks, D. (2004). "How to influence states: Socialization and international human rights law". Duke Law Journal 54, 612–703.
Green, S.P. (2002). "Plagiarism, norms, and the limits of theft law: Some observations on the use of criminal sanctions in enforcing intellectual property rights". Hastings Law Journal 54, 167–242.
Guzman, A.T. (2002). "A compliance-based theory of international law". California Law Review 90, 1823–1887.
Hadfield, G.K. (1999). "A coordination model of the sexual division of labor". Journal of Economic Behavior and Organization 40, 125–153.
Harder, D.W. (1995). "Shame and guilt assessment, and relationships of shame- and guilt-proneness to psychopathology". In: Tangney, J.P., Fischer, K.W. (Eds.), Self-Conscious Emotions: The Psychology of Shame, Guilt, Embarrassment, and Pride. Guilford Press, New York.
Hardin, R. (1991). "Hobbesian political order". Political Theory 19 (2), 156–180.
Harris, M. (1974). Cows, Pigs, Wars and Witches: The Riddles of Culture. Random House, New York.
Hasen, R.L. (1996). "Voting without law?" University of Pennsylvania Law Review 144, 2135–2179.
Hayden, R.M., Anderson, J.K. (1979). "On the evaluation of procedural systems in laboratory experiments: A critique of Thibaut and Walker". Law and Human Behavior 3, 21–38.
Hayek, F.A. (1973). Law, Legislation, and Liberty, Volume 1. University of Chicago Press, Chicago.
Hetcher, S.A. (1999). "Creating safe social norms in a dangerous world". Southern California Law Review 73, 1–86.
Hetcher, S.A. (2004). Norms in a Wired World. Cambridge University Press, Cambridge.
Hirshleifer, D. (1995). "The blind leading the blind: social influence, fads, and informational cascades". In: Tommasi, M., Ierulli, K. (Eds.), The New Economics of Human Behavior. Cambridge University Press, Cambridge, pp. 188–215. (Chapter 12).
Hirshleifer, D., Rasmusen, E. (1989). "Cooperation in a repeated prisoner's dilemma with ostracism". Journal of Economic Behavior and Organization 12, 87–106.
Hirshleifer, J. (1978). "Natural economy versus political economy". Journal of Social and Biological Structures 1 (4), 319–337.
Hirshleifer, J. (1987). "On the emotions as guarantors of threats and promises". In: Dupre, J. (Ed.), The Latest on the Best: Essays on Evolution and Optimality. MIT Press, Cambridge, Mass.
Holmes, O.W. Jr. (1897). "The path of the law". Harvard Law Review 10 (8), 457–478. (http://onlinebooks.library.upenn.edu/webbin/gutbook/lookup?num=2373.)
Horne, C. (2001). "The contribution of norms to social welfare: Grounds for hope or pessimism?" Legal Theory 7, 159–177.
Hume, D. (1751). Enquiry Concerning the Principles of Morals. Hackett Publishing. (http://onlinebooks.library.upenn.edu/webbin/gutbook/lookup?num=4320.)
Kahan, D.M. (1996). "What do alternative sanctions mean?" University of Chicago Law Review 63, 591–653.
Kahan, D.M. (1998). "Social meaning and the economic analysis of crime". Journal of Legal Studies 27, 609–622.

Kahan, D.M. (1999). "Strategies for private norm enforcement in the inner city". UCLA Law Review 46, 1859–1872.
Kahan, D.M. (2000). "Gentle nudges vs. hard shoves: Solving the sticky norms problem". University of Chicago Law Review 67, 607–645.
Kahan, D.M. (2001). "The limited significance of norms for corporate governance". University of Pennsylvania Law Review 149, 1869–1900.
Kahan, D.M. (2002). "Signaling or reciprocating? A response to Eric Posner's law and social norms". University of Richmond Law Review 36, 367–385.
Kahan, D.M. (2003). "The logic of reciprocity: trust, collective action, and law". Michigan Law Review 102 (1), 71–103.
Kahneman, D., Slovic, P., Tversky, A. (Eds.) (1982). Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge.
Kaplow, L. (1995). "A note on subsidizing gifts". Journal of Public Economics 58 (3), 469–477.
Kaplow, L., Shavell, S. (2001a). "Moral rules and the moral sentiments: toward a theory of an optimal moral system". Harvard Law and Economics, Discussion Paper No. 342. (http://ssrn.com/abstract=293906.)
Kaplow, L., Shavell, S. (2001b). "Any non-welfarist method of policy assessment violates the pareto principle". Journal of Political Economy 109, 281–286.
Kaplow, L., Shavell, S. (2002a). Fairness versus Welfare. Harvard University Press, Cambridge, Mass.
Kaplow, L., Shavell, S. (2002b). "Human nature and the best consequentialist moral system". Harvard Center for Law, Economics, and Business Discussion Paper No. 349.
Katz, A. (1996). "Law, economics and norms: Taking private ordering seriously". University of Pennsylvania Law Review 144, 1745–1763.
Kindleberger, C. (1983). "Standards as public, collective and private goods". Kyklos 36 (3), 377–396.
Kim, P. (1999). "Norms, learning and law: Exploring the influences on worker's legal knowledge". University of Illinois Law Review 1999 (2), 447–515.
Kirsch, M.S. (2004). "Alternative sanctions and the federal tax law: Symbols, shaming, and social norm management as a substitute for effective tax policy". Iowa Law Review 89, 863–939.
Klein, B., Leffler, K. (1981). "The role of market forces in assuring contractual performance". Journal of Political Economy 89, 615–641.
Kreps, D., Milgrom, P., Roberts, J., Wilson, R. (1982). "Rational cooperation in the finitely repeated prisoners' dilemma". Journal of Economic Theory 27, 245–252.
Kübler, D. (2001). ""On the regulation of social norms". Journal of Law, Economics and Organization 17 (2), 449–476.
Kuran, T. (1995). Private Truths, Public Lies: The Social Consequences of Preference Falsification. Harvard University Press, Mass.
Kuran, T. (1998). "Ethnic norms and their transformation through reputational cascades". Journal of Legal Studies 27, 623–659.
Kuran, T., Sunstein, C.R. (1999). "Availability cascades and risk regulation". Stanford Law Review 51, 683–768.
Landa, J.T. (1981). "A theory of the ethnically homogeneous middleman group: An institutional alternative to contract law". Journal of Legal Studies 10 (2), 349–362.
Landa, J.T. (1994). Trust, Ethnicity, and Identity. University of Michigan Press, Ann Arbor.
Lederman, L. (2003). "The interplay between norms and enforcement in tax compliance". Ohio State Law Journal 64, 1453–1513.
Lerner, J., Tirole, J. (2002). "Some simple economics of open source". Journal of Industrial Economics 50, 197–234.
Macaulay, S. (1963). "Non-Contractual relations in business". American Sociological Review 28, 55–70.
Mahoney, P.G., Sanchirico, C. (2001). "Competing norms and social evolution: Is the fittest norm efficient?" University of Pennsylvania Law Review 149, 2027–2062.
Mahoney, P.G., Sanchirico, C. (2003). "Norms, repeated games, and the role of law". California Law Review 91, 1281–1329.

Massell, G.J. (1968). "Law as an instrument of revolutionary change in a traditional milieu: The case of Soviet Central Asia". Law and Society Review 2 (2), 179–228.
Maynard-Smith, J., Parker, G.A. (1976). "The logic of asymmetric contests". Animal Behavior 24, 159–175.
McAdams, R.H. (1995). "Cooperation and conflict: The economics of group status production and race discrimination". Harvard Law Review 108, 1003–1084.
McAdams, R.H. (1996). "Group norms, gossip, and blackmail". University of Pennsylvania Law Review 144, 2237–2292.
McAdams, R.H. (1997). "The origin, development and regulation of norms". Michigan Law Review 96, 338–433.
McAdams, R.H. (2000). "An attitudinal theory of expressive law". University of Oregon Law Review 79, 339–390.
McAdams, R.H. (2001a). "Conventions and norms: philosophical aspects". In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of Social and Behavioral Sciences, vol. 4. Pergamon, Oxford, pp. 2735–2741.
McAdams, R.H. (2001b). "Signaling discount rates: Law, norms, and economic methodology". Yale Law Journal 110 (4), 625–689.
McGowan, D. (2001). "Legal implications of open-source software". University of Illinois Law Review 2001, 241–304.
Milhaupt, C.J., West, M.D. (2000). "The dark side of private ordering: An institutional and empirical analysis of organized crime". University of Chicago Law Review 67, 41–98.
Mill, J.S. (1859). On Liberty, Dover Publications. (http://www.bartleby.com/130/.)
Miller, G.P. (2003). "Norm enforcement in the public sphere: The case of handicapped parking". George Washington Law Review 71, 895–933.
Miller, G.P. (2004). "Norms and interests". Hofstra Law Review 32 (2), 637–674.
Murphy, K. (2004). "The role of trust in nurturing compliance: A study of accused tax avoiders". Law and Human Behavior 28 (2), 187–209.
Ng, Y. (1999). "Utility, informed preference, or happiness: following Harsanyi's argument to its logical conclusion". Social Choice and Welfare 16 (2), 197–216.
Ordeshook, P.C. (2002). "Are 'Western' constitutions relevant to anything other than the countries they serve?" Constitutional Political Economy 13 (1), 3–24.
Ostrom, E. (1990). Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press, Cambridge.
Ostrom, E. (1991). "Rational choice theory and institutional analysis: toward complementarity". The American Political Science Review 85 (1), 237–243.
Painter, R.W. (2001). "Rules lawyers play by". New York University Law Review 76, 665–749.
Pettit, P.N. (1990). "Virtus normativa: rational choice perspectives". Ethics 100 (4), 725–755.
Picker, R.C. (1997). "Simple games in a complex world: a generative approach to the adoption of norms". University of Chicago Law Review 64, 1225–1287.
Polinsky, A.M., Shavell, S. (2000). "The fairness of sanctions: Some implications for optimal enforcement policy". American Law and Economics Review 2 (2), 223–237.
Posner, E.A. (1996a). "The legal regulation of religious groups". Legal Theory 2 (1), 33–62.
Posner, E.A. (1996b). "The regulation of groups: the influence of legal and nonlegal sanctions on collective action". University of Chicago Law Review 63, 133–197.
Posner, E.A. (1998). "Symbols, signals, and social norms in politics and the law". Journal of Legal Studies 27, 765–797.
Posner, E.A. (1999). "Family law and social norms". In: Buckley, F. (Ed.), The Fall and Rise of Freedom of Contract. Duke University Press, Durham.
Posner, E.A. (2000a). "Law and social norms: the case of tax compliance". Virginia Law Review 86, 1781–1820.
Posner, E.A. (2000b). Law and Social Norms. Harvard University Press, Cambridge, Mass.
Posner, R.A. (1992). Sex and Reason. Harvard University Press, Cambridge, Mass.

Posner, R.A. (1997). "Social norms and the law: an economic approach". American Economic Review, Papers and Proceedings 87, 365–369.
Posner, R.A., Rasmusen, E. (1999). "Creating and enforcing norms, with special reference to sanctions". International Review of Law and Economics 19, 369–382.
Priest, G.L. (1997). "The common law process and the selection of efficient rules". Journal of Legal Studies 6, 65–82.
Rabin, M. (1993). "Incorporating fairness into game theory and economics". American Economic Review 83 (5), 1281–1302.
Rachlinski, J.J. (2000). "The limits of social norms". Chicago-Kent Law Review 74, 1537–1567.
Ramseyer, J.M. (1994). "Learning to love Japan: Social norms and market incentives". San Diego Law Review 34, 263–267.
Rasmusen, E. (1996). "Stigma and self-fulfilling expectations of criminality". Journal of Law and Economics 39, 519–544.
Rasmusen, E. (2002). "An economic approach to adultery law". In: Dnes, A., Rowthorn, R. (Eds.), Marriage and Divorce: An Economic Perspective. Cambridge University Press, Cambridge, pp. 70–91. Chapter 5.
Rasmusen, E. (2006). Games and Information, 4th edn. Blackwell Publishers, Oxford.
Rasmusen, E., Stake, J. (1998). "Lifting the veil of ignorance: Personalizing the marriage contract". Indiana Law Journal 73, 454–502.
Richman, B. (2004). "How communities create economic advantage: Jewish diamond merchants in New York". Harvard Law and Economics Discussion Paper No. 384. (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=349040.)
Robinson, P.H., Darley, J.M. (1997). "The utility of desert". Northwestern University Law Review 91, 453–499.
Rock, E., Wachter, M. (2002). "Meeting by signals, playing by norms: Complementary accounts of nonlegal cooperation in institutions". University of Richmond Law Review 36, 423–442.
Rose-Ackerman, S. (1999). Corruption and Government: Causes, Consequences, and Reform. Cambridge University Press, Cambridge.
Rothschild, M. (1974). "A two-armed bandit theory of market pricing". Journal of Economic Theory 9 (2), 185–202.
Rubin, P.H. (1977). "Why is the common law efficient?" Journal of Legal Studies 6, 51–63.
Rubin, P.H. (1982). "Evolved ethics and efficient ethics". Journal of Economic Behavior and Organization 3, 161–174.
Sartorius, C. (2002). "The relevance of the group for the evolution of social norms and values". Constitutional Political Economy 13, 149–172.
Schwartz, W.F., Baxter, K., Ryan, D. (1984). "The duel: can these gentlemen be acting efficiently?" Journal of Legal Studies 13, 321–355.
Scott, R.E. (2003). "A theory of self-enforcing indefinite agreements". Columbia Law Review 103 (7), 1641–1699.
Scott, R.E. (2000). "The limits of behavioral theories of law and social norms". Virginia Law Review 86, 1603–1645.
Scott, E.S., Scott, R.E. (1995). "Parents as fiduciaries". Virginia Law Review 81, 2401–2476.
Selten, R. (1978). "The chain-store paradox". Theory and Decision 9 (2), 127–159.
Shavell, S. (2002). "Law versus morality as regulators of conduct". American Law and Economics Review 4 (2), 227–257.
Singer, P. (1981). The Expanding Circle: Ethics and Sociobiology. Farrar, Straus and Giroux.
Skeel, D.A. Jr. (2001). "Shaming in corporate Law". University of Pennsylvania Law Review 149, 1811–1868.
Smith, A. (1776). The Wealth of Nations. The Liberty Fund, Indianapolis. (http://www.econlib.org/library/Smith/smWN.html.)
Smith, A. (1790). The Theory of the Moral Sentiments. The Liberty Fund, Indianapolis. (http://socserv2.socsci.mcmaster.ca/~econ/ugcm/3ll3/smith/moral.html.)

Stephen, J.F. (1873). "Liberty, Equality, Fraternity: And Three Brief Essays". In: Posner, R.A. (Ed.). University of Chicago Press, Chicago. (http://my.execpc.com/~berrestr/stelib.html.)
Strahilevitz, L.J. (2000). "How changes in property regimes influence social norms: commodifying California's carpool lanes". Indiana Law Journal 75, 1231–1296.
Strahilevitz, L.J. (2003). "Charismatic code, social norms, and the emergence of cooperation on the file-swapping networks". Virginia Law Review 89, 505–595.
Strotz, R.H. (1955). "Myopia and Inconsistency in dynamic utility maximization". Review of Economic Studies 23 (3), 165–180.
Sugden, R. (1986). The Economics of Rights, Co-operation and Welfare. Blackwell Publishers, Oxford.
Sugden, R. (1998). "Normative expectations: the simultaneous evolution of institutions and norms". In: Ben-Ner, A., Putterman, L. (Eds.), Economics, Values, and Organization, vol. 73. Cambridge University Press, Cambridge.
Sunstein, C.R. (1996). "On the expressive function of law". University of Pennsylvania Law Review 144, 2021–2053.
Sunstein, C.R., Schkade, D., Kahneman, D. (2000). "Do people want optimal deterrence?" Journal of Legal Studies 29, 237–253.
"Symposium: Law, economics, and norms" (1996). University of Pennsylvania Law Review 144 (May).
"Symposium: Social norms, social meaning, and the economic analysis of law" (1998). Journal of Legal Studies 27 (June).
"Symposium: Corporate law and social norms" (1999). Columbia Law Review 99 (June).
"Symposium: The legal construction of norms" (2000a). Virginia Law Review 86 (November).
"Symposium on norms, law, and order in the city" (2000b). Law and Society Review 34.
"Symposium: Norms and corporate law" (2001). University of Pennsylvania Law Review 149 (2001) 1735–2191.
Tangney, J.P. (1995). "Recent advances in the empirical study of shame and guilt". American Behavioral Scientist 38, 1132–1145.
Teichman, D. (2005). "Sex, shame, and the law: An economic perspective on Megan's Law". Harvard Journal on Legislation 42, 355.
Thibaut, J., Walker, L. (1975). Procedural Justice: A Psychological Analysis. Wiley, New York.
Trivers, R.L. (1971). "The evolution of reciprocal altruism". Quarterly Review of Biology 46 (1), 35–57.
Vandenbergh, M. (2003). "Beyond elegance: a testable typology of social norms in corporate environmental compliance". Stanford Environmental Law Journal 22, 55–144.
Wells, G., Petty, R. (1980). "The effect of overt head movements on persuasion". Basic and Applied Social Psychology 1 (3), 219–230.
Wendel, W.B. (2001). "Nonlegal regulation of the legal profession: Social norms in professional communities". Vanderbilt Law Review 54, 1955–2055.
West, M.D. (1997). "Legal rules and social norms in Japan's secret world of sumo". Journal of Legal Studies 26, 165–201.
Whitman, Do.G. (1998). "Hayek contra Pangloss on evolutionary systems". Constitutional Political Economy 9, 45–66.
Wilson, E.O. (1980). Sociobiology. Harvard University Press, Cambridge, Mass.
Wilson, J.O. (1993). The Moral Sense. Simon and Schuster, New York.
Zerbe, R.O. Jr. (2001). Efficiency in Law and Economics. Edward Elgar, Aldershot, New York.
Zywicki, T.J. (2003). "The rise and fall of efficiency in the common law: A supply-side analysis". Northwestern Law Review 97, 1551–1633.

Cases and Statutes

Brown v. Board of Education, 347 U.S. 483 (1954).
Donovan v. Fiumara, 114 N.C. App. 524 (1994).
Goldfarb v. Virginia State Bar, 421 U.S. 773 (1975).

Jenkins v. Georgia 418 US 153 (1974).
Model Penal Code §223.4 (1985).
Pabalk Ticaret Limited Sirketi v. Norsolor S.A., ICC Award of Oct. 26, 1979, No. 3131, 9 Y.B. Commercial Arb. 109 (1984).
Texas and Pacific Railway Company v. Behymer, 189 U.S. 468 (1903).
The 1793 Constitution of France.
The 1795 Constitution of France.
The Uniform Commercial Code.
The T.J. Hooper, 60 F2d 737 (2d Cir. 1932).
Titus v. Bradford, 20 A. 517 (Pa. 1890).

Chapter 21

EXPERIMENTAL STUDY OF LAW

COLIN CAMERER
School of Social Sciences, California Institute of Technology

ERIC TALLEY*
School of Law, University of California, Berkeley

Contents

* Camerer is Rea A. & Lela G. Axline Professor of Business Economics, California Institute of Technology, and Talley is Professor of Law, UC Berkeley and Senior Economist, RAND Corporation. Many thanks to Jennifer Arlen, Mitch Polinsky, Steve Shavell, and participants at a conference at Harvard Law School for helpful comments and discussions. Jennifer Lam provided excellent research assistance. All errors are ours.

Handbook of Law and Economics, Volume 2
Edited by A. Mitchell Polinsky and Steven Shavell

DOI: 10.1016/S1574-0730(07)02021-X

Abstract

This chapter surveys literature on experimental law and economics. Long the domain of legally minded psychologists and criminologists, experimental methods are gaining significant popularity among economists interested in exploring positive and normative aspects of law. Because this literature is relatively new among legally-minded economists, we spend some time in this survey on methodological points, with particular attention to the role of experiments within theoretical and empirical scholarship, the core ingredients of a well done experiment, and common distinctions between experimental economics and other fields that use experimental methods. We then consider a number of areas where experimental evidence is increasingly playing a role in testing the underlying foundational precepts of economic behavior as it applies to law, including bargaining in the shadow of the law, the selection of suits for litigation, and the investigation of jury and judge behavior. Our survey concludes by offering some suggestions about what directions experimental economists might push the methodology in the study of legal rules.

Keywords

Behavioral economics, experimental methods, experimental law and economics, replicability, generalizability, expected utility (EU), rational choice, equilibrium, quantal response equilibrium (QRE), cognitive hierarchy (CH), Coase theorem, endowment effect, self-serving bias, jury, judge, hindsight bias, norms

JEL classification: A12, C91, C92, C93, K10, K40

1. Introduction

Few methodological approaches to the study of law have received more recent attention than experimental methods. Virtually absent from the pages of law reviews and law and economics journals just a decade ago, experimental studies (or articles purporting to be inspired by the results of such studies) have become a veritable staple consumption good for today's legal scholars.

In many respects, the emergence of experimental methods to analyze law should not be terribly surprising, particularly within law and economics (and affiliated fields). Over an even longer period of time, experimental methods have become relatively well established in both economics and political science proper—two disciplines that served as central foci in generating insights that inform the law and economics literature. In addition, the emerging field of "behavioral economics" has begun to synthesize findings from economics, political science and psychology into a more unified theory of individual and multi-person decision theory. See, for example, Camerer (2006), Camerer, Loewenstein, and Rabin (2004). Because psychologists have long depended primarily on experimental methods, these interdisciplinary approaches were a natural fit for such methodological emphases. Indeed, during the last five years, a sub-discipline of "behavioral law and economics" (or BLE[1]) has emerged that largely echoes the approach in behavioral economics. This very sub-field has similarly enjoyed significant popularity in legal scholarship over the last decade.[2]

Experimental methods are but one methodological approach within the field of law and economics, and by far the most recent to take root. Most of the initial insights emanating from within law and economics during the 1960s and 1970s came from applications of core insights from microeconomic theory. Game theorists similarly claimed some analytical terrain during the 1980s and 1990s, incorporating insights from repeat play, asymmetric information, and evolutionary selection models into the analysis of law. During this period, empirical methods also began to emerge, particularly as methods for collecting and analyzing data from the court system became more reliable and feasible. In many ways, empirical methods have proven a helpful means for testing the numerous predictions made within theoretical law and economics.

Nevertheless, as we elaborate below, empirical approaches suffer from the fact that it is often difficult to stage (much less to observe by happenstance) a truly natural experiment in the real world that implies clear causal conclusions. Because laboratory approaches excel in just this respect, at the very least good experimental designs are likely to provide a complementary and confirmatory check on empirical methods. Accordingly, our enterprise in this essay is threefold: (a) to articulate more specifically how and where experimental methods fit into the larger tapestry of legal studies from

[1] See, e.g., Jolls, Sunstein, and Thaler (1998) for a review.

[2] A recent Westlaw search, for example, turned up a total of 580 law review articles discussing "behavioral law and economics" during the last seven years.

an economic perspective; (b) to describe contributions that have already been made in the field; and (c) to suggest future courses of inquiry that may well prove fruitful.

Before proceeding, a few caveats about our inquiry deserve specific mention. Although the foci of our inquiry may prove helpful to a number of different audiences, our contribution is intended to resonate most centrally with economists who are interested in legal applications of experimental economics. This target audience may pull within its ambit both seasoned experimentalists who tend not to focus on legal applications, as well as economists who—while not experimentalists by nature—are interested in testing experimentally predictions from non-experimental areas in law and economics.[3] In addition, for those endeavoring to evaluate a piece of experimental law and economics scholarship (e.g., referees, reviewers, and the like), our essay may help provide a background against which to assess the relative creativity, novelty, and substantive contribution of the work in question.

A second (and important) caveat to our treatment concerns how we make distinctions on subject matter and scope. Because experimental law and economics has a relatively short pedigree (only 20 years or so by even the most generous genealogical measures), it has necessarily drawn from and built upon a vast body of disparate research that used experimental approaches to law, but did not centrally concern economic inquiries *per se*. Fields such as law and psychology, criminology, and legal sociology have routinely employed experimental methods to gain purchase on causal claims about aspects of human legal behavior. These literatures, in contrast, have relatively lengthy pedigrees, and it would be virtually impossible to do justice to all of them within the confines of the current essay. We have chosen, therefore, to concentrate largely on the experimental literature developed by economists, except insofar as outside literature has informed the general approach applied by researchers in experimental law and economics. It is important to emphasize that this focus does not in any way discount the many contributions made by these fields for experimental studies, but instead is an artifact of the more targeted scope of this chapter.[4]

Our analysis proceeds as follows. Section 2 discusses general motivational and methodological issues surrounding experimental methods in law, including the purpose of experiments, the core ingredients of a well done experiment, and the central role that experimental evidence is increasingly playing in testing the underlying foundational precepts of economic behavior as it applies to law. Section 3 then considers a number of specific legal settings in which experimental approaches have proven particularly valuable, such as the study of bargaining in the shadow of the law, the analysis of suit and settlement, and the investigation of jury and judge behavior. In each of these contextual applications, experimentalists have informed the state of academic inquiry

[3] Our contribution should not be relied on exclusively, however. Indeed, those new to experimental methods might well benefit from the insights of an experienced co-author, and from other overviews of the experimental approach [such as Croson (2002)].

[4] Those readers looking for a more general treatment would do well to consult the Volume 4 of the University of Illinois Law Review (2002), which is dedicated entirely to experimental methods in law *writ large*.

substantially, and in many cases experimental approaches have played a central role in spawning affiliated theoretical or empirical literatures. Finally, Section 4 concludes by offering a series of observations about where experimental approaches might still be fruitfully expanded or applied in novel ways to the study of law.

2. Motivation and methodology for experimental law and economics

An experiment is the creation of a situation controlled by the experimenter (to some degree), for the purpose of testing a general theory or establishing causality [see Croson (2002) for a "how to" manual aimed at lawyers].[5] As noted above, experimentation, field observation, and theory are ideally complements of one another: Done well, each enhances the marginal productivity of the other enterprise. Moreover, the conditions of the experiment match the assumptions of a theory being tested. (Psychologists sometimes refer to this match as "internal validity.")

A central feature of any sound experimental design is *control*. The crucial features of control are (a) the ability to assign subjects randomly to treatments, and (b) the ability to operationalize features of theory which are often difficult to observe or adjust for econometrically in field data (and also to create situations of theoretical interest which do not occur naturally or regularly). Random assignment is perhaps the most foundational element of a valuable experiment. It is important to be sure that subjects are assigned their respective tasks and roles in a truly random fashion. For example, suppose that subjects are told to sit wherever they like as they arrive, and those in the front row of a lab will act as buyers and those in the back row as sellers. Even though assignment is nominally random, if early-arriving subjects sit in the front row then buyers will be those who are more punctual and latecomers will self-assign themselves as sellers. If, as seems plausible, early arrivers and late arrivers are likely to have differential proclivities, preferences, and the like, then the assignment protocol would not be truly random. It would be easy to draw spurious conclusions about the effects of buyer-seller interactions, when those effects are an artifact of differential arrival times.

After control, a second core desideratum of experimental methods (in both the physical and social sciences) is the *replicability* of an experiment's results. The procedural care that goes into designing a replicable study disciplines experimenters, permits accumulation of regularity, and facilitates tests of robustness to small design changes. Consequently, experimental economists frequently follow a written script meticulously, recording any *ad lib* instructions that are given (often in an experimental log). Such procedures not only allow later researchers to recreate the original experimental setting,

[5] Smith (1982) is a seminal discussion of principles of economic experimentation. A useful recipe book is Friedman and Sunder (1993). Bergstrom and Miller (1999) show how to teach an economic principles course using simple experiments. Cumulated regularities are summarized in Davis and Holt (1993) and Kagel and Roth (1995). In Section 2.3 below we describe some ways in which experimentation in psychology and economics have differed traditionally.

but they are also a prudent way to ameliorate a related problem: the lack of underlying "blindness" of an experiment to the hypotheses being tested. There are numerous subtle ways in which an experimenter's bias toward or against an underlying theory could influence the experiment's results, as well as other subject-experimenter interactions (e.g., male subjects may behave differently in the presence of a female experimenter). Replication by other researchers (who ideally have different theoretical commitments) is an effective check against experimental bias.

Finally, the analysis of experimental data generally follows the rules of good statistical analysis of any type of data: Sample sizes must be large enough to make statistical tests powerful and meaningful, replication effects must be as independent as possible, and researchers should attempt to conduct multiple parametric and nonparametric tests. One complication that is perhaps more symptomatic of experimental studies is the fact that responses are not independent because subjects interact. If subjects are bargaining in pairs, for example, then data can be analyzed both at the individual level and the pair-wise (dyadic) level. Experimental economists frequently make conservative assumptions about independence. For example, in some cases each experimental session is treated as a separate data point, and tests are conducted with each session-wide summary statistic as a single datum. An early tradition, which is unfortunately waning, is to invite subjects back for a second repetition of the same experiment to see whether experienced subjects behave differently than inexperienced ones. When equilibria are complicated, one often finds that experienced subjects are closer to equilibrium predictions[6] than inexperienced ones. This does not imply, of course, that the data from inexperienced subjects are uninteresting. The ability to compare the two groups just helps establish the boundaries of where the equilibrium prediction might work poorly or well.

2.1. Purpose of experiments

Like any methodological approach, experiments can have a number of different purposes. Perhaps most centrally, however, experiments facilitate sharp testing of theory when predictions depend delicately on assumptions that cannot easily be observed in the field. For example, experimental methods in game theory have proved especially useful because predictions often depend on fine details of what players know about others' strategies and how they value consequences [e.g., Camerer (2003); Camerer, Ho, and Chong (2004)]. Hundreds of experiments have shown where equilibrium concepts predict well and predict poorly. These data have also provided lots of raw material to motivate new theories about how learning, natural limits on strategic thinking, and social preference like reciprocity explain strategic behavior.

[6] Note that "equilibrium predictions" need not coincide with equilibria emanating solely from rational choice models. Indeed, as we elaborate below, many elements of behavioral economics are themselves amenable to equilibrium concepts.

Sometimes experiments simply document regularity "pre-theoretically," as an inspiration for theorizing. For example, early experiments established that even small groups of traders (e.g., as few as three buyers and three sellers) rapidly converge to Pareto-optimal competitive allocations in decentralized trading. This result is surprising because formal models of competitive markets had previously assumed infinitely many "price takers," thereby circumventing game-theoretic strategizing environments. The experimental regularity shows that surprisingly competitive results can be created by very small numbers of traders when exchange is centralized. This regularity then provoked theories of how strategic interaction would work in small-numbers settings [Wilson (1985); Cason and Friedman (1993)], and how simple adaptive rules could lead to convergence [Easley and Ledyard (1993); Gode and Sunder (1997)], although this is still largely an unsolved problem.

Experiments have also proved useful in prescriptive domains, helping policymakers understand how certain targeted policy interventions are likely to work when implemented [Plott (1987); Roth (2002)]. The inspiration here is very much akin to experimentation in the physical sciences, such as testing of airplane wing designs in wind tunnels, or testing ship designs in "tow tanks" with simulated oceanic waves. These experiments do not guarantee that a wing or ship which performs well in a wind tunnel or tow tank will be the best design in the air or at sea, but they can weed out bad designs at a low cost. A good recent example is experimental input to the design of telecommunication spectrum auctions, which significantly influenced the actual designs (as did auction theory) in the PCS auctions of the late 1990's, first in the US then later in many other countries [Milgrom (2004, p. 25)]. Experiments could be useful in a similar way, for providing wind-tunnel tests of proposed legal reforms.

2.2. *Generalizability*

Because experiments often study behavior in a specific situation, it is important to establish a clear basis for generalizing from the results of an experiment to a specific domain of interest. As Posner (1998) notes:

> The problem of extrapolating to normal human behavior from behavior in unusual experimental settings ... is obvious ... One would like to know the theoretical or empirical basis for supposing that the experimental environment is relatively similar to the real world. That would be the first question an experimental scientist would address.

Psychologists use the term "external validity" to describe the extent to which the experimental conditions match those of the setting the results are meant to generalize to. We prefer the term "generalizability" as a reminder that the external, naturally occurring world is complex; there is usually not a single "external world" that is different than the artificial experimental world. Furthermore, since experimental facts are meant to be part of a three-way dialogue including theory and field data, the crucial component of

generalizability is whether a theory draws a clear distinction between an artificial experimental envirionment and a naturally-occurring one. For example, an experimentalist may be willing to generalize from behavior of law student subjects in the lab to that of experienced attorneys in the courtroom (or the proverbial courthouse steps), because the theories of interest do not predict any specific effect of experience, so testing them with law students is a legitimate test. That is, if a theory purports to be general, and if its assumptions are carefully reflected in an experimental design (sometimes referred to as "internal validity"), then the criticism that the experiment was not meant to apply to students, for example, is really an admission that the theory's domain of application was not completely specified. A corollary implication of this view is that criticism of an experiment's generalizability is only productive if it is phrased as an hypothesis about how behavior would differ if the design were changed—i.e., the criticism should be in the form of an alternative design and a prediction about how (and why) the design change would matter. Then, the criticism that law students' experimental behavior will differ from that of experienced attorneys is constructive only if it leads to the hypothesis that replicating the experiment with experienced attorneys will yield a different result, and an explanation for how that hypothesis is derived.

Nevertheless, perhaps because of its nexus to real-world institutions, experimental work in law and economics tends to invite enhanced scrutiny along two lines: (a) whether the experiment utilized an appropriate experimental subject pool and (b) whether the experiment employed appropriate remunerative stakes.

As to the first criticism, one can do much to replicate the significant contextual details of a legal decision-maker's action, but the normative and prescriptive implications for legal applications may still be elusive if the subject pools cannot be used to predict the behavior of "real people" who actually make decisions in such contexts. A number of studies have attempted to address this concern, comparing behavior of student subjects against seasoned professionals in specific contexts. Ball and Cech (1996) provide a relatively comprehensive (if now a bit dated) survey of this research, and find that at least within a majority of studies, experienced professionals typically behaved in abstract experiments much like college student subjects did. While such research provides a possible basis for one's *a priori* belief that students and experienced legal professionals behave similarly, it bears noting [as did Ball and Cech (1996)] that as an experimental design grows less abstract—and more practical—the likelihood of deviations between student subjects and professionals increases.

There are a number of approaches for dealing with this concern (see Harrison and List, 2004 for a general discussion). First, one could attempt to recruit solely professional subjects "from the wild" to participate in a laboratory experiment. Such approaches enhance generalizability, but are often both cumbersome and expensive to implement. Indeed, researchers must not only traverse a more attenuated path to con-

tact professional subjects, but then must also must remunerate them more handsomely because of their greater opportunity costs of time.[7]

An alternative (and less expensive) approach is to utilize richer information about cross-sectional variation within student subject pools to construct calibrated predictions about a target population possessing a different demographic composition. A leading example of this approach, in fact, has come from a decidedly legal application. Harrison and Lesley (1992) devised an approach for using student subjects to mimic the actual results of a much larger (and more costly) study assessing contingent valuations for purposes of formulating damages estimates in the *Exxon Valdez* litigation. Carson et al. (1992) had used relatively sophisticated survey methods (based on a fully probabilistic sample of the U.S. population) to infer contingent valuation estimates. In contrast, Harrison and Lesley (1992) collected data from a much smaller group of students at a single university, taking care to record a number of demographic variables (such as age, sex, household income, marital status, and the like), which enabled them to develop a simple statistical model to predict behavior as a function of those observables. They then used the estimated coefficients of this calibrated model to predict (by interpolation or extrapolation) the responses of the target population *as a function* of their (possibly differing) socio-economic characteristics.[8] This simple technique relies on the assumption that the effect of a variable—marital status, for example—is additive and does not interact with student status.

Because of the greater focus that experimental law and economics pays to specific, contextual applications, issues relating to generalizability are likely to remain at the forefront of such applications. In this respect, more studies comparing the two groups in legal experiments would be useful. It is notable that many of the studies below, conducted by legal scholars, do use law students, so that a basic step toward generalizability toward more experienced attorneys has been taken.

2.3. *Psychology and economics experimental conventions*

It is important to note that experimental conventions in psychology and economics have historically been quite different [see Camerer (1997); Loewenstein (1999); Hertwig and Ortmann (2001)], though conventions are being mixed in a fusion as experimenters cross traditional boundaries. Experimental economists tend to insist that subjects' earnings depend on their choices. Most comparisons between no performance incentives, and low and high incentives, show that paying some performance-based incentive does

[7] Most experimenters would be happy to conduct such experiments, by the way, if subjects were available and willing and granting agencies increased grant budgets adequately.

[8] Significantly, a variation of this approach would integrate artefactual experiments as well, allowing researchers to combine samples drawn from convenience subjects with "real world" actors. So doing would further enable the researcher to test directly the representativeness of the convenience population, after controlling for observable demographic traits.

not usually alter summary statistics like mean responses, but sometimes reduces variance in responses, boredom, and "presentation effects" in which subjects exhibit socially desirable behavior when it is cheap to do so [e.g., subjects are less altruistic and risk-preferring when playing for real money; see Camerer and Hogarth (1999); Hertwig and Ortmann (2001)]. Paying very large sums, compared to modest sums, typically does not alter behavior much. But a growing body of evidence from paying modest sums, by the standards of developed countries, in foreign countries where incomes are lower will provide more data on whether paying very high stakes makes a difference.

Experimental economists also regard deception, which is particularly common in social psychology, as a last resort. The taboo against deception is enforced because of a fear that repeated deception undermines experimental credibility in general, which is a public good for all experimenters, and may also be transparent to savvy subjects.

Experimental economists also generally prefer abstract descriptions of the experimental setting (to avoid non-pecuniary motives which may be activated by labeling an action "defect," say, rather than "choose row B"). Abstract descriptions are not used by default out of a belief that context does not matter—in fact, the belief is quite the opposite. Unless a particular theory predicts an effect of the contextual description, specifications of lifelike context are treated as a Pandora's box of nuisance variables with unpredictable effects. Finally, experimental economists are more fastidious about reporting all their data in a raw form (to permit skeptical readers to draw their own conclusions), and typically make their instructions and data available—typically on a website, nowadays—to permit low-cost replicability (which also signals the experimenter's faith in the replicability of their results). Psychologists are more inclined to gather a wide range of cognitive measures, like response times, demographic data about subjects, psychometric tests, subjects' self-reports about motives for their behavior, and brain imaging [e.g., Camerer, Loewenstein, and Prelec (2005), Sanfey et al. (2006) and the November 2004 issue of the Royal Transactions of the Philosophical Society B, on "law and the brain"].

These distinctions between the two approaches have blurred recently, particularly as experimental economists become more interested in the influence of the description of the experimental setting on behavior. Our view is that good experimentation combines the best of both approaches—extreme care in enabling exact replication, paying some incentive for performance to ensure thoughtful responses and reduce outliers (American subjects are typically paid about triple the minimum wage), full disclosure of all the data with website archiving, self-reported "debriefings" after the experiment, and measuring demographic and cognitive variables, when those data are easy to gather and potentially informative (even if not the main focus of study).

2.4. Behavioral economics

A theoretical arena where experimental methods have grown increasingly important in recent years (for both economics in general and economic analysis of law) is in the field of *behavioral economics*, which relaxes strong assumptions about rationality, willpower and self-interest in decision- and game-theoretic settings in order to make better predic-

tions. Since experiments have played an important role in demonstrating behavioral anomalies and provided raw stylized facts for inspiring new theories, and because some of these concepts are useful in understanding experimental results discussed below, it is perhaps worthwhile to briefly review that area of discourse here [see also Camerer, Loewenstein, and Rabin (2004); Jolls, Sunstein, and Thaler (1998); Jolls (2007)].

The central concepts in behavioral economics introduce psychological complications to traditional economic theories of choice over risk and time, social preferences toward outcomes of others, how equilibration occurs, and deviations in information processing from Bayes' rule (see Camerer, 2006).

One area of active research is risky decision making, where the predominant mode used in law and economics is expected utility (or EU) theory. EU theory, which traces its roots at least as far back as von Neumann and Morgenstern (1944) [and in one form, to Bernoulli (1738)] has been one of the foundational bedrocks of economics and game theory since the mid-twentieth century. The EU approach represents preferences analytically for an individual who faces a lottery Y, which pays her (possibly vector valued) allocation y_i with probability p_i, where $i \in \{1, \ldots, M\}$ indexes the differing payoff contingencies. If the agent's preferences over lotteries (1) constitute a complete ordering; (2) are continuous; and (3) are independent of common consequences then her preferences in each contingency can be represented by a scalar utility function u, so that her expected utility is given by:

$$U(p, y) = \sum_i p_i \cdot u(y_i) \tag{1}$$

Moreover, it is common in rational choice theory to extend the EU framework to dynamic settings. If the outcome of each lottery Y represents non-durable consumption allocations, then such an extension is relatively simple. Consider, for example, a variation on the above framework in which, in period t, the individual receives an allocation y_{it} with probability p_{it}. Under the same assumptions as above, if along with an additional assumption that (4) the individual's utility at time t is invariant of her utility at some time before t, then her utility over consumption bundles is time consistent and invariant, and expression (1) reduces to:

$$U(p, y, t) = \sum_t \sum_i \delta^t \cdot p_{it} \cdot u(y_{it}), \tag{2}$$

where δ corresponds to a (time consistent) personal discount factor for the individual.

Much of experimental economics is devoted to calibrating parameters within the utility framework offered above. However, behavioral economists have endeavored to expand the above formulation to include other factors that are not generally within (nor always consistent with) the conventional EU framework. Under a generalized expected utility (GEU) framework capturing these elements, expression (2) might have the following structure.

$$U(p, y, t) = \sum_i \pi(p_{i0}) \cdot u(y_{i0} - z_0) + \sum_{t>0} \sum_i \beta \cdot \delta^t \cdot \pi(p_{it}) \cdot u(y_{it} - z_t) \tag{3}$$

Equation (3) replicates the principal components of (2), but adds to it a number of other parameters that behavioral economists often wish to test. The parameter β represents an added multiplicative discount factor reflecting the possibility that the individual exhibits an exaggerated preference for immediate outcomes (sometimes called "present bias") discounting any future payoff with an additional discount factor β (Laibson, 1997). A parameter $\beta < 1$ will generate a dynamic inconsistency, because future rewards at time T receive little weight in current decision making, but effectively receive a boost when time T actually arrives (because the weight β is no longer is applied). In addition, the above formulation reflects the possibility that the agent utilizes subjective probability assessments $\pi(p)$ that diverge from the objective (or "true") probabilities, and also may not be updated in a manner consistent with Bayes' rule. Finally, expression (3) reflects the possibility from prospect theory [Kahneman and Tversky (1979)] that the agent may set a baseline reference point (z_t), assessing utility based on whether her realized level of consumption falls short of or exceeds that baseline. In this approach, descriptions of a gamble's possible outcomes, which are "framed" as differences from various reference points, can lead to choices that depend on the reference point ("framing effects").[9] In prospect theory the utility function is thought to exhibit a disproportionate aversion to losses, compared to equal-sized gains, a property called "loss-aversion" ($u(x) < -u(-x)$ for $x > 0$ in the original formulation; cf. Koszegi and Rabin, 2006). Note that expression (3) is not as much of a departure from conventional EU theory as it is a generalization of it. Indeed, EU theory is simply a special case within the above framework. A significant portion of experimental research in economics is now devoted to attempts to calibrate a more general model, such as that illustrated above, and explore its implications theoretically and to explain field data and business practices.

In the realm of strategic thinking, a workhorse concept is equilibrium—all players choose an optimal response given beliefs about the others' choices which are accurate. Behavioral game theorists have weakened the concept of equilibrium in two directions. One direction is "quantal response equilibrium" [QRE, e.g., McKelvey and Palfrey (1998); Goeree and Holt (2001)], which assumes that players deviate from best responses, but make large deviations less often than small deviations, and that players are aware of the propensity for deviations. A different direction is cognitive hierarchy (CH) models in which players may iterate through varying numbers of steps of strategic thinking [e.g., Camerer, Ho, and Chong (2004)]. Both of these generalizations of the standard concept of equilibrium can potentially explain when simple error-free equilibrium models predict accurately, and when they do not. The central difference is that in both the QRE and CH approaches, strategies that would not be played in a conventional equilibrium are sometimes played, and even small amounts of such behavior might influence even rational players. These theories therefore provide a tool to investigate when modest

[9] Note that expression (3) could be generalized even more. For example, the reference point z_t could systematically vary over time or by contingency.

departures from rationality will either become magnified or become erased by strategic interaction with other agents. Neither approach has yet been applied to law, however.

3. Applications

We now turn to consider specific applications of experimental approaches within salient legal settings. While time and space considerations prevent us from sampling the entire field, we highlight below those applications that (in our estimation) have proven to be particularly valuable for the study of law and economics. Our first set of applications considers the experimental analysis of bargaining and the Coase theorem, and how economic incentives, information, and entitlement structure affect the ability and proclivity for individual actors to reallocate their rights optimally. The second set of applications concerns experimental studies of the litigation and settlement process—an area that has shed considerable light on our understanding of the processes that select cases for litigation. Finally, we consider experimental analyses of jury and judge behavior. In each of these contextual applications, experimentalists have unambiguously informed the state of academic inquiry, and in many cases experimental approaches have played a central role in spawning affiliated theoretical or empirical literatures.

3.1. Contracting, legal entitlements, and the Coase theorem

It is difficult to imagine a precept of law and economics that is more central than the much-heralded Coase theorem [Coase (1960)]. The theorem (in at least one version of its various forms) posits that in the absence of significant transaction costs, the underlying manner in which legal rights are allocated is unimportant for efficiency purposes, since self-interested parties will tend to reallocate rights efficiently through bargaining. Coasean logic is, in fact, foundational to all contracting, and to the widespread modeling principle in economics that efficient organizations and institutions will thrive. It is therefore not surprising that experimental approaches in law and economics have been particularly focused on contracting behavior.

Perhaps contributing to the interest in experimental tests of the Coase theorem is the concept's own fluidity. As a conceptual premise, the Coase theorem is as easy to understand as it is difficult to apply. Indeed, as Coase himself recognized, the most interesting cases are those in which transaction costs are significant, and bargaining in the shadow of the law need not be efficient. Experimental law and economics scholars have devoted considerable attention to numerous manifestations of the Coase theorem, in part to identify what those cases are. Although their contributions are too numerous to catalog here, we can at least provide a sampling.

3.1.1. Simple bargaining environments with perfect information

Perhaps the most natural experimental environment in which to test the predictions of the Coase theorem is in a simple two-party bargaining framework. Early pioneering

work in the lab [e.g., Hoffman and Spitzer (1982, 1985)] largely confirms the zero-transaction costs predictions of the Coase theorem in simple experimental settings. The Hoffman/Spitzer experimental design involved a designated "controller" who could unilaterally choose among various social allocations of money, which differed both in their aggregate level of compensation and in their distribution. Each allocation was simply a line in a table stating how much the controller and the other player got, and the total of those payments. The Coasean prediction is that the line with the highest total would be chosen, and the controller would demand a sidepayment making her table payment plus sidepayment equal to the highest sum she could receive by acting unilaterally. The experimenters varied the determinants of how one was designated a controller, ranging from simple random designation to an "earned" right (earned by winning a series of backwards induction games). Finally, the parties were allowed to contract for sidepayments prior to the controller's choice of social allocation. The authors found that while the method for determining the controller significantly affected distributions between the parties, nearly all the dyads were able to contract into the socially optimal outcome. However, when the controller designation (a/k/a property right) was allocated randomly, the controllers generally did not demand a large sidepayment (earning less than they could be acting unilaterally). This work (and various follow-on efforts) gives reason to be sanguine about the ability of parties to overcome endowment effects, but also suggests a role for fairness in the distribution of gains from efficiency.

3.1.2. Private information

One factor that early Coase theorem experiments described above did not attempt to control for was the information structure of the bargaining environment. This omission is potentially significant, since it is well known from the theoretical bargaining literature [e.g., Myerson and Satterthwaite (1983)] that private information leads to generic inefficiencies. However, experiments by Radner and Schotter (1989) show that privately informed agents generally trade more often than predicted by theory, so the inefficiencies are smaller than predicted, particularly when bargaining is conducted face-to-face rather than through computer interfaces.

More recent experimental work [e.g., McKelvey and Page (2000)] has confronted the challenge of studying private information more directly, testing the Coase theorem in contexts where asymmetric information pervades the bargaining environment. McKelvey and Page (2000), for example, find that property rights can be "sticky" under asymmetric information, in that efficiency-enhancing transfers from low valuers to high valuers frequently fail to be consummated unless the mutual gain is significantly greater than zero. This finding is largely consistent with the predictions of asymmetric information game theory.

In a related twist on this approach, Croson and Johnston (2000) studied Coasean bargaining experiments involving asymmetrically informed parties, but the experimenters varied the way that the underlying legal entitlement was *protected*. For some subjects, a legal entitlement was protected by a property rule (which gives its owner the right

to enjoin conflicting use by others). For other subjects, the right was protected by a less certain rule (such as a probabilistic entitlement or a liability rule that allows the nonowner to appropriate the right non-consensually in exchange for paying damages). Croson and Johnston (2000) find that partial entitlements of this type (particular uncertain ones) to lead to more efficient bargaining outcomes than strong property rights, a prediction that is consistent with rational choice theory predictions with asymmetric information [Ayres and Talley (1995); Johnston (1995)].

3.1.3. Endowment effects

During the period in which the Coase theorem experiments were ongoing, research motivated through law and psychology was exploring a related phenomenon that has come to be known as the "endowment effect," a term that embodies experimental findings about how *ex ante* possession appears to affect valuation decisions. Explicitly, the endowment effect reflects experimental (and empirical) evidence that the maximum amount a person would be willing to pay to procure a good is often significantly less than the minimum amount she would be willing to accept to part with the same good, in contrast with most assumptions of the rational actor model that mere possession does not affect value. It is a phenomenon that has been detected in numerous experimental settings [for a relatively recent meta-analysis, see Horowitz and McConnell (2002)].

The endowment effect was originally studied by prospect theorists in economics and psychology [e.g., Kahneman, Knetsch, and Thaler (1990)], inspired by buying-selling price gaps observed in contingent valuations of non market goods (such as hunting licenses). Legally oriented experimental scholars soon recognized (and exploited) its relevance as well. Indeed, the existence of significant endowment effects may have important implications for both our positive understanding of legal rules and how such rules should be designed optimally [Cohen and Knetsch (2002)]. Most generally, because the endowment effect retards efficient trade, it is more incumbent on legal orderings to calibrate allocations efficiently from the outset. As such, considerably more thought would have to go into determining how to set default rules in contracting [Korobkin (1998)], whether to protect entitlements with strong or weak protections [Rachlinski and Jourden (1998)], and in whom to vest entitlements to begin with.

The set of challenges introduced by the endowment effect is made more intriguing by the fact that it appears to be a fairly context-dependent phenomenon. In particular, the effect appears *most* pronounced in situations where the entitlement at question has few market substitutes [Shogren et al. (1994)], when subjects believe their entitlement was the result of merit rather than luck [Loewenstein and Issacharoff (1994)]; and when subjects have had little practice or other familiarization with the experimental design [Plott and Zeiler (2005)]. A theory of reference point formation by Koszegi and Rabin (2006) suggests that the anticipated trading environment establishes a reference point, which influences preferences through reference-dependence, and creates choices that fulfill the anticipation in a "personal equilibrium." Similarly, Heifetz, Segev, and Talley (2007) develop a theoretical model suggesting that a confrontational content may

enhance proclivities towards bargaining "toughness." Both approaches help capture the idea that experienced traders who plan to sell goods don't feel a loss when selling them, and hence show no endowment effect [List (2003)], as predicted by Kahneman et al. in their original 1990 article.[10] A variant of this sort of theory may help to explain why subjects tend not to display endowment effects in some corporate agency-cost frameworks [Arlen, Spitzer, and Talley (2002)].

The contextual caveats noted above are important for legal policy makers, since exporting lessons from the laboratory to the outside world is often a hazardous business. Nevertheless, there appear to be many legal environments where the endowment effect is a plausible and important phenomenon. A principal challenge for experimental economists interested in legal applications of the endowment effect, then, is to formulate more sophisticated theories of the circumstances under which context matters.

A second challenge in this area that has yet to be significantly developed is the intersection between theories of asymmetric information and endowment effects in legal bargaining contexts. Many of the existing contributions in one field or the other can be justifiably criticized for concentrating too myopically on one account or the other, without attempting to discern between them. This is particularly problematic, since many of the situations in which the endowment effect is most pronounced correspond precisely to situations where private information is likely to be a factor.[11]

3.2. Litigation and settlement

Another robust area for experimental research in law and economics has been on the litigation process itself. The process of trials is the unique foundational feature of the law as an institution, and it should therefore not be surprising that it has garnered attention from not only legal scholars but also from other social scientists interested in strategic interaction.

Perhaps the longest standing line of research in the trial process was begun by psychologists interested in the effects of optimism and "self-serving bias" in affecting decisions within risky environments. Building on pioneering work in psychology, a number of legal, economics, and psychology scholars have explored the degree to which individual litigants appear to form expectations about trial in a manner that favors their own case. In perhaps the most familiar set of experiments [Babcock et al. (1995); Babcock,

[10] Most experimental evidence suggests that the endowment effect is not present when the underlying right is solely or principally a store of value. A few experiments, however, have detected an effect when the underlying value is itself uncertain [e.g., Van Dijk and van Knippenberg (1996)]. Kahneman, Knetsch, and Thaler (1990, p. 1328) clearly anticipated this effect of experience, noting that "there are some cases in which no endowment effect would be expected, such as when goods are purchased for resale rather than for utilization."

[11] For example, as Shogren et al. (1994) demonstrate, the endowment effect is strongest in situations where there are few if any ready market substitutes for a good. Similarly, the strategic importance of private valuation is substantially reduced (and often eliminated) when there are numerous market substitutes for a good in question.

Loewenstein, and Issacharoff (1997); Babcock and Loewenstein (1997)], subjects were instructed to act as bargainers, and read a narrative involving a personal injury dispute between two potential litigants, in which damages were known to be $100,000 but liability was in doubt. The experimental protocol is notable because it used thirty-one pages of transcript, depositions and exhibits from an actual personal injury case (a car hitting a motorcyclist), rather than an abstract description. While a group of control subjects read these materials without knowing which side they would ultimately represent, a treatment group read the materials after being informed they would ultimately represent either the plaintiff or defendant. Information was then elicited from all subjects about (a) what amount of money was likely to be the most "fair" to remunerate the plaintiff; and (b) what amount of money they predicted would be awarded by a real judge (who had read the facts previously and who had issued an independent judgment[12]). Finally, the parties were afforded an opportunity to settle their case during a thirty-minute period that preceded the non-consensual imposition of an outcome. Delay was costly because each five-minute period that passed cost both parties some money.

In virtually all permutations of this experimental setting, subjects who were informed of their role beforehand exhibited economically and statistically significant differences from the control group. For example, treatment subjects differed from control counterparts both in their assessment of the fair outcome *and* of the judge's ultimate decision by approximately $18,000, while control subjects actually exhibited no difference (or even a *negative* difference). The amount of the self-serving gap in expectations about the judge's decision was correlated with the propensity to settle, and the length of time to settlement. More to the point, control subjects exhibited a significantly higher settlement rate, and a significantly faster time to settle when they could reach an agreement.

More recent experimental efforts to test the robustness of this finding have provided evidence that self-serving biases are also present when the extent of damages (rather than liability) serves as a source of potential disagreement. Babcock and Loewenstein (1997), also find that the same qualitative variety of self-serving bias also can be found in attorneys and experienced negotiators (albeit sometimes in a smaller magnitude). In other recent studies, Babcock and Pogarsky (1999, 2001) find that it is possible to manipulate the "gap" between plaintiffs' and defendants' assessments of a case by imposing a damages cap that constrains the range of feasible settlements. Additionally, the authors find, quite interestingly, that the imposition of an exceedingly *generous* cap on damages had the reverse effect, increasing disagreement and discouraging bargaining. This finding suggests a possible interaction between litigant "optimism" on the one hand, and anchoring effects on the other (a cognitive behavior discussed more fully below).[13] Finally, at least some preliminary work has been done exploring processes by which procedures may de-bias litigant optimism. For instance, Babcock, Loewenstein,

[12] Subjects in these studies were awarded monetary prizes for the accuracy of their predictions about the judge's behavior.

[13] See Subsection 3.3, infra.

and Issacharoff (1997) find that a simple but effective de-biasing instruction is to urge litigants to consider the weaknesses in their own case or the real possibility that the judge may rule for the other side. Interestingly, urging litigants to consider the strengths of the other side's case did not work (and sometimes backfired).[14] Debiasing (and reduction in costly delay) also occurs when subjects are assigned to their bargaining role *after* reading the case materials, which implies that the self-serving bias is created by encoding of the case facts as they are being digested, rather than by the role assignment per se. Still few experimental results exist, however, mapping out the robustness of de-biasing mechanisms.

The literature on self-serving biases in litigation plays a particularly interesting role in providing experimental support for a key theoretical account within law and economics about how cases proceed to litigation. The famous Priest and Klein (1984) hypothesis about litigation posited a theoretical model in which relatively "optimistic" litigants fail to settle their cases, but pessimistic ones litigate. The self-serving bias literature presents credible evidence for a version of the Priest-Klein hypothesis in which *all* litigants exhibit some optimism, although some exhibit more than others. Interestingly, many of the same predictions from their theoretical mode (such as the well known 50% plaintiff victory prediction at trial) can emerge from models of self-serving biases. This may be an area where theoretical work by law and economics scholars helped to motivate later research by non-legal scholars about litigation behavior.

An approach complementary to that reflected in the self-serving bias literature for analyzing suit and settlement comes from the theoretical literature positing that information asymmetries retard settlement [e.g., Reinganum and Wilde (1986); Spier (1994)].[15] Under this approach, some parties act as tough bargainers not because they are overly optimistic, but rather because they possess proprietary information about the strength of their case. Similar to the endowment effect literature, the experimental results in self-serving bias literature may, in part, embody some aspects of private information. For example, subjects pre-informed of their hypothetical roles may selectively search for facts that support their client's case, glossing over those that are either neutral or support the other side's case. Entering negotiation, then, each side may have some informational advantages over the other. To date, there appear to be few experimental designs that are capable of disaggregating informational from cognitive impediments to settlement.[16]

[14] A likely explanation for why this treatment did not work is that subjects generated half-hearted lists of strengths in the other side's case, then were unimpressed by these lists.

[15] While information asymmetries are undoubtedly important, note that the debiasing effect of assigning bargaining roles after reading case facts in Babcock, Loewenstein, and Issacharoff (1997) shows clearly that the creation of self-serving bias while encoding information is important too. That is, two parties reading the same case facts with different roles in mind can create a *perception* asymmetry (like fans rooting for opposite teams watching a sports event both thinking the refereeing is biased against them) which will have similar implications to a true information asymmetry.

[16] A few experimental studies find results consistent with the informational account of settlement failure in analyzing the English versus American rules on fee shifting. See Coughlan and Plott (1997).

As noted earlier, one important use of experiments is to help design or evaluate how well legal institutions work, in settings where we can control for endogeneous adoption and evaluate efficiency directly. Babcock and Landeo (2004) did bargaining experiments to examine the effects of pre-settlement escrows [as studied by Gertner and Miller (1995)]. In their experiments a plaintiff learns of a damage amount drawn from a commonly-known distribution; in the interesting case the defendant knows only the distribution. Both plaintiff and defendant make secret settlement offers. If the offers overlap they settle immediately; otherwise two rounds of bargaining proceed, incurring fixed costs. If no settlement is reached they "go to trial" and the plaintiff is awarded the actual damage. They found that the escrow mechanism increases settlement from 49% to 69% and reduces pre-trial legal costs by about half. Furthermore, the results are roughly consistent with many numerical predictions of the Gertner-Miller model.

In another recent effort, Babcock and Landeo (2004) analyze how the imposition of award splitting between the state and a civil plaintiff (an increasingly popular type of tort reform) affects settlement, incidence of trial, and injurer precautions in an experimental setting. The authors attempted to test a numerical parameterization of the model developed by Landeo and Nikitin (2005), which predicts that the imposition of such reforms should reduce levels of care, increase injury rate, reduce incidence of litigation (conditionally and unconditionally), and increase settlement. Contrary to the theoretical predictions, the authors found that treatment subjects operating under split award reforms did not exhibit care levels appreciably different from a control group. On the other hand, once an injury occurred, settlement rates increased and litigation expenditures decreased, in accordance with theoretical predictions. The authors posit that the complexity of the relationship between care taking and litigation may dampen much of the *ex ante* feedback effects that procedural tort reforms could generate in theory. This is a potentially important observation, as it suggests that certain types of procedural tort reforms may be effective in reducing the costs of litigation without significantly sacrificing care and precaution levels.[17]

3.3. Adjudication, jury behavior and judge behavior

Another fertile area of research for experimental law and economics scholars has been in the behavior of juries and judges. These domains are especially important for studying the effect of rationality limits on aggregate behavior, because a small number of biased individuals could disproportionally influence the collective outcome, depending on the jury communication and voting rules and how judges are elected and evaluated. Furthermore, concerns about generalizability are muted in experiments like these because it is not hard to construct an artificial jury-like environment which rather closely resembles actual deliberations among jurors in naturally-occurring cases (compared to

[17] Obviously, it would be desirable to understand the robustness of this result before using it as a basis for tort reforms.

concerns about students told to "pretend you are a judge," or of simple experiments to evoke the dramatic emotions present in litigation).

Unlike each of the foregoing topic areas, however, the study of judges and juries has a significantly more disorderly nature to it. While there are a number of interesting individual findings in this area, they are more difficult to weave into a larger, unified descriptive tapestry. We therefore content ourselves merely with describing some of the more thought-provoking contributions that exist here.

3.3.1. Jury behavior, voting, and social pressure

Game theorists have long been interested in jury decision-making, and in many instances theoretical contributions in the field have inspired subsequent efforts by experimentalists to test the theory in the laboratory. In one notable example of this trend, a relatively recent literature (emanating largely from economists and political scientists) has called into question the relative wisdom of unanimous vote requirements in the jury setting. In a well-known article, Feddersen and Pesendorfer (1998) argue that strategic voting by juries may lead to *more* convictions under a unanimity rule than under a majority rule, since individual jurors under a unanimity rule may feel particularly hesitant to be the pivotal hold-out in a vote [Feddersen and Pesendorfer (1998)]. Others [such as Coughlan (2000)] have challenged this assertion, noting that most juries engage in deliberation, and through such pre-vote deliberation juries are significantly more likely to transmit important information to others, thereby decreasing the chances of such perverse predictions. A recent experimental study by Guarnaschelli, McKelvey, and Palfrey (2000) attempts to test that claim in an experimental setting of hypothetical jurors. They find, consistent with the Fedderson and Pesendorfer thesis, that juries are more likely to vote strategically under a unanimity rule; however, the aggregate magnitude and nature of this strategic distortion does not appear consistent with the Nash prediction of higher conviction rates under a unanimity rule.

The tendency of juries to vote strategically has interesting parallels to the early pioneering work in psychology on social pressures to conform. Solomon Asch's well-known experiment in which confederates are successfully able to influence a minority to state an obvious falsehood, while not directly about law, has been cited by many as an example of the non-Bayesian biases that juries often face [see Asch (1956); Kuran and Sunstein (1999)]. Models of "rational conformity" or cascades, due to inferring information from behavior of others, could be applied to these settings as well. To date, however, there have been few efforts by experimentalists in law and economics to incorporate this phenomenon into a broader analysis of strategic voting.

3.3.2. Hindsight biases

Another important cognitive heuristic that has been shown to be important in legal settings is hindsight bias: the tendency for individual decision-makers to be overconfident about the *ex ante* predictability of a particular event, after knowing that such an event

did in fact come to pass. In more pedestrian parlance, the hindsight bias is roughly akin to the practice of "Monday morning quarterbacking"—proclaiming foreknowledge of a solution to a difficult problem that has since become obvious in hindsight. Motivated by early experimental results in psychology [e.g., Fischhoff (1975)], a number of legally oriented scholars developed experimental settings to determine whether jurors are also subject to hindsight biases [e.g., LaBine and LaBine (1996); Kamin and Rachlinski (1995)]. These experiments (which sometimes involved financial incentives) asked jurors to assess whether a particular act of nonfeasance constituted negligence. Those in the control condition were given information about the relative costs and benefits of action, while those in the treatment condition were given the identical set of facts, along with information that harm had actually occurred subsequent to the defendant's nonfeasance. In each study, subjects in the treatment group were consistently much more willing to find the existence of negligence than those in the control group.

Prescriptively, the finding of a significant hindsight bias among jurors may provide some basis for a number of concrete institutional responses, such as altering the legal nature of negligence, obviousness, or other legal standards that turn on retrospective assessment of *ex ante* likelihood. However, even if such bias exists, little is known about how to minimize or eliminate the bias with procedural changes in court. In this respect, some experimental findings [e.g., Viscusi (1999, 2001)] indicate that judges are substantially less susceptible to hindsight biases than are jurors, suggesting that encouraging bench trials may be an opportune way to reduce the hindsight bias from legal proceedings. This conclusion, however, is not free from debate, and other studies [e.g. Guthrie, Rachlinski, and Wistrich (2001)] find that judges are also susceptible to hindsight biases (albeit perhaps to a lesser degree than jurors). Additionally, there may be simple forms of jury instructions that have the effect of diffusing (or at least dampening) the bias. For example, one study finds that a warning to jurors to avoid second guessing the defendant's actions or being "Monday-morning quarterbacks" substantially reduced the prevalence and degree of hindsight bias in subject jurors [Stallard and Worthington (1998)].

3.3.3. Anchoring

Finally, as noted above, at least some of the experimental literature on attorneys has touched on a phenomenon known as the "anchoring effect," a behavioral regularity that is more routinely highlighted in jury experiments. Anchoring refers to the process by which an individual decision maker adjusts insufficiently from a reference point that she subsequently uses as an initial starting point or anchor for arriving at a final decision. The effects of anchoring appear to be especially strong in the context of damage awards given by juries. These decisions are typically made complex by the lack of familiarity that juries have with the range of damages that are appropriate within a given class of cases. The lack of an "upper support" on the set of damages can invite manipulations by interested parties to distort jury decision-making. For example, it is now well documented in a number of experimental studies that both statutory damage caps

as well as specific monetary requests made by plaintiff attorneys can provide an anchor from which jurors may work, in awarding both compensatory damages and (in particular) punitive damages. Higher anchors can lead to higher eventual damages [Hastie, Schkade, and Payne (1999); Robbennolt and Studebaker (1999)]. And, while anchoring appears to be predominantly a danger for juries and courts, as noted previously litigants themselves may also be susceptible to the effect [Babcock and Pogarsky (1999)]. As with the hindsight bias, anchoring effects appear to be more pronounced with lay juries than with professional judges. Keep in mind that this observation does not imply that anchors have little effect simply because more highly trained professionals are more impervious to them, because jury trials are still largely the norm.

3.4. Legal rules and legal norms

One fruitful application of experimental methods to legal applications considers how legal rules and extra-legal behavior norms interact with one another (see McAdams and Rasmusen, 2007, in this Handbook, Chapter 20). Mitchell (1999) gives a useful (though typically open) definition of a norm: "... norms tell us what we should do under a given set of circumstances and are therefore obligatory upon those who wish to participate in the society which is at least partly constituted by such norms." As they might be translated into economic language, a norm is a social rule which people prefer to follow—through internalized guilt, or because of external social sanctions that result from norm violation.

Most conventional accounts of legal rules in economic analysis view legal rules as instrumental devices to change incentive structures, and in so doing, induce different types of strategic behavior. However, another important (and under-analyzed) role that legal rules can play is to affect the extent how (and when) law interacts with non-legal (and non-contractual) forms of incentives. A law that is rarely enforced, or is costly to enforce, such as a speed limit on an unpatrolled stretch of highway, can therefore create a norm which people obey simply because norms are supposed to be obeyed.

Perhaps the simplest domain for investigating this relationship is in simple game-theoretic frameworks. Bohnet and Cooter (2001), for example, devised a set of three simple multi-person coordination games (repeated numerous times) in order to test the interaction effects of legal rules and other norms of behavior. Of the three strategic settings explored, the first two had unique Nash equilibria, while the third was a pure coordination game with multiple equilibria. In each setting, the experimenters imposed a small, probabilistic legal sanction on a treatment group for individuals who deviated from a prescribed strategy profile (which corresponded in all cases with the socially efficient profile). The size of the sanction was both modest, and, in the case of the treatment group, was offset precisely by an altered fundamental payoff structure to the game that equalized the *expected* payoffs from each strategy to those of a control group (which faced no sanctions). The authors found that within the first two strategic settings, the sanction had no discernible effect on the proportion of subjects who played the prescribed strategy. However, in the pure coordination game, the sanction induced

significantly more subjects to converge to the efficient profile, and to do so more rapidly. These results are consistent with the view that the behavioral aspirations "expressed" by legal rules can provide a form of "focal point" [Schelling (1960)] that can facilitate a coordinated choice among available equilibria.

In a similar vein, McAdams and Nadler (2005) tested the effects of expressive law in less cooperative version of a coordination game: the "hawk-dove" game (also called "chicken"). In a typical version of the hawk-dove, players can both play the passive "dove" strategy and each earn 1; or if one plays hawk and another dove, the total of 2 goes entirely to the hawk player and the dove players gets 0. If both play hawk they each get −1. There are two efficient equilibria, in which one player chooses hawk and another dove and they divide the total of 2 unevenly. Rather than testing the effects of small legal sanctions, the authors used a type of "dignitary" manipulation involving a public expression of either (1) an efficient pure-strategy profile, selected randomly, or (2) a designation of one player as a "leader," which indirectly made more salient the leader's preferred pure strategy equilibrium profile. The authors found that either type of expression substantially altered the ability of the parties to realize payoffs on the Pareto frontier (although the designated leader manipulation had the strongest effects).[18]

These experiments show the capacity of law to act as a coordinating or correlating device to create "psychological prominence" (in Schelling's apt phrase) of one equilibrium over another, when there are multiple equilibria. Another potential expressive role of law is not as a coordinating device to lead to selection of one among many equilibria, but rather as a norm activating device that alters equilibrium play away from any (conventionally conceived) equilibrium. Although the studies discussed above do not appear to find evidence of such an effect, some recent work has generated counterexamples that warrant additional attention. Tyran and Feld (2006), for example, study a game of public goods provision in which, in the absence of intervention, self-interested players would tend to free ride off one another, overinvesting in private goods relative to public goods. Against this baseline, the authors introduce two types of exogenous legal sanction against free riding—one in which the monetary sanction is so large as to make full contribution a dominant strategy; and one in which the monetary sanction is mild, so that (at least in theory) free riding is still a dominant strategy. In addition to that manipulation, however, Tyran and Feld (2006) introduce a manipulation along one other dimension: they allow subject cells to choose (by constructed majority vote) whether to adopt a particular sanction (i.e., mild or severe), thereby giving rise to the *endogenous* imposition of law.

The authors find that, relative to respective control groups, an exogenously imposed legal sanction tends to deter free riding (above the baseline case) only when it is relatively severe. Mild sanctions do not appreciably enhance relative contribution rates.

[18] McAdams and Nadler actually used two manipulations in their designated "leader" condition. Under the first, the leader was chosen randomly. Under the second, subjects were told that the leader was chosen on the basis of "merit"—performance on a simple pre-test given by the authors. The differences between these two manipulations were statistically (and economically) insignificant, however.

On the other hand, when legal sanctions are endogenously imposed, contribution rates under both a mild and severe sanction increase sharply (relative to the exogenous case), and are appreciably higher than in cases where no law is present. These results, the authors argue, appear consistent with the view that legal rules—if viewed as having legitimacy—can shape norm activation beyond simply providing focal points. Endogeneity conveys legitimacy. Much more can be done in this area to test the robustness of such results, and the likely nuances of what constitutes a legitimate (and hence, obeyed) legal rule.[19]

It is important to note that much of the work described above comes against the backdrop of a large literature outside of law in which norms of cooperation and reciprocity frequently emerge in an experimental setting, contrary to traditional rational actor predictions. Fehr, Fischbacher, and Gachter (2002), for example, report on experiments uncovering a "strong reciprocity" norm among parties in anonymous interactions with one another. In a series of experiments, they document numerous instances in which experimental subjects are willing to spend personal resources to reward observed good behavior and punish bad behavior, even when such behavior did not directly affect them, but some third party. Indeed, Tyran and Feld's study (described above) finds that contribution rates are significantly above zero (the rational choice prediction) in the absence of law. [For a more comprehensive review of similar results outside legal applications, see Ledyard (1995).]

A promising area of inquiry in the area of law and norms concerns whether law and norms function as complements or substitutes for one another. Does the introduction of law (exogenous or endogenous) tend to amplify existing legal norms or crowd them out? Although research in this area is still relatively sparse, Bohnet, Frey, and Huck (2001) present a notable study of the interaction between legal and non-legal enforcement in a contractual setting involving moral hazard. Varying exogenously the intensity of expected legal sanction associated with shirking, the authors find that intermediate to severe sanctions tend to act as substitutes for norms of cooperation, inducing behavior that more closely corresponds with rational actor theories. On the other hand, they found that mild forms of legal enforcement tend to catalyze norm activation, sometimes inducing relatively large rates of cooperation. In this context, then, it would appear that law and norms operate principally as substitutes, but can be complements over certain domains. There are also notable parallel concerns in experimental organizational economics. Inside firms, the question is whether creating "high-powered" performance-based incentive contracts—which are thought to motivate lazy workers who are naturally likely

[19] For example, Tyran and Feld (2006) attempt to control for possible selection effects by exploring whether the endogenous vote creates biased populations of habitual cooperators (who are likely to be "yes" voters) and habitual non-cooperators (who are likely to be "no" voters). They find that even dissenters tend to adopt strategies mimicking those in the majority, regardless of whether the dissenter voted yes or no. Nevertheless, other parts of the results remain curious: such as the fact that contribution rates appear significantly lower in the endogenous conditions with no law than in the exogenous conditions with no law.

to shirk—actually "crowd out" or negatively substitute for homegrown reciprocal motivations which would lead to minimal shirking even absent explicit penalties [e.g., Fehr and Falk (2002)]. In a sense, firm contracts can be seen as expressions of "internal law" and so there may be some fruitful parallels between expressive role of firm contracts and cultural norms, and the expressive role of societal law.

4. Looking ahead

Although the universe of experimental studies of law is now becoming sizeable and is still growing, there is still much "low hanging fruit" for legally oriented experimentalists to harvest. Although we have neither the time nor space in this essay to offer an exhaustive inventory, future research possibilities include both methodological and substantive dimensions, which we briefly explore below.

Methodologically, as noted above, few experimental studies in law attempt to discern how subjects behave when repeating experiments. Not only would such information convey significant information about how learning, acclimation and experience affect performance, but it would also contribute to the power of experimental findings, by allowing researchers to make "within subject" treatment inferences and subject-based fixed effects that are not easily observable with strict cross-sectional analysis (see, e.g., Plott and Zeiler, 2005). Another promising methodological approach is to use internet-based experimental instruments to measure individual responses to experimental protocols. While this form of data collection has obvious drawbacks (such as a reduction in control and likely more selection bias among responders), it has the redeeming attributes of allowing experimenters to access a much broader cross section of subjects than is frequently available in university settings, and generating very large samples [see, e.g., McCaffrey and Baron (2005)].

In a similar vein, a rapidly emerging development in experimental economics is the use of field experimentation, in which some elements of control are imposed on a naturally occurring situation. In one study of this sort, Camerer (1998) placed large ($500–1000) bets at a horse racing track (and cancelled them at the last minute) to determine whether such activity would cause cascade effects in other bettors [e.g., Bikhchandani, Hirshleifer, and Welch (1992)]. Control came from the fact that which of two matched-pair horses were bet on was randomized by a coin flip.[20] Another cross-cultural field experiment used simple bargaining games in 15 small-scale societies to investigate the link between cultural practices and fairness norms [Henrich et al. (2004)].

Field experimentation is not completely foreign to the economic analysis of law. For example Ayres (1991) explored the prevalence of racial profiling in new car markets by

[20] Similarly, Lucking-Reiley (1999) created internet auctions for "Magic" playing cards with different auction structures to test predictions about the influence of reserve prices and other variables on bids. Control came from the fact that subjects who entered the different auctions presumably did not self-select which one to enter.

sending a racially diverse set of confederates into automobile dealerships to bargain for new cars using identical bargaining strategies and mannerisms. Using visits to identical dealerships to establish control, Ayres finds that racial minorities tended to pay both higher first offers and higher final prices than white males. Bertrand and Mullainathan (2004) conducted a similar "audit" study, sending employers resumes that were otherwise identical except for applicant names (which had either strong black or white associations), to investigate discrimination in hiring.

The principal attraction of field experiments is that they extend and target the reach of the experimental method to the very subject pools that one is most interested in studying: actual decision makers in the real world. However, a concomitant cost of this design approach (and one that draws frequent criticism) is the difficulty in procuring adequate informed consent from the target populations.[21] Consequently, researchers should generally tread carefully in designing field experiments that impose minimal burdens on their intended subject pools.

Substantively, experimental economists interested in law are in a particularly opportune position to make contributions about how law should respond to behavioral patterns that are frequently observed in experimental settings. Indeed, the legal milieu is one of the most elaborate and pervasive of contexts in which many (or even most) people interact. As noted above, we are still far from determining the contextual boundaries of a number of experimental findings (such as the endowment effect), and underlying questions about the generalizability of experimental findings still substantially hinders informed legal reform. More work by legal scholars can help to investigate the precise legal contexts in which such effects are largest, how they interact with the substantive underpinnings of legal rules and standards, and how legal rules may be best designed to avoid situations where individual decision making is likely to be untrustworthy. The "wind tunnel" analogy which has been used to guide design of actual policy in experiment economics for decades [Plott (1987)] is appropriate here. Most changes in law rest on a conjecture about the effect that changing the law will have. Experiments are a cheap way to weed out bad ideas by showing that the empirical conjectures embodied in them are wrong, or at least to shift the burden of proof to advocates of those changes.

Another area in which law-oriented experimental scholars are in a prime position to contribute to experimental law and economics is in the enterprise of discerning how (and whether) legal structure itself can help to de-bias individual decision-making [Jolls and Sunstein (2004)]. As noted above, for example, some studies have found that relatively simple manipulations can dampen—and in some instances eliminate—cognitive biases, such as jury instructions [e.g., Simon et al. (2001)][22] or the introduction of a fiduciary-like agency relationship [e.g., Arlen, Spitzer, and Talley (2002)]. Despite these isolated

[21] Harrison and List (2004) and Bertrand and Mullainathan (2004) provide more examples and a taxonomy of features of field experiments.

[22] Simon et al. (2001) find that simple "consider the other side" jury instructions can help to mitigate the effects and incidence of constraint-satisfaction reasoning (somewhat akin to cognitive dissonance).

findings, there is surprising little experimental work to date exploring on how legal institutions might play a role in accomplishing this task.

Doctrinally, a number of areas appear ripe for more study. For example, in the economic theory of law enforcement, a long-standing normative argument due to Becker (1968) argues that law should impose maximal fines in order to minimize enforcement costs. In application, however, this argument is confounded by a number of issues, including wealth constraints, risk aversion, the possibility of combining economic and non-economic sanctions, and the process by which potential victims, injurers, and enforcers "learn" about the underlying factors that affect their mutual interaction. In such settings, the simple Beckerian logic noted above may no longer hold (see, e.g. Garoupa and Jellal, 2004). Although parts of the literature described earlier are relevant to these questions, a more sustained experimental endeavor in this area is likely warranted. Other factors that might be included in such a program might include the attractiveness of strict liability versus fault-based doctrines, the effect of insurance and litigation costs, and the extent to which contractual relationships between the potential victim and potential injurer affects the ultimate incidence of such costs.

There are also doctrinal areas within the economic analysis of law that are now heavily theorized, and could use additional experimental calibration. For example, there is now a significant incomplete contracting literature analyzing how (and whether) negotiating parties choose to include express terms to cover future contingencies, how the parties invest in the relationship up until the time of performance, and how/whether they renegotiate the terms of their relationship (see Schwartz, 1998 for a review). Experimental approaches that test some of the core propositions from the incomplete contracts literature, while certainly not lacking now, still lag far behind the body of theoretical work. Moreover, empirical data on these sorts of contracting practices and behaviors is often difficult to obtain, thereby underscoring the role that experimental work can continue to play in this area.

Notwithstanding its relatively youthful pedigree, experimental methodology has come to occupy a position as a fully fledged component of economics proper. It is therefore hardly surprising, then, that experimental methods have also come to play an important role in studying legal institutions as well. Although the ultimate historical impact of experimental economics on the study of law has yet to be determined, experiments have already helped to create an economic understanding of law that is more descriptively accurate, and theoretically parsimonious and normatively nuanced than what would have been possible with theory and empirical methods alone.

References

Arlen, J., Spitzer, M., Talley, E. (2002). "Endowment effects within corporate agency relationships". Journal of Legal Studies 31, 1–37.

Asch, S.E. (1956). "Studies of independence and conformity: a minority of one against a unanimous majority". Psychological Monographs 70 (9), Whole No. 416.

Ayres, I. (1991). "Fair driving: gender and race discrimination in retail car negotiations". Harvard Law Review 104, 817–872.

Ayres, I., Talley, E. (1995). "Solomonic bargaining: dividing an entitlement to facilitate Coasean trade". Yale Law Journal 104, 1027–1117.

Babcock, L., Landeo, C.M. (2004). "Settlement escrows: an experimental study of a bilateral bargaining game". Journal of Economic Behavior and Organization 53, 401–417.

Babcock, L., Loewenstein, G. (1997). "Explaining bargaining impasse: the role of self-serving biases". Journal of Economic Perspectives 11 (1), 109–126.

Babcock, L., Pogarsky, G. (1999). "Damage caps and settlement: a behavioral approach". Journal of Legal Studies 28, 341–370.

Babcock, L., Pogarsky, G. (2001). "Damage caps, motivated anchoring and bargaining impasse". Journal of Legal Studies 30, 143–159.

Babcock, L., Loewenstein, G., Issacharoff, S., Camerer, C. (1995). "Biased judgments of fairness in bargaining". American Economic Review 85 (5), 1337–1343.

Babcock, L., Loewenstein, G., Issacharoff, S. (1997). "Creating convergence: debiasing biased litigants". Law and Social Inquiry 22, 913–925.

Ball, S., Cech, P. (1996). "Subject pool choice and treatment effects in economic laboratory research". In: Isaac, R.M. (Ed.), Research in Experimental Economics, vol. 6. JAI Press, Greenwich, CT, pp. 239–292.

Becker, G. (1968). "Crime and punishment: an economic approach". Journal of Political Economy 76, 169–217.

Bergstrom, T., Miller, J. (1999). Experiments with Economic Principles: Microeconomics. McGraw-Hill/Irwin, New York.

Bernoulli, D. (1738). "Specimen theoriae novae de mensura sortis". Papers Imp., Acad. Sci., St. Petersburg 5, 175–192.

Bertrand, M., Mullainathan, S. (2004). "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination". American Economic Review 94, 991–1013.

Bikhchandani, S., Hirshleifer, D., Welch, I. (1992). "A theory of fads, fashion, custom and cultural change as informational cascades". Journal of Political Economy 100, 992–1026.

Bohnet, I., Cooter R.D. (2001). "Expressive law: framing of equilibrium election?" Law and Economics, Working Paper Series, Paper 31.

Bohnet, I., Frey, B., Huck, S. (2001). "More order with less law: on contract enforcement, trust and crowding". American Political Science Review 95, 131–144.

Camerer, C.F. (1997). "Rules for experimenting in psychology and economics, and why they differ". In: Güth, W., Van Damme, E. (Eds.), Essays in Honor of Reinhard Selten. Springer-Verlag, NY, pp. 313–327.

Camerer, C.F. (1998). "Can asset markets be manipulated? A field experiment with racetrack betting". Journal of Political Economy 106, 457–482.

Camerer, C.F. (2003). Behavioral Game Theory. Princeton University Press, Princeton, NJ.

Camerer, C.F. (2006). "Behavioral economics". In: Blundell, R.M., Newey, W.K., Persson, T. (Eds.), Advances in Economics and Econometrics. Cambridge University Press, Cambridge.

Camerer, C.F., Hogarth, R.M. (1999). "The effects of financial incentives in economics experiments: a review and capital-labor-production framework". Journal of Risk and Uncertainty 18, 7–42.

Camerer, C.F., Ho, T., Chong, K. (2004). "A cognitive hierarchy model of one-shot games". Quarterly Journal of Economics 119, 3.

Camerer, C., Loewenstein, G., Rabin, M. (2004). Advances in Behavioral Economics. Princeton University Press, Princeton, NJ.

Camerer, C.F., Loewenstein, G., Prelec, D. (2005). "Neuroeconomics: how neuroscience can inform economics". Journal of Economic Literature.

Carson, R.T., Mitchell, R.C., Hanemann, W.M., Kopp, R.J., Presser, S., Ruud, P.A. (1992). A Contingent Valuation Study of Lost Passive Use Values Resulting From the Exxon Valdez Oil Spill. Attorney General of the State of Alaska, Anchorage.

Cason, T.N., Friedman, D. (1993). "An empirical analysis of price formation in double auction markets". In: Friedman, D., Rust, J. (Eds.), The Double Auction Market: Institutions, Theories, and Evidence. Addison-Wesley, Reading, MA, pp. 253–283.
Coase, R.H. (1960). "The problem of social costs". Journal of Law and Economics 3, 1–44.
Cohen, D., Knetsch, J. (2002). "Judicial choice and disparities between measures of economic value". Osgood Hall Law Journal 30, 737–770.
Coughlan, P.J. (2000). "In defense of unanimous jury verdicts: communication, mistrials, and strategic voting". American Political Science Review 94 (2), 375–393.
Coughlan, P.J., Plott, C.R. (1997). "An experimental analysis of the structure of legal fees: American Rule vs. English Rule". Caltech Working Paper. (http://econpapers.hhs.se/paper/cltsswopa/1025.htm.)
Croson, R. (2002). "Why and how to experiment: methodologies from experimental economics". University of Illinois Law Review 2002, 921–945.
Croson, R., Johnston, J.S. (2000). "Experimental results on bargaining under alternative property rights regimes". Journal of Law, Economics and Organization 16, 50–73.
Davis, D.D., Holt, C.A. (1993). Experimental Economics. Princeton University Press, Princeton, NJ.
Easley, D., Ledyard, J.O. (1993). "Theories of price formation and exchange in double oral auctions". In: Friedman, D., Rust, J. (Eds.), The Double Auction Market: Institutions, Theories, and Evidence. Addison-Wesley, Reading, MA, pp. 253–283.
Feddersen, T., Pesendorfer, W. (1998). "Convicting the innocent: the inferiority of unanimous jury verdicts under strategic voting". American Political Science Review 92, 23–35.
Fehr, E., Falk, A. (2002). "Psychological foundations of incentives". European Economic Review 46, 687–724.
Fehr, E., Fischbacher, U., Gachter, S. (2002). "Strong reciprocity, human cooperation, and the enforcement of social norms". Human Nature 13, 1–25.
Fischhoff, B. (1975). "Hindsight $\neq$ foresight: the effect of outcome knowledge on judgment under uncertainty". Journal of Experimental Psychology 1, 288–299.
Friedman, D., Sunder, S. (1993). Experimental Methods: A Primer for Economists. Cambridge University Press, Cambridge.
Garoupa, N., Jellal, M. (2004). "Dynamic law enforcement with learning". Journal of Law, Economics and Organization 20, 192–206.
Gertner, R.H., Miller, G.P. (1995). "Settlement escrows". Journal of Legal Studies 24, 87–122.
Gode, D.K., Sunder, S. (1997). "What makes markets allocationally efficient?" Quarterly Journal of Economics 112, 603–630.
Goeree, J., Holt, C. (2001). "Ten little treasures of game theory and ten intuitive contradictions". American Economic Review 91, 1402–1422.
Guarnaschelli, S., McKelvey, R., Palfrey, T.R. (2000). "An experimental study of jury decision rules". American Political Science Review 94 (2), 407–423.
Guthrie, C., Rachlinski, J., Wistrich, A. (2001). "Inside the judicial mind". Cornell Law Review 86, 777–830.
Harrison, G., Lesley, J.C. (1992). "Must contingent valuation surveys cost so much?" Journal of Environmental Economics and Management 31, 79–95.
Harrison, G., List, J.A. (2004). "Field experiments". Journal of Economic Literature 42 (4), 1–90.
Hastie, R., Schkade, D.A., Payne, J.W. (1999). "Juror judgments in civil cases: effects of plaintiff's requests and plaintiff's identity on punitive damage awards". Law and Human Behavior 23, 445–470.
Heifetz, A., Segev, E., Talley, E. (2007). "Market design with endogenous preferences". Games and Economic Behavior 58, 121–153.
Henrich, J., Boyd, R., Bowles, S., Gintis, H., Fehr, E., Camerer, C. (2004). Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies. Oxford University Press, Oxford.
Hertwig, R., Ortmann, A. (2001). "Experimental practices in economics: a methodological challenge for psychologists?" Behavioral and Brain Sciences 24 (3), 383–403.
Hoffman, E., Spitzer, M. (1982). "The Coase theorem: some experimental tests". Journal of Law and Economics 25, 73–93.

Hoffman, E., Spitzer, M. (1985). "Entitlements, rights and fairness: an experimental examination of subjects' conceptions of distributive justice". Journal of Legal Studies 14, 259–297.
Horowitz, J.K., McConnell, K.E. (2002). "A review of WTA/WTP studies". Journal of Environmental Economics and Management 44, 426–447.
Johnston, J.S. (1995). "Bargaining under rules versus standards". Journal of Law, Economics and Organization 11, 256–281.
Jolls, C. (2007). "Behavioral law and economics". In: Diamond, P., Vartainen, H. (Eds.), Behavioral Economics and Its Applications. Princeton University Press, Princeton, NJ.
Jolls, C., Sunstein C. (2004). "Debiasing through law". University of Chicago Law and Economics, Olin Working Paper No. 225.
Jolls, C., Sunstein, C.R., Thaler, R. (1998). "A behavioral approach to law and economics". Stanford Law Review 50, 1471–1550.
Kagel, J.H., Roth, A.E. (1995). Handbook of Experimental Economics. Princeton University Press, Princeton, NJ.
Kahneman, D., Tversky, A. (1979). "Prospect theory: an analysis of decision under risk". Econometrica 47 (2), 263–291.
Kahneman, D., Knetsch, J., Thaler, R. (1990). "Experimental tests of the endowment effect and the Coase theorem". Journal of Political Economics 98, 1325–1348.
Kamin, K.A., Rachlinski, J. (1995). "Ex post $\neq$ ex ante: determining liability in hindsight". Journal of Law and Human Behavior 19, 89–104.
Korobkin, R. (1998). "The status quo bias and contract default rules". Cornell Law Review 83, 608–687.
Koszegi, B., Rabin, M. (2006). "A theory of reference-dependent preferences". Quarterly Journal of Economics 121, 1133–1165.
Kuran, T., Sunstein, C. (1999). "Availability cascades and risk regulation". Stanford Law Review 51, 683–768.
LaBine, S.J., LaBine, G. (1996). "Determinations of negligence and the hindsight bias". Journal of Law and Human Behavior 20, 501–516.
Laibson, D. (1997). "Golden eggs and hyperbolic discounting". Quarterly Journal of Economics 62, 443–478.
Landeo, C.M., Nikitin, M. (2005). "Split-award tort reform, firm's level of care and litigation outcomes". University of Alberta, Department of Economics, Working Paper.
Ledyard, J.O. (1995). "Public goods. A survey of experimental research". In: Kagel, J.H., Roth, A.E. (Eds.), Handbook of Experimental Economics. Princeton University Press, Princeton, NJ, pp. 111–194.
List, J. (2003). "Does market experience eliminate market anomalies?" Quarterly Journal of Economics 118 (1), 41–71.
Loewenstein, G. (1999). "Experimental economics from the vantage point of behavioral economics". Economic Journal 109, 25–34.
Loewenstein, G., Issacharoff, S. (1994). "Source dependence in the valuation of objects". Journal of Behavioral Decision Making 7, 157–168.
Lucking-Reiley, D. (1999). "Using field experiments to test equivalence between auction formats: magic on the Internet". American Economic Review 89 (5), 1063–1080D.
McAdams, R., Nadler, J. (2005). "Testing the focal point theory of legal compliance: expressive influence in an experimental Hawk/Dove game". Journal of Empirical Legal Studies 2 (1), 87–123.
McAdams, R., Rasmusen, E. (2007). "Norms in law and economics". In: Polinsky, A.M., Shavell, S. (Eds.), Handbook of Law and Economics. Elsevier, Amsterdam. Chapter 20.
McCaffrey, E.J., Baron, J. (2005). "The political psychology of redistribution". UCLA Law Review 52, 1745–1792.
McKelvey, R.D., Page, T. (2000). "An experimental study of the effect of private information in the Coase theorem". Experimental Economics 3, 187–213.
McKelvey, R.D., Palfrey, T.R. (1998). "Quantal response equilibria for extensive form games". Experimental Economics 1, 9–41.
Milgrom, P. (2004). Putting Auction Theory to Work. Cambridge Press, Cambridge.
Mitchell, L.E. (1999). "Understanding norms". University of Toronto Law Journal 49, 177–245.

Myerson, R., Satterthwaite, M. (1983). "Efficient mechanisms for bilateral trading". Journal of Economic Theory 29, 265–281.

Plott, C.R. (1987). "Dimensions of parallelism: some policy applications of experimental methods". In: Roth, A.E. (Ed.), Laboratory Experimentation in Economics: Six Points of View. Cambridge University Press, Cambridge, pp. 193–219.

Plott, C.R., Zeiler, K. (2005). "The willingness-to-pay/willingness-to-accept gap, the endowment effect, subject misconceptions and experimental procedures for eliciting valuations". American Economic Review 95, 530–545.

Posner, R.A. (1998). "Rational choice, behavioral economics, and the law". Stanford Law Review 50, 1551–1576.

Priest, G., Klein, B. (1984). "The selection of disputes for litigation". Journal of Legal Studies 13, 1–55.

Rachlinski, J.J., Jourden, F. (1998). "Remedies and the psychology of ownership". Vanderbilt University Law Review 51, 1541–1582.

Radner, R., Schotter, A. (1989). "The sealed-bid mechanism: an experimental study". Journal of Economic Theory 48, 179–220.

Reinganum, J., Wilde, L. (1986). "Settlement, litigation, and the allocation of litigation costs". RAND Journal of Economics 17, 557–566.

Robbennolt, J.K., Studebaker, C.A. (1999). "Anchoring in the courtroom: the effects of caps on punitive damages". Journal of Law and Human Behavior 23, 353–373.

Roth, A.E. (2002). "The economist as engineer: game theory, experimentation, and computation as tools for design economics". Econometrica 70 (4), 1341–1378.

Sanfey, A.G., Loewenstein, G., McClure, S.M., Cohen, J.D. (2006). "Neuroeconomics: cross-currents in research on decision-makeup". Trends in Cognitive Science 10, 108–116.

Schelling, T.C. (1960). The Strategy of Conflict. Harvard University Press, Cambridge, Mass.

Schwartz, A. (1998). "Incomplete contracts". In: New Palgrave Dictionary of Economics and Law. Palgrave, Macmillan, Hampshire, UK.

Shogren, J.F., Shin, S., Kliebenstein, J., Hayes, D. (1994). "Resolving differences in willingness to pay and willingness to accept". American Economic Review 84 (255), 259–264.

Simon, D., Pham, L., Le, Q., Holyoak, K. (2001). "The emergence of coherence over the course of decision-making". Journal of Experimental Psychology, Learning, Memory and Cognition 27, 1250–1260.

Smith, V.L. (1982). "Microeconomic systems as an experimental science". American Economic Review 72 (5), 923–955.

Spier, K. (1994). "Pretrial bargaining and the design of fee-shifting rules". RAND Journal of Economics 25, 197–214.

Stallard, M.J., Worthington, D.L. (1998). "Reducing the hindsight bias utilizing attorney closing arguments". Journal of Law and Human Behavior 22, 671–681.

Tyran, J.-R., Feld, L.P. (2006). "Achieving compiliance when legal sanctions are non-deferrent". Scandinavian Journal of Economics 108, 135–156.

Van Dijk, E., van Knippenberg, D. (1996). "Buying and selling exchange goods: loss aversion and the endowment effect". Journal of Economic Psychology 17, 517–524.

Viscusi, W.K. (1999). "How do judges think about risk?" American Law and Economics Review 1, 26–62.

Viscusi, W.K. (2001). "Jurors, judges, and the mistreatment of risk by the courts". Journal of Legal Studies 30, 107–142.

von Neumann, J., Morgenstern, O. (1944). Theory of Games and Economic Behavior. Princeton Univ. Press, Princeton, NJ.

Wilson, R. (1985). "Incentive efficiency of double auctions". Econometrica 53 (5), 1101–1116.

Further Reading

Asch, S.E. (1995). "Opinions and social pressure". In: Aronson, E. (Ed.), Readings About the Social Animal. Worth Publisher, San Francisco, p. 13.

Hoffman, E., McCabe, K., Schachat, K., Smith, V. (1994). "Preferences, property rights, and anonymity in bargaining games". Games and Economic Behavior 7, 346–380.

Kahneman, D., Schkade, D., Sunstein, C.R. (1998). "Shared outrage and erratic awards: the psychology of punitie damages". Journal of Risk and Uncertainty 16, 49–86.

Kahneman, D.J., McCaffery, E.J., Spitzer, M.L. (1995). "Framing the jury: cognitive perspectives on pain and suffering damages". Virginia Law Review 81, 1341–1420.

Kelman, M., Fallas, D.E., Folger, H. (1998). "Decomposing hindsight bias". Journal of Risk and Uncertainty 16 (251), 260–267.

Rachlinsky, J. (1998). "A positive theory of judging in hindsight". University of Chicago Law Review 65, 571–625.

Schkade, D., Sunstein, C.R., Kahneman, D. (2000). "Deliberating about dollars: the severity shift". Columbia Law Review 100, 1139–1175.

Sunstein, C.R., Kahneman, D., Schkade, D. (1998). "Assessing punitive damages". Yale Law Journal 107, 2071–2153.

Chapter 22

THE POLITICAL ECONOMY OF LAW

MCNOLLGAST*

Department of Political Science, University of California, San Diego; Department of Economics, Stanford University; and Department of Political Science, Stanford University, and Hoover Institution

Contents

* Professor of Political Science, University of California, San Diego; Professor of Economics, Stanford University; and Senior Fellow, Hoover Institution, and Ward C. Kreps Family Professor of Political Science, Stanford University.

Handbook of Law and Economics, Volume 2
Edited by A. Mitchell Polinsky and Steven Shavell

DOI: 10.1016/S1574-0730(07)02022-1

Abstract

In the 1980s scholars began applying Positive Political Theory (PPT) to study public law. This chapter summarizes that body of research and its relationship to other schools of legal thought. Like Law and Economics, PPT of Law uses sequential game theory to examine how rules and procedures shape policy and evaluates these outcomes from the perspective of economic efficiency. Like the Legal Process School in traditional legal scholarship, PPT of Law focuses on how the structure and process of legislative, bureaucratic and judicial decision-making influences the law and evaluates these procedures using the principle of democratic legitimacy; however, rather than using procedural norms derived from moral and political philosophy to evaluate procedures, PPT of Law conceptualizes the decision-making procedures of government as rationally designed by elected officials to shape the policies arising from decisions by executive agencies, the courts, and future elected officials. After summarizing this theory, the essay turns to applications of this approach in administrative law and statutory interpretation.

Keywords

Positive political theory, governance, rule of law, government institutions, policy-making processes, judicial review

JEL classification: H11, K23, K40

1. Introduction

The political economy of law is a branch of Law and Economics that applies positive political theory (PPT)—optimizing models of individual behavior applied to political decision making—to study the development of law. PPT of Law is primarily a positive theory of rational strategic behavior in the presence of imperfect information that seeks to explain and predict the content of the law. These theoretical predictions are derived from information about the preferences of citizens, elected officials and government civil servants and the design of relevant political institutions, including electoral processes and the legislative, executive and judicial branches of the government. PPT of Law also includes a normative component that evaluates the effects of the structure and processes of governance in terms of economic efficiency, distributive justice and democratic legitimacy. PPT of Law also is relevant to other consequentialist normative theories of law because it provides a positive theory of the link between political institutions and policy outcomes.

This essay summarizes the assumptions, arguments and conclusions of PPT of Law. In legal scholarship, most studies of the law focus on the courts, judges, cases and judicial doctrine. While the judiciary is an important source of law, judicial doctrines and decisions do not constitute all of law. Most law is set forth in legislation, executive decrees and bureaucratic decisions, yet these sources of law have not been as extensively studied as judicial law. As Staudt (2005, p. 2) observes:

> Although scholars have spent much time and energy debating questions such as how the judiciary should interpret statutes, how agencies should enforce statutes, or why, as a normative matter, Congress should write an altogether different statute, few have delved into the complex web of congressional players, rules, and practices that impact the initial decision to adopt the law and the decision to maintain it in the long-term.

The purpose of focusing on legislatures, the chief executive and the bureaucracy is threefold. First, we seek to understand the role and influence of the executive and legislative branches in creating law. Second, we seek to understand the interactions among these branches of government and the courts—how each branch constrains and influences the law-making activity of the others. Third, we seek to demonstrate that law is not primarily the domain of the judiciary. Because the other branches influence judicial decisions, even judge-made law cannot be understood by treating the courts in isolation.

To this end, PPT of Law examines each major political institution that is part of the law-making process. The analysis begins with elections, which induce preferences on elected officials and are the principal means by which citizens influence policy. Next, we examine decision-making by legislatures, the president, and the bureaucracy. We study these institutions separately for two reasons. First, as noted, each is an important source of law. Second, in order to evaluate these institutions as sources of law, we need to understand the extent to which they respond to citizen interests. The legitimacy of

these sources of law depends on the extent to which they are responsive to citizens, as opposed to interest groups or the personal ideology of decision makers.

After reviewing the executive and legislative branches, we turn to the courts. PPT of Law provides insights about how judges make decisions and create judicial doctrine, and hence about the content of law. Of particular interest is how the other branches influence judicial law-making by forcing the courts to act strategically in developing doctrine and deciding cases.

Before discussing each major institution that makes law, we first review the main schools of legal thought, explaining the differences in the structure of their arguments. The positive and normative approaches of PPT are best understood when placed within the broader context of other important approaches to the study of law and policy.

2. Schools of legal thought

Since the earliest days of English legal scholarship (Coke, 1608 and Hobbes 1651, 1971) legal scholars have debated the question: "What is and/or ought to be the law?" During the last century, this debate was expanded to address the more vexing question: "Who has and/or should have the authority to make, interpret, and apply the law?" The schools of legal thought that contend to understand law and to shape its creation and use can be distinguished by how they answer these questions.

At any point in time, a society inherits a mutual understanding of what law is, say L_0. This understanding may be subject to uncertainty, so that each member of society, i, believes that the state of law is really $L_0 + u_i$, where u_i is a random variable. The institutions of society then determine who participates in interpreting (reducing the variance of u_i) and changing (altering the value of L_0) the law. The "what is" question addresses reducing u_i to explicate L_0 more clearly, while the "what ought" question identifies the optimal law, L^*. The "who has authority" question seeks a cause-effect explanation for why the law is L_0, and the "who should have authority" question identifies those who ought to make the law, presumably because they are most likely to move the law from L_0 towards L^*.

Until the last third of the 20th Century, scholars made few attempts to ground the answers to these questions in coherent theories of the behavior of participants in the process of governance, whether voters, elected officials, civil servants or judges. For the most part, answers to these questions were based on either philosophical or religious arguments, or simple observation of who appeared to have the power to make law that had to be obeyed.

The "what is and/or ought to be the law" questions have three contending answers: law as nature, law as process, and law as policy. *Traditionalist* legal thought does not separate "is" from "ought." Traditionalists regard law as exogenous to politics, society and individual mortals. To traditionalists, law emerges from a source outside of human manipulation, such as God's will, nature or an abstract system of moral philosophy. Law is "good" if it is consistent with these external standards, regardless of its policy

implications. Law that is not "good" is not really law in that it need not be obeyed, and in some cases ought to be disobeyed out of duty to "good law."

By contrast, *Realists* see law as constructed and manipulated by humans to serve earthly purposes. Most modern Realists are consequentialists in that they regard law as policy—a statement of the purposes and obligations of government to be evaluated on the basis of its effects. To these Realists, "good law" is law that produces normatively compelling policy outcomes. Economists will recognize Law and Economics as a form of Realism, wherein the normative objective is economic efficiency.

Another branch of Realism, the *Legal Process School*, is Kantian in that it focuses on law as a means to obtain social purposes, without specifying the social goal. The Legal Process School focuses on the procedural architecture that defines the policy-making process. "Good law" is law that satisfies principles of good decision-making processes that are derived from normative democratic theory, such as assuring rights of participation and according respect to all individuals.

"Who makes the law" is a practical question about the distribution of authority in society. To Traditionalists, law is created outside the context of human institutions and decisions, perhaps by a divinity or simply inherited as part of the natural order, like the physical laws of nature. To Realists, people who have political power create the law. Political power is institutionalized by law that sets forth the rules and procedures of the political and legal system. This component of law also is created by those with power, usually to solidify their authority.

In democratic societies, many players have a role in making law as Realists define it. Voters elect legislators, and sometimes executive officials and judges, and in so doing influence the development of law through their choices among candidates for office. Sometimes voters even pass laws themselves (e.g. through initiatives or referenda). Legislators enact statutes. Where one is present, an independent chief executive vetoes legislation and issues decrees or executive orders. Elected legislatures and chief executives delegate law-making authority to bureaucrats, who then issue rules and regulations, decide how to enforce the law, make expenditure decisions, and produce public goods. Finally, the courts interpret law, resolve conflicts within the law, and make new law, typically when established law is vague, incomplete or contradictory. In some societies, the power given to all of these players depends on a form of higher law, or Constitution, that establishes rules and allocates authority for making law, including amending the Constitution.

"Who should make law?" is fundamentally a question about the legitimacy of the law, and therefore the circumstances under which law should be obeyed. This question also has three contending answers: popular sovereignty (supremacy in creating law should be given to citizens or their elected representatives), judicial sovereignty (supremacy in creating law should be given to the judiciary), and expert sovereignty (supremacy in creating law should reside in the hands of technically trained bureaucrats). The first answer views legitimacy as arising from popular consent, and so is related to the liberal theory of justice and normative democratic theory. In essence, popular sovereignty theories evaluate law on the basis of the extent to which it arises from the consent of the gov-

erned. The other two answers view legitimacy as arising from authorities who possess appropriate skills and/or values, such as religious leaders, judges, technicians or royalty, regardless of the popularity of their decisions among citizens. From combinations of these answers emerge eight major schools of legal thought.[1]

2.1. Traditionalists

The oldest school of legal thought is the *Traditionalist* (or *Classical*) *School*, and finds its most complete expression in Anglo-American law in Langdell (1871, 1880). Traditionalism is the pinnacle of formalism, focusing exclusively on the internal structure of law regardless of its consequences. This focus on internal structure implicitly assumes that law is separate from politics and other worldly pursuits.

Following Coke (1608) and Blackstone (1765–1769), Traditionalists argue that law emerges from inherited cultural norms, such as Saxon traditions, God's command, or nature. According to Traditionalists, humans do not make law; however, some humans must interpret the inherited law and decide how it applies to daily life. In this sense, humans make law, and, according to Traditionalists, those who make law should be "oracles" who are trained in appropriate traditions and are independent of outside influences, including those arising from the political process. In some societies law is thought to emanate from deities, and legislators and judges must be selected from or approved by the clergy, as is the case in Islamic Law states such as Iran.

2.2. Realism

Legal Realism is a broad category of schools of legal thought. All positive Realist legal theories regard law as made by humans to serve the objectives of those who make law, and all normative Realist theories evaluate law according to the extent to which it conforms to some version of a liberal theory of justice. But Realist schools differ in their assumptions, logic and conclusions in addressing the core positive and normative questions about the development of the law.

The first Realists, though not known by that name, were from the *Sociological Jurisprudential School* (SJS), represented most clearly by Holmes (1881, 1897), Cardozo (1922), Pound (1931) and H.L.A. Hart (1961). SJS replaced Traditionalists in Anglo-American law. SJS argues that because law has social consequences, it ought to be regarded as an element of, or input to, policy. In this view, law should be evaluated on the basis of whether it improves society according to democratic principles, implying that law should serve the interests of most citizens while respecting individual rights. Although acknowledging the connection of law to the welfare of citizens, SJS, like Traditionalism, relies on philosophical reasoning or observations of cultural norms, not theory or facts about how citizens behave or perceive their interests, to evaluate policies.

Modern heirs of Holmes and Pound go one step farther, treating law as policy itself, i.e., an allocation of resources or a division of winners and losers by use of force

[1] For surveys see Horwitz (1992); Posner (1990); Eskridge and Frickey (1994).

(Llewellyn, 1930, 1931, 1960; see also Landis and Posner, 1975 and Posner, 1990).[2] Modern Realism includes five modern branches of legal scholarship: mainstream Political Science, Public Choice, Legal Process, and two overlapping offshoots of Legal Process, Law and Economics and PPT of Law.

2.2.1. Mainstream Political Science

Mainstream Political Science (MPS) is a type of modern Realism, although political scientists do not always adopt the democratic normative standards of SJS and other Realist schools. That is, mainstream political scientists typically assume that law is policy made by humans according to their values and preferences. MPS is not the same as PPT for two reasons. First, MPS does not use the economic approach of goal-directed rational choice to examine political decisions. Second, MPS has no standards for evaluating policy outcomes other than counting support and opposition or applying moral and political philosophy to a particular policy issue. Thus, MPS measures expressions of preferences through votes and public opinion surveys, and seeks the roots of these expressions by correlating them with socioeconomic measures to ascertain how political values and preferences are created and transmitted.

Work in MPS deals with all the relevant political actors, including casting votes by citizens, enacting statutes (making policy) by legislators, implementing statutes by the executive and the bureaucracy, and deciding cases by judges. Because PPT research on voters, legislators, and the executive branch builds on and has extensive overlap with MPS, we include the latter's contributions in these areas in subsequent sections that focus on PPT of law.

Research on the judiciary in MPS views judicial decisions as expressing the preferences of judges, and seeks to determine the sources of these preferences. One MPS group, the Attitudinalists, searches for judicial preferences in the personal characteristics and values of judges. The pioneering studies by Pritchett (1948), Schubert (1959, 1965), Nagel (1961, 1969) and Spaeth (1963) developed many of the techniques used to study judges' attitudes.[3] Another MPS group regards judicial preferences as derived from the political process in much the same way as politics influences the preferences of elected officials and bureaucrats. Other MPS scholars look for the source of judicial preferences in public opinion (Cook, 1977; Kuklinski and Stanga, 1979; Barnum, 1985; Caldeira, 1987, 1991; Marshall, 1989). Still others look to interest groups (Galanter, 1974; O'Connor, 1980; O'Connor and Epstein, 1983; Epstein, 1985; Sunstein, 1985; Macey, 1986; Caldeira and Wright, 1988; Kobylka, 1991; Epstein and Kobylka, 1992).

[2] Progressives fit within the Realist School, but we will reserve our discussion of their contribution to later in this essay.

[3] More recent works in this paradigm include Tanenhaus et al. (1963), Giles and Walker (1975), Rohde and Spaeth (1976), Baum (1980, 1988), Carp and Rowland (1983), Segal (1984), Carter (1988), Songer and Reid (1989), Perry (1991), Segal and Spaeth (1993), Songer, Segal, and Cameron (1994), Kobylka (1995) and Songer and Lindquist (1996).

The MPS work that is closest to PPT is the self-designated "Neo-Institutionalist" School (see Epstein, Walker, and Dixon, 1989). Following Peltason (1955) and Dahl (1957), these scholars regard court decisions as derived from the individual policy preferences of judges, but these preferences are constrained and directed by the institutional structure of the judiciary and its relation to the rest of the political process.[4] These scholars regard Supreme Court justices as mediating their own policy preferences according to the norms of jurisprudence and democratic legitimacy. They also regard justices as behaving strategically through their interactions with each other and with lower courts. While each judge seeks to achieve the best feasible outcome, the institutions of the court, such as procedures for assigning the task of writing opinions and the shared norm of precedent, affect both their goals and their strategies (Epstein and Knight, 1998).

2.2.2. *Public Choice*

Public Choice is a modern branch of Realism because it also assumes that law is policy that serves the interests of those in power (see Farber and Frickey, 1991; "Symposium on Public Choice and Law," 1988).[5] Public Choice also is another close relation to PPT because of its use of economic analysis to study politics; however, it has some unique elements that causes scholars in both camps to regard themselves as not part of the other. The distinctive features of the Public Choice School are a strong form of the liberal theory of justice that comes very close to Libertarianism (in fact, some leaders of the Public Choice School are Libertarians), an equally strong suspicion of democratic processes for producing policies that respect this theory, and an absence of concern for distributive justice.

Public Choice scholars regard the normative purpose of government as maximizing a combination of freedom and wealth, implying that the role of government is to ensure individual liberty, to protect private property, and to promote economic efficiency. The goal of economic efficiency is defined by the strong Pareto Principle: a hypothetical social state is superior to the status quo and ought to be adopted if its makes some better off while harming no one. Public Choice rejects the weak Pareto Principle, i.e. that a policy is preferred if the winners could compensate the losers and still experience a net gain from the change, on the grounds that it does not respect liberty or property. In Public Choice liberty and property rights always trump distributive justice.

Public Choice theory is highly skeptical about the efficacy of democracy for achieving economic efficiency, enhancing personal liberty and protecting private property.

[4] Examples of New Institutionalist scholarship are Adamany (1973, 1980), Funston (1975), O'Brien (1986), Gates (1987, 1992), Marks (1988), Epstein, Walker, and Dixon (1989), Gely and Spiller (1990, 1992), Rosenberg (1991), Eskridge and Ferejohn (1992a, 1992b), George and Epstein (1992), Schwartz (1992), Spiller (1992), Spiller and Spitzer (1992), Zuk, Gryski, and Barrow (1993), Cameron (1994), Schwartz, Spiller, and Urbiztondo (1994), Epstein and Walker (1995), Epstein and Knight (1996), and Knight and Epstein (1996).

[5] We define the Public Choice School narrowly, as the term is used in economics, rather than broadly, as in much legal scholarship (e.g., Farber and Frickey, 1991) that regards all work applying microeconomic reasoning to study law and politics as Public Choice.

One Public Choice critique of democracy is that decisions are driven by rent-seeking elected officials and by special interest groups who essentially buy policy from politicians (Buchanan, 1968; Buchanan, Tollison, and Tullock, 1980). Public Choice theory regards policy as purchased by the highest bidder, usually by sacrificing efficiency, liberty and property rights, and therefore as lacking a compelling normative defense.

Another Public Choice critique of democracy is that collective choice is a meaningless concept from both a positive and normative perspective. The basis for this critique is one of the cornerstones of PPT, the Condorcet paradox and the Arrow Impossibility Theorem (Arrow, 1951). Condorcet (1785, 1989) was the first to observe that majority-rule voting can lead to intransitive and unstable social decisions, even though each person votes non-strategically according to a stable, transitive preference ordering. Arrow's Impossibility Theorem, which we discuss more fully in the section on elections, states that if rational individuals have different values and objectives, all social decision process are normatively ambiguous (see also Chipman and Moore, 1976) and, without arbitrary rules that restrict the decision-making process, unstable (McKelvey, 1976, 1979; Cohen and Matthews, 1980). Public Choice scholars infer from these theoretical results that all collective decisions reflect either the imposition of agenda control by someone in power or the random result of an inherently chaotic process (Riker, 1982).

Public Choice challenges the normative legitimacy of all forms of law, whether legislative, judicial or administrative (see Farber and Frickey, 1991 and Eskridge, 1994 for reviews). Some Public Choice scholars conclude that the only solution to these problems is to shrink the scope and power of government and to require unanimous consent to adopt coercive law.

2.2.3. The Legal Process School and its cousins

Another branch of modern Realism is the Legal Process School. The origins of this school lie in a dissatisfaction with the form of Realism that was dominant in the 1950s and 1960s. This version of Realism thought of law solely as the expression of power, and had largely abandoned the normative component that was prevalent among Traditionalists and early Realists. To bring a normative grounding back to the law, the founders of the Legal Process School, while agreeing that law is policy, proposed that law acquires legitimacy from the process by which it is made (Bickel, 1962; Fuller, 1964; Hart and Sacks 1958, 1994; and Wechsler, 1959). "Neutral principles" inform the construction of the legal process to ensure that law-making, whether legislative, administrative or judicial, is reasonable and serves the common good. In *The Legal Process*, Hart and Sacks (1994) did not adopt either popular or judicial sovereignty, but rather see law as a holistic institutional system in which courts, legislatures, and administrative agencies interact to make policy. Indeed, the subtitle of Hart and Sacks' famous 1958 manuscript is *An Introduction to Decision-Making by Judicial, Legislative, Executive and Administrative Agencies*. Hart and Sachs argued that if the design of this system follows principles of representativeness and fairness, the process is legitimate and the policies it produces are in the interest of society.

The Legal Process School is closely related to two other schools within Realism: Law and Economics and PPT of Law. Law and Economics does not have an articulated theory of political legitimacy, and so does not take a position on the issue of popular versus judicial sovereignty (see Posner, 1986; Cooter and Ulen, 1988; Polinsky, 1989; Romano, 1993, 1994; Craswell and Schwartz, 1994; Schuck, 1994; and Shavell, 1987). Nevertheless, Law and Economics research, following other Realists, typically assumes that the purpose of law is to promote collective welfare. In Law and Economics, the normative goal is to increase economic efficiency. But unlike Public Choice, Law and Economics generally uses the weak form of the Pareto Principle: policy change is desirable if the winners could fully compensate the losers and still be better off, regardless of whether compensation actually is paid. Thus, Law and Economics scholars are comfortable with policies that improve overall welfare by reducing transactions costs, the dead-weight loss of monopoly, or the incentive to engage in socially undesirable behavior, even if the losers (e.g., a monopoly that is divested or regulated, or a firm that is barred from producing an unsafe or polluting product) are not compensated.

Law and Economics employs positive microeconomic theory, which assumes rational, self-interested behavior, to predict the policy outcomes that will arise from a set of legal rules, and welfare economics to evaluate alternative approaches to the law for solving the same problem. The essential feature of work in Law and Economics, and arguably its most important contribution to legal scholarship, is the application of sequential game theory to explore the consequences of law, using a two-step analysis:

> Stage I: society adopts law to constrain and to direct rational, self-interested behavior.
> Stage II: members of society maximize their selfish interests, given the law that shapes their incentives.

Socially desirable rules parallel the accomplishment of perfectly competitive markets as perceived initially by Adam Smith (1776): channel individual greed so that it leads to maximum social welfare. This dictum is almost identical to Madison's argument in Federalist 10 that in designing government institutions ambition must be made to counteract ambition. Hence, Law and Economics typically analyzes a legal rule (e.g., cost-plus regulation, formulas for compensating breach of contract, tort liability standards) to identify its incentives, to characterize the efficiency of the behavior arising from these incentives, and to propose an alternative that, if not perfectly efficient, at least is better.

PPT of Law is a close relative to Law and Economics. In fact, PPT of Law can be conceptualized as attacking a loose end in Law and Economics: why rational actors who greedily maximize their personal welfare in the second stage of the game altruistically adopt legal rules in the first period that constrain their subsequent behavior in order to maximize social welfare. PPT of Law also extends Law and Economics into new areas by using its method to study a broader array of legal issues, such as administrative procedures, statutory interpretation and judicial doctrine.

Like Law and Economics, PPT of Law employs microeconomic theory to study the development of legal rules and institutions. The underlying assumptions are that political actors, like participants in private markets, are rational and goal-directed, and that government institutions, including the electoral process, shape their incentives.

As with Law and Economics, PPT of Law uses sequential game theory as its core analogy, but in PPT the process has four stages, not two.[6] In the first stage, citizens vote for candidates. In the second stage, elected officials (legislators and, where relevant, independent executives) produce law that empowers bureaucrats. In the third stage, a bureaucratic official makes decisions to elaborate and to enforce the law as authorized by statutes or decrees (e.g., Executive Orders). In the fourth stage judges make decisions on the basis of their own preferences, subject to the constraints and incentives that are established by pre-existing law (judicial precedent, statutes, the Constitution).[7] In each stage, decisions reflect "rational expectations" in that choices are based on expectations of the future behavior of decision-makers in subsequent stages. Because the four-stage game is repeated, in the fourth stage courts make decisions in expectation that all other actors will have a chance to respond to them.

The study of regulation has played a central role in the development of both Law and Economics and PPT, and both schools cite the early works on the economic theory of regulation as part of their canon (e.g., Stigler, 1971; Posner, 1974; for a survey of this work, see Noll, 1989). The economic theory of regulation grew out of a desire to explain a key finding of early Law and Economics research, which is the divergence between normative Law and Economics (welfare maximization in the presence of market failures) and the actual effect of some regulation (cartelization and cross-subsidization). This research first focused on rules issued by the agency, then on legislation and oversight by the agency's principals, the legislators, and, finally, the decisions by the legislature to create the administrative procedures of agencies and the jurisdiction, powers and procedures of the courts as a means of influencing the actual policies that emerge from agencies and courts.

PPT of Law differs from the Legal Process School in two important ways. First, PPT, along with Law and Economics, argues that legal processes are designed to achieve policy objectives, and not as ends in themselves to satisfy neutral principles. PPT and Law and Economics are consequentialist in that they evaluate processes on the basis of their outcomes. Second, PPT extends Law and Economics by providing an alternative answer to the question "Who makes law?" In particular, PPT accords more weight to the role of citizens and elected officials, in other words the processes of democratic policy making, and less weight to the role of bureaucrats and judges, in other words the processes of policy implementation, than does the Legal Process School.

[6] Of course, in some cases, a stage may be missed, such as when a Constitutional challenge is raised against a statute, or when voters create a statute through the initiative.

[7] Each of the four stages is further divisible into a sequence of substages. For example, in a hierarchical judiciary, decisions are made sequentially by courts at each level. This elaboration of PPT of Law is examined in subsequent sections.

PPT argues that the choice of structure and process is directly related to the choice of substantive policy. Choice of legal process—that is, the design of institutions that make and enforce policy—is a substantive political choice that is directly connected to policy objectives and outcomes (c.f. Noll, 1976, 1983; Shepsle and Weingast, 1981; McCubbins, Noll, and Weingast, 1987), not some "neutral" choice that is independent of policy content and based on principles unrelated to policy objectives. According to PPT, elected officials design the structure and process of agency decision-making and judicial review to make bureaucratic and judicial decisions accountable to legislative and executive authority (Wilmerding, 1943; Shapiro, 1964; Fiorina, 1977a, 1977b, 1979; Fiorina and Noll, 1978; Weingast, 1984; McCubbins and Schwartz, 1984; McCubbins, 1985; McCubbins and Page, 1987; Ferejohn, 1987; McCubbins, Noll, and Weingast, 1987, 1989; Moe, 1989; Kiewiet and McCubbins, 1991; Eskridge and Ferejohn, 1992a, 1992b; Ferejohn and Weingast, 1992a, 1992b; Cohen and Spitzer, 1994, 1996; Lupia and McCubbins, 1998). In brief, delegation through administrative processes and judicial enforcement is not an abdication of policy-making authority by elected officials, but is a means for assuring that their policy objectives are carried out.

PPT of Law and the Legal Process School share a procedural norm, democratic legitimacy, which means that policy ought to be responsive to the preferences of citizens. If elected officials influence the decisions of unelected officials (bureaucrats and judges), then law that emanates from stages three and four of the law-making process has indirect democratic legitimacy (McCubbins, Noll, and Weingast, 1987; Kiewiet and McCubbins, 1991; Lupia and McCubbins, 1998). These later-stage decisions have direct legitimacy if first-stage decision makers (voters) influence second-stage decisions by elected officials. If statutes and decrees that are crafted by elected officials have democratic legitimacy, the ability of these officials to control third and fourth stage decisions confers democratic legitimacy on this part of law as well.

2.3. The foundations of PPT of law

PPT of Law draws its methods from positive political theory (c.f. Riker and Ordeshook, 1973), an interdisciplinary field in economics and political science that seeks to model and explain political behavior. All PPT models are based on assumptions about how people respond to complexity and competition for resources (including ideas). PPT models share three foundational assumptions.

1. ***Rationality***—PPT models assume rationality, at least in the weak sense. That is, individual behavior is purposive, and decisions are made to advance these purposes (c.f. Ferejohn and Satz, 1994; Cox, 1999; Lupia, McCubbins, and Popkin, 2000). While many theorists assume the self-interest principle (i.e., individuals selfishly maximize their own welfare), this assumption is not essential to rational actor theory (Noll and Weingast, 1991). All that rational actor theory requires is that individuals can make comparisons between any two alternatives, deem one better, worse or the same as the other, and make decisions based on such preferences that are weakly transitive (A preferred to B and B preferred to C implies C not preferred to A).

2. ***Strategic behavior***—PPT assumes that individuals recognize that the consequences of their actions can depend on and affect the actions of others, and take this dependence into account when making decisions. PPT uses games as an analogy to specify how choices and consequences are jointly determined by multiple actors, and characterizes people's choices as strategies within a game.
3. ***Component analysis***—while an individual simultaneously engages in many different interactions with many different people (i.e., people play many games at once), PPT assumes that studying the actions of individuals one interaction at a time produces useful insights. This assumption parallels analysis in Law and Economics, and economics generally. Moreover, it closely corresponds to the concept of "factoring" in cognitive psychology, which refers to the observation that humans tackle complex problems by segmenting them into a sequence of simpler problems. PPT assumes that real social behavior can be explained and predicted on the basis of studying more simple interactions of individuals in a specific decision-making setting.

Each of these assumptions foments debate, particularly the rationality assumption. Assessing this debate is beyond the scope of this essay (c.f. Green and Shapiro, 1994; *Critical Review*, 1995; Cox, 1999; Lupia, McCubbins, and Popkin, 2000).

3. Elections, representation and democratic legitimacy

The responsiveness of elected officials to the values and preferences of citizens is central to both democratic theory and the theory of law. The question of what sort of democracy we *should* have is informed by positive, analytical answers to the questions about what sort of democracy we *can* and *do* have, and how changes in the details of democratic institutions would influence the law that emanates from it. PPT of Law is about making policy in a democracy, and it inevitably must address whether the policies that emanate from democratic processes are normatively attractive. PPT provides two answers to those who challenge the normative significance of its results, given that individual goals may not deserve respect.

First, if institutions are evaluated on the basis of their policy outcomes, and if people usually do behave rationally as defined here, then PPT is normatively important because it links institutions to consequences, regardless of whether the goals of rational actors are normatively attractive. A necessary preamble to "what ought to be" in institutional design is "what are the consequences" of specific forms of institutions. One need not believe that individual choice is normatively interesting to find useful a good positive theory of how individual behavior is shaped by institutions.

Second, normative democratic theory argues that the best method of governance bases policy decisions on individual expressions of political preferences through voting. The relevance of PPT to normative democratic theory is that it examines the extent to which a specific set of political institutions truly are democratic, i.e. that voting actually affects the content of the law so that one can say that policy has the consent of the governed.

PPT of Law asks whether the policy emanating from the four-stage game has democratic legitimacy, which means that law can be said to reflect the preferences of citizens or, if not, whether it fails to do so only because of other conflicting values, such as protection of liberty and property. PPT provides an understanding of the extent to which the values and preferences of citizen/voters are transmitted to their elected representatives, and whether these preferences are then embodied in the law.

3.1. Elections and democratic legitimacy

PPT of law begins with elections for two reasons. First, PPT of Law addresses the normative issue of democratic legitimacy by examining whether, as a matter of positive theory, electoral institutions enable citizens to influence policy. Second, PPT of Law must encompass elections to the extent that elections shape the preferences and behavior of elected officials, bureaucrats and judges. The development of normative and positive theories of the linkages between elections and law are not the only reasons to study voting behavior, so our review of this research is selective and incomplete.

If the preferences of all citizens influence policy, they do so through elections. Three necessary conditions for elections to influence policy are as follows. First, the electoral process must produce elected officials who broadly represent or respond to the preferences of citizens. Second, the legislative process must yield statutory law that broadly reflects the preferences of legislators that are, in turn, derived from or represent the preferences of voters. Third, the law-making actions of elected officials must be carried out by the players in subsequent stages of the game, bureaucrats and judges. This section addresses the first condition. Sections on the legislature, the chief executive, the bureaucracy and the courts discuss the second and third conditions.

Whether the first necessary condition is satisfied depends on the nature of elections. To ensure that elected officials are responsive to the preferences of all citizens, the power to influence elections and hence candidates must be distributed universally and equally. In addition, elections must be competitive in that entry to run for office must be sufficiently easy that incumbents who pursue unpopular policies will attract opposition candidates who will advocate and implement more popular policies.

PPT of democratic elections argues that the presence of political competition leads to the election of candidates who are broadly responsive to citizen preferences. The simplest and most commonly used positive theory of elections is the one-dimensional spatial model of majority-rule decision-making, sometimes called the Black-Downs model after the pioneering work of Downs (1957) and Black (1958). This theory assumes that candidates and voters are arranged spatially on a one-dimensional continuum and that each voter has "single-peaked" preferences. A single-peaked preference refers to a preference ordering in which each voter has a "most preferred" point on the policy continuum and the desirability of other policies to a voter is inversely proportional to their distances from the voter's most desired policy.

If elections are limited to a single candidate and citizens must either vote for that candidate or not vote at all, elections have no effect on policy because the sole candidate

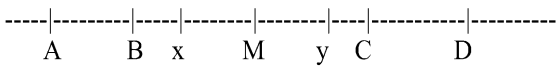


Figure 3.1. Majority-rule equilibrium in one dimensional policy space.

can take any position on the continuum, obtain some votes, and so win the election. But if elections are competitive, the one-dimensional spatial model produces two well-known results: the "median voter theorem" and the "positive responsiveness theorem."

The median voter theorem states that if two candidates run for office, can espouse any position on the continuum, and are motivated solely to win the election, and if citizens are uninformed about the policies that a candidate will adopt, then a candidate who takes the position that is most preferred by the median voter will defeat a candidate who takes any other position, so that the ideal point of the median voter will be the policy that is adopted by the winner. Although the logic of the median voter theorem is widely understood, it is useful to set forth the simple model here because we extend it to illustrate other issues about the politics of public law in other sections. Figure 3.1 depicts a configuration of ideal points, A, B, C, D and M, in one dimension for a polity of five voters. In this example M is the most preferred policy of the median voter because an equal number of voters have ideal points on either side of M. In addition, x and y are hypothetical policy positions of candidates.

If the distance from the ideal point is inversely proportional to the utility of a policy proposal to that voter, then the two voters having ideal points A and B prefer policies to the left of the median voter and would vote for a candidate who proposes x against a candidate who proposes M; however, the other three would vote for M. Likewise, the voters with ideal points C and D would prefer policy to move to the left of M, and would vote for proposal y, but the other three would prefer M. If candidates themselves have no preferences over policies (they simply seek to win the election), each has an optimal strategy to propose M. If only one candidate proposes M, that candidate will obtain three votes regardless of the proposal by the other. If both candidates select M, each has a probability of $1/2$ of being elected, but regardless of which candidate wins, the policy that is implemented will be M. Hence, the equilibrium outcome in majority-rule democracy is the most preferred policy of the median voter.

The positive responsiveness theorem states that a shift in the preference of a voter either will cause the majority-rule equilibrium to shift in the same direction or will have no effect on the equilibrium. Put another way, policy can not shift in the opposite direction of a change in the ideal point of a voter. This theorem arises from performing comparative statics analysis on the median voter equilibrium. If a shift in preferences causes a change in either the median voter's most preferred position or the identity of the median voter, then the winning policy position will move in the same direction as the shift in preferences. For example, in Figure 3.1, suppose the ideal point of the voter who formerly most preferred C to move to the point represented by x. This shift makes this person the new median voter and x the new median voter equilibrium. Likewise, if the preferred outcome for M switched to proposal x, then the identity of the median

voter would not change but x would be the new equilibrium. But if the position of the person who formerly most preferred C switched only to proposal y, the majority-rule equilibrium still would be M.

These two theorems provide a positive theoretical basis for democratic legitimacy. Together they imply that the preferences of citizens affect the policy preferences of elected officials if all citizens have equal opportunity to vote, if voters are informed, and if candidates adopt the policies that they espouse in a campaign. The theoretical basis for believing that candidates will do more or less what they say they will is that elections are repeated, so that if voters believe that candidates have not lived up to their promises, they will vote against incumbents who seek re-election. Thus, the key issues in whether the outcome of democratic elections confers legitimacy on the policies that are adopted by elected officials is whether all citizens have equal access to the polls and all voters are sufficiently informed to evaluate candidates reasonably accurately. If voter participation is biased—say, if citizens whose most preferred policies are A and B above are unable to vote—then the median voter will not represent the median of citizen preferences. If voter evaluation errors are small, policy outcomes will be responsive to voter preferences, but if these errors are large, policy outcomes will have no coherent relationship to the underlying preferences (utility) of citizens.

A more complex spatial theory, beginning with Davis, Hinich, and Ordeshook (1970), relaxes the assumption that policy is a choice in a single dimension. This theory begins with two facts: governments adopt many policies (the textbook example is the choice between guns and butter), and citizens differ in their policy preferences, including the importance they assign to more policy output compared to more private consumption. To capture these facts in a model, the theory represents elections as a choice in a multi-dimensional policy space, which presents the danger of instability and unpredictability due to the Condorcet paradox (majority-rule cycles).[8]

Although the multidimensional spatial model generally lacks an equilibrium outcome that corresponds to the median voter theorem, it does predict a centralizing tendency of winning policy positions. In the standard multidimensional model, the preferences of each citizen are characterized by an ideal point and a utility function in which utility is inversely proportional to distance from the ideal point. In this model, the Pareto Set is the smallest compact subset of points that contains the most-preferred outcome, or ideal point, of every citizen. The Pareto Set has the property that in an election involving a candidate who takes a position outside the Pareto Set, at least one position in the Pareto Set is unanimously preferred to that position. Majority-rule instability arises because each alternative in the Pareto Set can be defeated by some other alternative, although never unanimously. Thus, in competitive majority-rule elections in which candidates are motivated to win, the winning platform will be in the Pareto Set, but will not be stable over a sequence of elections.[9]

[8] In the one-dimensional spatial model transitive individual preferences lead to transitive social preferences under majority rule if individual preferences are single-peaked.

[9] PPT has sought largely unsuccessfully to identify a smaller set that contains all feasible majority-rule outcomes. See Banks, 1991; Epstein, 1998 and Penn 2006.

Elections in a multidimensional space also obey positive responsiveness, albeit in a weaker form than in the one-dimensional model. Because the composition of the Pareto Set is determined by the collection of the ideal points of citizens, the set of potentially winning platforms changes if citizen preferences change. If a change in preferences causes the Pareto Set to shift so that it no longer contains the status quo, then policy will move into the Pareto Set, thereby tracking the general shift in preferences.

The significance of voting theory is that it establishes a weak form of democratic legitimacy. Elections do respond to shifts in preferences, and the power of citizens whose preferences are most completely satisfied arise not from their identity or position, but from the fact that their preferences are in the middle of the distribution. Nevertheless, the choices arising from majority rule clearly can not lay claim to social optimality, as many critics of democracy have shown. The next section addresses these critiques and assesses the extent to which they undermine the democratic legitimacy of elections.

3.2. Critiques of democratic elections

Realists in economics, law and political science have developed a long litany of criticisms of democracy as an effective method of making decisions. These criticisms all are related to the same fundamental theoretical result: the outcome under majority-rule democracy either is unstable (no equilibrium exists), or, even if the outcome is an equilibrium, it does not necessarily (or even probably) maximize social welfare. The following discussion pinpoints the causes of these problems, and how their significance depends on the design of political institutions.

3.2.1. Tyranny of the majority

Perhaps the best-known critique of majority-rule is the possibility of a "tyranny of the majority," which refers to a circumstance in which a majority extracts a small gain but in so doing imposes an enormous cost on a minority. This problem arises primarily because voting transmits little information about the intensity of preferences. If citizens vote for one alternative over another, all that one can infer is that the intensities of their preferences are sufficient to offset the cost of voting. Thus, a majority with moderately intense preferences can impose its will on a minority with very intense preferences. In the absence of side payments (one side purchases the votes of its opponents), majority-rule is unlikely to pick policies that maximize social welfare because of the inherent difficulty in weighing the gains of the victors against the losses to the vanquished.

Despite this problem, democratic theory, in requiring a test of the consent of the governed to legitimate a policy, is not without a normative defense because of the absence of compelling alternatives. If the preference intensities of every individual are measurable and are described by convex functions over all feasible bundles of private goods and public policies, one can then identify an optimal social state. Unfortunately, the lesson of the Arrow Impossibility Theorem is that the only decision-making process that would select that state is a dictatorship run by a perfectly informed altruist. As long as

those who make policy choices are not perfect altruists, no mechanism exists for selecting policies that achieve the social optimum—not the market mechanism because of its indifference to whether the initial allocation of endowments corresponds to differences among citizens in their abilities to derive welfare from income, and not the surrogate for a market mechanism in the public sector, benefit-cost analysis. Thus, the Arrow Impossibility Theorem is a counsel of despair for creating institutions that are capable of picking optimal policy.

Nevertheless, the failure of an institutional system to attain the optimal policy is not fatal to all normative inquiry about the performance of alternative decision-making processes if two conditions hold. First, normative analysis must be able to rule out some outcomes as worse than others, even though it can not produce a complete preference ordering over all possible outcomes. Second, positive analysis must be able to compare alternative decision-making procedures in terms of their abilities to avoid bad outcomes and select acceptably good ones. The Arrow Impossibility Theorem does not say that one can never determine whether one social state is better than another, or that all institutions are equally inept at avoiding bad outcomes. Instead, it says that as a general proposition one can not always determine which of two social states is socially more desirable, and that no decision-making institution always implements the optimal social state. For example, the compensation test (the weak Pareto Principle) can be conclusive, but in some circumstances it is not. As shown by Besley and Coate (1997, 1998), the multidimensional spatial model does produce policy outcomes that are not strictly Pareto dominated by other alternatives with respect to their effects in the current election cycle, although they may not be efficient when one takes into account their effects across multiple election cycles.

The exploration of the meaning of Arrow Impossibility Theorem adds context to both PPT and normative democratic theory. The consent of the governed as a criterion for the legitimacy of policy links to the spatial model in that it confers normative approval on a process in which the set of outcomes predicted by the theory excludes those that are unanimously regarded as inferior to others. Moreover, constitutional democracy, with its guarantees of certain individual rights combined with democratic decision making, can be interpreted as a system in which actions to provide valuable public goods are feasible, but are unlikely to impose enormous harm on anyone unless their preferences are widely at variance with the rest of society.

3.2.2. Imperfect information

A potential problem with democratic decision making arises from the unreality of the assumptions that voters know the positions of candidates, candidates know the preferences of voters, and all voters participate equally and independently in the election. The transmission of citizen preferences to the preferences of elected officials is subject to distortions if these assumptions are relaxed. This section examines the distortions can arise from imperfect information, as examined initially by Downs (1957).

PPT provides a rich interpretation of the information problem in democratic elections as well as an understanding of how citizens deal with this problem. One important insight from Downs (1957) is that uninformed citizens (because a single vote is unlikely to be decisive in an election in which more than a few voters participate) are "rationally ignorant"—that is, they have no instrumental incentive to become informed, or even to vote if doing so is costly. Nevertheless, candidates and their intense supporters have an incentive to reduce the participation costs of voters who are likely to favor them, such as by supplying free information and providing transportation to the polls. Other inexpensive signals are available to voters, such as the party of the candidate, the candidate's career record in and out of public office, and, for incumbents, the general state of the nation.

Fiorina (1981a, 1981b) explains that, in the absence of information about the likely policy preferences of candidates, the optimal voting strategy for a rational voter is "retrospective voting:" to keep a tally of positive and negative evaluations and, when an election occurs, to vote for the incumbent if the running score is positive. If citizens use a high discount rate, this strategy simplifies to observing the state of the nation at the time of the election and voting for incumbents if the voter is better off now than at the time of the last election but against them otherwise. Retrospective voting emphasizes the importance of repeated elections in forcing candidates to be responsive to citizens.

Political parties play an especially important role in overcoming information problems. Parties focus on increasing their overall power in the government, not on winning a particular seat, and as a result have an incentive to nationalize elections by appealing to a broad range of citizens. Parties perform this role by taking actions that connect imperfectly informed citizens to politicians, such as by developing a collective brand name, raising money collaboratively, and arranging for cooperation among members on policy goals (Petrocik, 1981; Cox, 1987).

If citizens rely on interested parties to provide information, one danger is that these groups will provide false or misleading information that will cause citizens to vote against their actual preferences. Cue theory analyzes how voters effectively can use the information that they acquire from easily accessible signals, such as parties, interest groups and other citizens, to inform their decisions while minimizing the danger of manipulation (Berelson, Lazarsfeld, and McPhee 1954; Downs, 1957; Schelling, 1960; Popkin, 1991; Lupia and McCubbins, 1998). In cue theory political parties play an especially prominent role because parties are easily identified and have a strong incentive to secure their reputations among voters (c.f., Cox and McCubbins, 1993, 2005).

3.2.3. Mobilization bias

Mobilization bias refers to systematic over-representation of some preferences relative to others in political decision-making, which includes voting, lobbying, litigating and participating in administrative processes. Mobilization bias arises because some preferences are more easily aggregated and represented by organizations that seek to influence policy through political participation. Mobilization bias is closely connected to the con-

cept of "salience" in MPS, whereby citizens are said to consider only a few issues in a campaign that, at the time, are most important (salient) to them. To focus attention on a few issues is one response to the problem of incomplete information and rational ignorance, as analyzed in Downs (1957). If an issue is salient only to a small minority, a candidate can gain votes among them without sacrificing votes among the majority by advocating policies of importance to them. From the perspective of aggregate social welfare intense per capita preferences among a small group do not necessarily offset less intense per capita preferences among a large majority, so that a candidate's optimal strategy does not necessarily lead to policies that do more good than harm.

Olson (1965) takes this argument further to identify the types of policy preferences that are more likely to be effectively represented. Holding the aggregate intensity of preferences for alternative policies constant across groups, a group's preferences are more likely to be represented if the group is smaller (hence that group has a higher per capita stake), the group is already organized for another purpose (e.g., a firm, a trade association, a union, a church, or an outdoor club as in the case of the Sierra Club), and the preferences among group members are more homogeneous.

Mobilization bias does not necessarily distort policy. For example, Pluralists (c.f. Dahl, 1967) observe that mobilization bias has the advantage of causing advocates of policies to generate information to inform both voters and decision makers. As long as groups representing a variety of policy positions are organized, then, in Madison's terminology, "ambition will counter ambition," leading to a negotiated policy decision that does not fully satisfy any of the organized groups.

In some obvious cases, conflicting preferences are not equally mobilized, in which case the preferences transmitted to candidates for election are distorted. For example, the preference of voters for federal construction projects in their home district, so-called pork barrel expenditures, may only reflect the salience of the large local expenditure for their particular project and the lack of salience of the low individual tax price for projects in other communities. Hence, this form of mobilization bias can cause voters to respond positively to programs that do most of them more harm than good (c.f. Weingast, 1979; Weingast, Shepsle, and Johnson, 1981).

The likely importance of mobilization bias depends on how the electoral system is designed. Democracies exhibit a variety of methods for dividing citizens into constituencies for choosing legislatures. Because representation systems aggregate citizen preferences in different ways, the preferences among legislators that are induced by citizen preferences through elections also differ according to the design of the system of representation. Consequently, the nature and extent of pathologies arising from mobilization bias differs according to how citizens are organized into constituencies for electing representatives.

As a general proposition, smaller constituencies (implying a larger number of elected representatives) are likely to be more homogeneous with respect to their economic interests and their non-economic values, and therefore more likely to produce elected representatives who differ from each other more widely in the policy preferences that they will bring to government policy-making. For more universal policies that are salient

to many voters, narrow constituencies are not likely to bias the pattern of representation in the government; however, narrow constituencies also make it easier for a group with less access to information and lower turnout, or a group with atypically intense preferences in a particular policy domain, to influence an election. Hence, a larger number of smaller constituencies can increase the extent to which the induced preferences of legislatures emphasize narrow policies to benefit a small fraction of the population. The legislators then have a further incentive to form a coalition that delivers targeted benefits to each of their small constituencies.

The pathology arising from this system of representation is the tendency to focus on policies such as pork barrel, where the benefits but not the costs of projects are salient in a majority of districts. But small districts have other consequences that may be more important. Small districts can give representation to small groups with atypical preferences that otherwise would not be represented in the legislature and so would stand no chance of being part of a controlling coalition. Moreover, in small districts citizens are more likely to be familiar with candidates, so that votes are more informed. Thus, the tendency to provide pork barrel projects is properly viewed as the price associated with having a legislature that is a more representative cross-section of the entire population.

In the U.S. House of Representatives and the dominant legislative branches in Canada, the United Kingdom and France, the nation is divided into a large number of distinct geographic districts, each of which is represented by a single legislator. But other nations use different methods of converting votes into legislators. Italy has geographic districts, but each elects several legislators, as did Japan before the reforms of the mid-1990s. In this system citizens cast a single vote, so that each elected legislator has a distinct, non-overlapping constituency in the same geographic area. Typically the most popular candidates receive a large number of "wasted votes" (votes in excess of the number needed to elect them), which enables other candidates to be elected with relatively few votes. For example, compare a district of 2K voters electing two representatives with two separate districts, each containing K voters. In the latter case, candidates need to receive roughly K/2 votes to win a two-candidate race in each district. But in the former case, if one candidate receives K votes in a four-candidate race, the second winning candidate may receive as few as roughly K/3 votes.

Holding constant the number of legislative seats, the main difference between single-member and multi-member districts is that the latter eliminate the necessity for a small group with intense preferences to be geographically concentrated in order to be decisive in an election. As a result, multi-member districts are more likely than single-member districts to enable a group with distinct preferences to achieve representation and to create an induced demand for pork-barrel projects, c.f. McCubbins and Rosenbluth (1995); Cox and Rosenbluth (1996); Cox (1997). This outcome is achieved at the cost of underrepresenting citizens who favor the most popular candidates.

Another method for assuring representation of small groups is to institutionalize their representation [Lijphart (1977, 1996, 1999)]. Minority representation can be assured by reserving seats for them. In India, some legislative seats are reserved for women and for members of "scheduled castes and tribes." This representation requirement has changed

the bundle of policies that are adopted by local governments in favor of health and education (Besley and Burgess, 2002; Besley et al. 2004).

Many Western Europe nations and Japan since reform use a proportional representation system, whereby citizens vote for parties, seats are allocated among parties in proportion to their votes, and parties decide who will fill the seats that they win. Proportional representation from large constituencies reduces the electoral payoff from policies that cater to narrow constituencies, and so reduces the influence of mobilization bias and the political attraction of pork barrel. Proportional representation also substantially increases the role of parties, which orients campaigning more towards national as opposed to local issues.

Whereas the U.S. House is designed to represent relatively small constituencies, the Senate is composed of representatives of states. In small states, Senate constituencies are equal to one or two House districts, so that representation is not likely to differ much between Senators and House members. But most states have several House districts, and a few have twenty or more. In these states, Senators represent a broader and typically more heterogeneous constituency, and hence are less likely to be able to generate majority support by adopting platforms that appeal to the small, mobilized groups that may be decisive in House elections.

Many democracies, including the U.S., elect an independent Chief Executive who also has law-making powers, and some, including many U.S. states, elect several independent executives, each with authority in specific areas of policy. The U.S. President and U.S. governors are elected from constituencies of the whole—all voters within the jurisdiction. As a result, votes for the President and governor, like votes for senators in larger states, are less likely to reflect the narrow, parochial interests of a group than can be influential in some House districts or in Senate elections in small, homogeneous states. These votes are more likely to be determined by issues that are salient to a large number of citizens, and so less likely to cause narrow, parochial interests to influence the policy preferences of an elected executive.

The U.S. national government grants law-making authority to officials that are elected from constituencies that represent different ways to aggregate the preferences of the same citizens. As explained by Madison, this system was deliberately designed to add stability to national policy and to provide checks and balances against the weaknesses and dangers in each form of constituent representation. Specifically, the House is generally more "representative" (in the sense of office holders with heterogeneous preferences) than the Senate and the President, but the latter are generally more oriented towards national rather than local issues.

The effectiveness of this system of checks and balances hinges on how elected officials interact to produce law, for the process of enacting statues determines the extent to which bargains among independently elected officials can be said genuinely to reflect the preferences of their constituents and, therefore, to have the consent of the governed. The next two sections analyze how the three forms of elected law-makers in the U.S. interact to produce law and the extent to which the law that they produce can be said to have democratic legitimacy.

4. The Positive theory of legislative politics

Legislatures are central to democracy because they have the authority to make law. The theory of how democracy works, and therefore the theory of law, revolves around understanding the legislature. If legislatures are corrupt, so too is democracy; if legislatures are representative, then government by the consent of the governed is at least feasible. Thus, an understanding of legislatures drives a theory of law and informs us about how we should interpret statutes, organize government and construct constitutions.

The view of legislatures in most theories of law is not flattering. Nearly all 20th Century jurisprudes agree that legislatures are suspect sources of law. For example, Posner (1990, p. 143) asks and answers the core question as follows:

> More fundamentally, how do we know that legislators really are better policy makers than judges? No doubt they could be—if only they could throw off the yoke of interest group pressures, reform the procedures of the legislature, and extend their own policy horizons beyond the next election. If they cannot do these things, their comparative institutional advantages may be fantasy.

Farber and Frickey (1991, p. 2) add:

> Sometimes the legislature is portrayed as the playground of special interests, sometimes as a passive mirror of self-interested voters, sometimes as a slot machine whose outcomes are entirely unpredictable. These images are hardly calculated to evoke respect for democracy.

Finally, Eskridge and Frickey (1994, pp. cxix–cxx), after reviewing the scholarly literature about the failure of legislative processes, ask:

> [W]hy should judges—or anyone else—defer to the legislature? It is easy to see ... that legal scholarship would start to favor judicial supremacy over legislative supremacy; civil, criminal, and voting rights, administrative, bankruptcy, and antitrust law would become increasingly independent of legislative desires.

This section discusses the contributions of positive political theory to understanding the effect of political institutions and legislative organization on the content of the law. Legislative organization, as we will see, is understood to be analogous to, among other things, town meetings, firms and football teams. Each of these analogies provides insight into how legislatures make decisions and, more importantly, whose interests and welfare they try to serve.

4.1. Understanding legislative politics

PPT seeks to explain whose preferences are reflected in statutes. As such, the theory of the legislature is an essential ingredient to addressing questions such as the legislative intent of a statute and the democratic legitimacy of the policies that it creates. In principle, legislatures are democratic bodies in which members have preferences over

alternative laws (policies), so that a theory of legislatures rests on a conceptualization of how the heterogeneous preferences of a decisive group of legislators (usually a majority) become aggregated into law. To the extent that legislatures really are democratic, the median voter and positive responsiveness theorems ought to apply. And, since the preferences of legislators are induced by elections, then the democratic legitimacy of majority-rule elections confers democratic legitimacy on legislatures. But are legislatures really miniature democracies that represent citizens?

In practice, the view that statutes arise from a democratic interaction among legislators is not the basis of most theories of a legislature. Instead, most theories hold that legislative authority is controlled (some say "seized") by some group of political actors. The theories differ according to who seizes power. Non-partisan theories point to congressional committees, interest groups, and/or the executive branch, while partisan theories point to political parties and their leaders. By contrast, PPT views legislatures as democracies, and their structure and process as selected by majority rule to serve the goals of its members. Thus, what others interpret as seizing power PPT sees as a delegation of limited and reversible authority to serve the majority's common end. The key questions addressed by this debate are whether responsible democratic governance is possible and how legislative outcomes should be interpreted and understood.

A key institutional feature of most legislatures is that the task of crafting legislation typically is delegated to a subset of the members. In most legislatures, and especially legislatures with an independently elected executive, the task of proposing legislation is assigned to committees. Some parliamentary systems do not have committees. In these cases the responsibility for proposing legislation usually is delegated to ministries, which in turn are managed by one or more members of the legislature (a minister and perhaps one or more deputy ministers). Conceptually this system can be viewed as one with very small committees.

Among the powers given to committees is agenda control. One form of agenda control is the *ex ante veto*, or the power to prevent proposals from being considered by the entire legislature (Shepsle, 1979; Shepsle and Weingast, 1987; Weingast and Marshall, 1988). Thus, in both houses of the U.S. Congress, all bills are referred to a committee, and rarely are they considered by the parent body unless the committee formally approves the bill, perhaps after extensive amendment in committee. Another form of agenda control is the *ex post veto*, whereby the committee has the authority to block enactment of the bill as amended and approved by the parent body (Shepsle and Weingast, 1987). For example, the U.S. House and Senate frequently pass different versions of the same bill, and then appoint a joint conference committee to iron out the differences. A legislative body can grant an ex post veto to a committee by allowing the committee to act as its representatives on the conference committee.

Although some models of the legislature are based on the assumption that agenda-setting power is delegated to committees by the chamber as a whole (Weingast and Marshall, 1988; Gilligan and Krehbiel, 1990; Krehbiel, 1991), others argue that in some cases committees have usurped their agenda-setting powers (Mayhew, 1974; Fiorina, 1977a; Smith and Deering, 1984; Shepsle and Weingast, 1987). These two types of

models have profoundly different implications with respect to the democratic legitimacy of legislative outcomes.

4.1.1. Non-partisan theories

Non-partisan theories of legislatures do not necessarily ignore political parties, but instead see them as unimportant manifestations of a more fundamental division of the legislature according to the policy preferences of its members. The fundamental building block of non-partisan theories is that legislators have heterogeneous preferences, which presumably reflects heterogeneity of preferences among citizens. Parties are simply groups of legislators that exhibit much less within-group heterogeneity than the legislature as a whole.

The baseline that we employ for analyzing more complex positive theories of legislatures is the simplest non-partisan theory in PPT, which ignores not only parties but also all other organizational features of a legislature. In this theory, a legislature is a group of equal legislators in which none has greater resources or agenda-setting power than any other. This idealized legislature is organized as if it were a town meeting or a social group, without order or rules, except those that define voting rights. The simplest non-partisan theory is the application of the one-dimensional spatial model of majority-rule decision-making to legislatures, as depicted in Figure 3.1, in which legislative outcomes are the ideal point of the median legislator. This model is the most widely used theory for understanding policy choice in legislatures, and is represented by the work of Riker (1962), Smith (1989) and Krehbiel (1998).

This simple theory has been extended to incorporate and explain the committee organization of legislatures (Krehbiel, 1991). According to this theory, the policy preferences of legislators (and constituents) are uncertain because of imperfect information about the consequences of changes in the law. Committees are a mechanism for legislators to divide labor, develop expertise, and collect relevant information. Each legislator bears the cost of becoming informed on only a relatively small part of policy. With special knowledge may come the ability to mislead less knowledgeable legislators into enacting laws that, with full information, a majority would oppose; however, this adverse consequence of legislative specialization can be overcome if the committee is broadly representative of the membership of the legislature. If the median voter on a committee has approximately the same ideal point as the median voter in the entire legislature, then the legislative outcome of the committee will be the majority-rule equilibrium in the legislature. The fact that committees include members of minority parties is regarded as evidence that committees are selected to be broadly representative in order to protect against strategic information manipulation by a committee.

The significance of the degree to which committees are representative of the legislature is apparent from considering various committee structures in the model depicted in Figure 3.1, and reproduced here is Figure 4.1. The entire legislature consist of five members, whose ideal points are A, B, C, D and M, three of whom form a majority-rule committee to propose legislation to the parent body. The status quo is SQ, and SQ* is

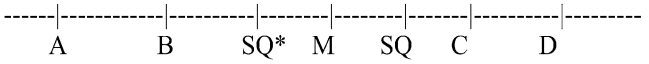


Figure 4.1. Committees as legislative agenda setters.

the policy that the median voter regards as equally attractive as the status quo. If the committee is the median voter plus one member from both the right and the left of M, say the members with ideal points A and C, then the committee will propose M, its majority-rule equilibrium, which also is the majority-rule equilibrium of the entire legislature. But if the committee contains two members from either the right or the left, then their proposed bill will not be M. For example, a committee of the members preferring A, C and D, respectively, will propose its majority-rule equilibrium, C, which could be enacted if the members with ideal points B and M are uninformed and if both of these voters are unconvinced by the protests of A.

Alternatively, if the committee anticipates that the median voter might catch on to the ruse and propose to amend C to M, the committee may exercise its *ex ante* veto and propose nothing. The majority of the committee prefers SQ to M, so that by exercising an *ex ante* veto, the committee achieves a preferred result. Or the committee may try to succeed by proposing C, but then exercise an ex post veto if things turn out badly. If the median voter does not figure out what the committee has done, the committee will obtain C; however, if the median voter proposes M, the committee can prevent a vote on the final bill and preserve SQ. Of course, the committee veto can be overridden; however, doing so is costly, because it eliminates the incentive for the committee to put forth the effort to become informed about this dimension of policy.

The preceding assumes that the committee's proposal can be amended by the whole legislature, which implies an "open rule"—that is, a legislative rule that members are permitted to propose any amendment during the course of floor debate. Most bills in the U.S. Senate are considered according to an open rule. An alternative is a "closed rule," under which either amendments are not permitted or a committee decides in advance which amendments will be considered. In the U.S. House of Representatives, bills usually are considered under a closed rule in which the House Rules Committee decides which amendments will be considered and the sequence of votes. In this case, even if other members are informed, the composition of the committee determines the final outcome. In Figure 4.1, any outcome in the interval [SQ*, SQ] is preferred by a majority to SQ. A committee that includes the members whose ideal points are A and B, plus any other member, can propose a policy slightly to the right of SQ* and receive majority support. A balanced committee, such as one containing the members with ideal points A, M and D, will propose M, which also will pass. Finally, a committee comprised of the members with ideal points at C and D plus any other member will propose nothing because it can not obtain majority support for any bill to the right of SQ. Note that if the median voter on the committee has an ideal point anywhere between SQ* and SQ, the committee will propose that member's ideal point, which will then pass.

The multidimensional spatial model also has been applied to study legislatures. These models view the organizational structure of legislatures—especially committees—as a means to overcome the instability of majority-rule outcomes, creating a "structure-induced equilibrium" where equilibrium would not exist otherwise.

One version of multi-dimensional theory interprets committees as a means of breaking down the dimensions of policy into a series of single dimensions, one for each committee, which then yields a unique median-voter equilibrium in each dimension. The stability of these equilibria are protected by "germaneness" requirements on proposed amendments, which are interpreted as preventing a legislator from creating instability by offering an amendment that introduces a second dimension into a proposed bill.

A related multi-dimensional non-partisan theory argues that the committee system is a means to facilitate vote trades and bargains among legislators (Mayhew, 1974; Fiorina, 1977a; Weingast, 1979; Shepsle and Weingast, 1981; Weingast and Marshall, 1988). Vote-trading is a mechanism for taking into account intensities of preferences. Suppose that a legislature is considering two issues (separate dimensions), each of which has the same preference configuration as shown in Figure 4.1. Suppose that on one issue legislators with ideal points C and D have intensely held preferences while the others do not feel strongly, while on another issue the legislator with ideal point B has intensely held preferences. If the legislature as a whole picks committees, a majority consisting of the legislators with most preferred points B, C and D would assign the members with ideal points M, C and D to the first committee, which would then propose C, and the members with ideal points A, B and M to the second, which would then propose B. The members with ideal points at B, C and D could "trade votes"—the member with ideal B agrees to vote for outcome C from the first committee if the members with ideals C and D agree to vote for outcome B from the second committee. Because each makes a small sacrifice for a large gain, all are better off from trading.

The normative implications of these theories are disputed. On the plus side, a committee system with vote-trading is a means for producing stable outcomes that take into account preference intensities when the alternative could be some combination of chaos and tyranny of the majority. If policy choices are multi-dimensional and, therefore, legislative outcomes are unstable due to preference heterogeneity, some mechanism for achieving stable legislative bargains is necessary for society as a whole to acquire valuable public goods and other desirable policy outcomes. The fact that many bargains may emerge from this process, some of which may be preferred to the actual outcome, is a normative quibble if the actual outcome is substantially better for society than doing nothing or having policy instability. Thus, a committee system that facilitates bargains and enforces vote trades can improve policy outcomes for most or all legislators.

On the negative side, a particular vote-trading agreement is typically one among many that could emerge. All agreements by a majority of legislators that produce policies within the Pareto Set are feasible coalitions, so that the particular agreement that emerges is not obviously superior to others that might have emerged but did not. Moreover, vote-trading coalitions can lead to policy excesses in all dimensions. If outcomes in each policy dimension are driven by legislators with atypically intense preferences,

policy outcomes will not reflect the preferences of the median or pivotal legislator as predicted by the Black-Downs model (this view dates to Wilson, 1885). An example of policy excess that is widely attributed to the committee structure of Congress is "pork barrel" bills, which overspend on public works because spending is distributively attractive to legislators and constituents (Ferejohn, 1974). In practice, the incentive to avoid exclusion from the list of approved projects can and often does cause legislators to agree to a coalition of the whole, in which each legislator who votes for the bill receives a project (Weingast, 1979) even when no project generates more benefits than costs.

Some scholars who emphasize the cost of the committee system see committees as groups of individuals who *seize* legislative power. In this view, the distribution of power in the legislature is as if it were an assortment of monopolies that use their market power to expropriate maximum profits. That is, each committee holds a monopoly over changes in policy on each dimension of the policy space (Shepsle, 1979; Shepsle and Weingast, 1987). Thus, policy is stable, bargains are enforceable and stable (Shepsle and Weingast, 1981; Laver and Shepsle, 1996), but policy outcomes hardly can be said to represent the majority will unless all committees are representative of the distribution of preferences in the legislature.

The process by which legislators build support constituencies among voters has led to the theory that committees are influenced or controlled by interest groups. One such theory is Pluralism (Bentley, 1949; Truman, 1951; Dahl, 1967), which takes a sanguine view of the process because it views policy as the outcome of bargains among many interest groups with conflicting interests. Another is the "political marketplace" in Public Choice, in which committees auction public policy "rents" to the highest-bidding interest groups (Becker, 1983; Buchanan, 1968; Peltzman, 1976; Posner, 1974; Stigler, 1971). Some scholars with the latter view argue that interest groups form "iron triangles" or "unholy trinities" with the committees and executive agencies that control the legislative agenda, or that interest groups "capture" congressional committees and executive agencies (Schelling, 1960; Lowi, 1969).[10]

Whether the sanguine view of the Pluralists or the darker view of Public Choice is more accurate depends in part on the process by which interest groups form and influence legislators. If interest groups from the spectrum of support and opposition to a policy are represented on a committee and participate in crafting its legislation, then committee bargains are likely to embody a balancing of interests, lending support to the view that committee bargains improve welfare. But if interest representation is biased on one side of an issue, the legislative bargain may harm the majority in service of a minority. Of course, even this outcome is normatively ambiguous, for a policy that is intensely desired by a few but mildly opposed by a majority can still increase social welfare.

[10] Nonetheless, as Ramseyer and Rasmussen (1994) observe, bribes in most modern democracies are not prevalent and tend to be small relative to the stakes.

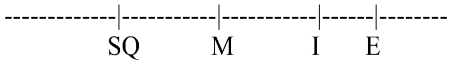


Figure 4.2. Legislative outcomes with executive agenda setters.

The theory of the legislature includes a role for the Chief Executive. Much as when Julius Caesar seized control of the government from the disorderly Senate, modern scholars sometimes see the "imperial" President as seizing control of government from legislatures. Adherents of this view assume that the legislature lacks the ability or the will to set its own agenda, and so cedes that power to the executive (Fiorina, 1974; Edwards, 1980; Sundquist, 1981). For example, some argue that the President or bureaucracy is able to influence legislative outcomes through "overtowering" knowledge (Weber, 1968), control over spending powers (Fisher, 1975) the appropriations process (Niskanen, 1971), or the issuance of executive orders (Moe and Howell, 1999a, 1999b). The executive-as-agenda-setter model is not applied to the U.S. as often as to European, Asian and Latin American legislatures, especially in countries in which the Chief Executive has the power to issue unilateral decrees.

The core analogy in these models is that the President or an agency is able to make take-it-or-leave-it proposals to the legislature (Niskanen, 1971).[11] Examples of legislatures that operate in this fashion are the French and European parliaments. A variant of Figure 4.1 that closely parallels the analysis of the role of committees under a closed rule illustrates the effect on legislative outcomes of granting such authority to the Chief Executive. In Figure 4.2, let SQ be the status quo, M be the most preferred position of the median voter in the legislature, and I be a position that the median voter finds equally attractive as SQ. Notice that the median legislator will prefer any proposal between SQ and I to SQ. If the ideal point of the Chief Executive lies anywhere in this range, the Chief Executive can implement it. If the Executive's most preferred outcome is E, which is to the right of I, the Chief Executive successfully can move policy just short of I. In all cases, the median legislator is not made worse off by the Chief Executive's proposal power, although most of the benefit of policy change is captured by the Chief Executive.

4.1.2. Partisan theories

Another strand of legislative research places parties at the center of analysis, arguing that parties control the legislative agenda. The electoral incentives of party members and the majority's ability to control outcomes lead the majority party to enact generally good public policy that represents the interests of voters. A political party or a coalition of parties that controls a majority of votes can seize the legislative agenda by cartelizing

[11] Romer and Rosenthal (1978) developed the first formal model of an agency agenda setter in analyzing the use of referenda to approve bond measures for school districts.

legislative procedure, keeping measures unfavorable to it off the agenda (Cox and McCubbins, 1993, 2002, 2005) and pushing their platform onto the agenda (Rohde, 1991; Aldrich, 1995; Aldrich and Rohde, 1998, 2001).

A system of single-member legislative districts with plurality voting, such as the United States, tends to have two effective parties (Duverger, 1954; Cox, 1997; Taagepera and Shugart, 1989). In this case, legislative authority is seized by the majority party, and understanding legislative organization and operation involves understanding *party* organization and operation. It is to a discussion of these theories we now turn.

A simple theory of partisan organization conceives of parties as fraternal gatherings of like-minded individuals (Young, 1966; Krehbiel, 2000). For example, some argue that parliamentary coalitions (Laver and Shepsle, 1996), committee decision-making in the German Bundestag (Saalfeld, 1997), and the boardinghouse origins of American political parties (Young, 1966) resemble a structureless social gathering. The implication for legislative organization and policy outcomes is that a majority party or coalition controls the legislature simply because its members are like-minded and can implement their harmonious goals by majority rule.

Another theory of partisan organization emphasizes preference heterogeneity within the party and the role of party leadership. One version of this theory draws an analogy between parties and armies, with generals (party leaders) in charge of the direction, promotion, and placement of the rank-and-file (backbenchers) (Gosnell, 1937; American Political Science Association, 1950; Cohen and Taylor, 2000). In these models, party leaders determine the organization and agenda of the legislature, and their preferences determine policy outcomes, so that party governance is a form of dictatorship. Some European parties and American party machines at the turn of the 20th century resemble the parties-as-armies model (Gosnell, 1937; Cohen and Taylor, 2000).

Recent approaches to understanding party organization see party leaders not as the principals of party members (as in the army model), but as the agents of party members in charge of solving collective-action problems within the party. Analogous to the theory of the firm in industrial organization, this approach argues that party members, recognizing their incentives to "free ride," empower a boss (party leader) to manage and discipline them such that they all can achieve the benefits of cooperation (Cooper and Brady, 1981; Sinclair, 1983; Cox, 1987; Stewart, 1989; Rohde, 1991; Maltzman and Smith, 1994; Binder, 1997). Rohde (1991) and Aldrich and Rohde (2001) argue that the amount of authority that backbenchers delegate to the party leaders waxes and wanes in accordance to the internal homogeneity (like-mindedness) of the party members' preferences and the heterogeneity between the majority party or coalition and the other (minority) parties in the legislature. According to this view, the importance of the party for legislative organization and output is conditional on the amount of authority given to party leaders by party backbenchers; legislative governance is thus "conditional party government."

Building on the "parties as firms" approach, another partisan theory conceptualizes party leadership as a team (Kiewiet and McCubbins, 1991; and Cox and McCubbins, 1993, 1994, 2002, 2005). In this approach, legislative leadership is collegial because it

is distributed among the majority party leadership—the Speaker, the majority leader, and so on—as well as among the chairs of the standing committees, especially control committees such as Rules, Appropriations, Budget, and Ways and Means.

Partisan models of legislatures are analogous to vote-trading models in non-partisan theories. A problem with vote-trading models is that, in a large legislature, coalitions emerge from bargaining among legislators and, as a result, many vote-trading coalitions are feasible. In partisan models, if a single party is in the majority, the membership of the vote-trading coalition is largely determined by party affiliation. And, in legislatures in which no party has a majority, the set of feasible coalitions is vastly reduced to the combinations of a few parties that could form a majority.

4.2. Delegation, monitoring and legislation

One major contribution of PPT is a better understanding of the legislative process. This process has three basic elements. First, because each legislature must allocate scarce plenary time, a substantial fraction of the rules, procedures, and structure of a legislature is devoted to defining how the legislature's agenda will be determined (Oleszek, 2004; Cox and McCubbins, 1993, 2005). Second, the rules must also proscribe what happens if no new law is passed, which scholars call the "reversionary policy." Usually, but not always, the reversionary point is the status quo; however, for bills authorizing expenditures and appropriating funds, the reversionary point normally is zero spending. Third, once plenary time is allocated and the reversionary policy is set, the legislature must have rules and procedures that dictate how a collective decision on policy change will be reached (Oleszek, 2004). These rules and procedures include the duties and powers of committees and the process for assigning members and jurisdiction to them.

Research to explain the structure and process of legislators focuses on two questions. The first is the extent to which legislatures actually delegate power, and the second is the mechanisms for controlling agents once authority has been delegated.

On the first question, Aldrich and Rohde (2001) suggest that the majority party will delegate more as the preferences of its members become more homogeneous. The logic behind this argument is that if all members of the majority party have very similar policy preferences, the policy that any one would adopt is "close enough" to the optimal policy of other party members that party members do not fear any significant cost to delegating authority; however, if party members have widely differing preferences, each member risks losing a great deal if authority is delegated to someone with very different policy objectives.

Laver and Shepsle (1996) examine a condition in which preferences do differ substantially by studying multi-party coalitions in European nations. They conclude that coalitions will determine which policy dimension matters most to each member of the coalition, and then will delegate control over each dimension to a party that values it highly by letting that party appoint the minister. This work highlights an important role of delegation: keeping the majority together, whether it is a single majority party or a

coalition of parties. To keep a majority together requires making certain that each partner has more to gain from remaining in the majority than by defecting to the minority.

Regarding the second question, much of the legislative process involves attempts to mitigate the problems that result from delegation inside the legislature, principally to committees and party leaders (Kiewiet and McCubbins, 1991; Cox and McCubbins, 1993, 2005). The purpose of these mechanisms is to capture the benefits of delegation without giving so much power to agents that agents become dictators. These mechanisms for controlling the behavior of agents have important effects on the flow of legislation.

Research on delegation deals with the ways in which majorities can exercise influence over the discretion of the agents to whom agenda authority is delegated. Controlling the legislative agenda involves creating and delimiting two powers. One power is the authority to put proposed policy changes on the legislative agenda, or *positive agenda power* (c.f. Shepsle and Weingast, 1981, 1987). The other power is the authority to keep proposed policy changes off the legislative agenda, and thereby to protect the status quo—or reversionary policy—from change, or *negative agenda power* (c.f. Cox and McCubbins, 2000, 2002, 2005). Negative agenda power is similar to an *ex ante* veto. Committees with positive agenda power have an *ex ante* veto because they can decide not to let a proposed bill within their jurisdiction reach the floor of the legislature; however, others can be given the power to block proposals independently of the power to make them. For example, the chair of a committee can refuse to allow a committee vote on the final version of a bill, and the Rules Committee can refuse to allocate floor time to a bill that passes out of committee.

There is, of course, an inherent tradeoff between the use of positive and negative agenda power. The more that the majority party distributes veto rights (at the expense of proposal rights), the harder it is to pass legislation. The more it distributes proposal rights (at the expense of veto rights), the greater the risk that some proposals will impose external costs on other members of the majority party—even to the point of adopting proposals that make a majority of the members of the majority party worse off. Thus, the majority party always faces the question of the optimal mix of veto and proposal powers (Cox and McCubbins, 2005; and Aldrich et al., in progress).

The simple model in Figure 4.3 illustrates these issues. Suppose that the ideal point of the median voter of the party is P and for the entire legislature is M, and that, like the Senate, bills are considered under an open rule so that if a bill is proposed, regardless of its initial content, it will be amended to be M and then passed. Note that by definition both P and M are members of the majority party. Assuming that the majority party is democratic, its policy position will be P; however, if it enacts P, it may cause the median voter (and other party members whose ideal points are to the right of M) either to be defeated or to defect. If membership in the majority party is valuable, the median voter in the legislature need not be fully satisfied with the policy outcome for the majority party to retain control. For example, if policy can be as far away from M as M^1 without causing the loss of the legislative median, then status quo policy SQ can be retained; however, if the median voter in the legislature can tolerate no deviation

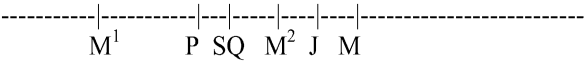


Figure 4.3. Agenda power and party control of legislatures.

from M beyond M^2, then an attempt by the majority party to preserve SQ will cause the party to lose power. The delegation problem for the party is to design a system in which self-interested agents will be willing to propose to amend SQ in the second case (making most party members worse off by moving policy away from the party median) but to retain SQ in the latter case.

The creation of two agents solves this problem. First, proposal power can be given to M, who will make a legislative proposal if SQ deviates from M in any direction. Second, veto power can be given to a member whose ideal point is represented by J, who is indifferent between M and M^2. This member will veto any proposal if the status quo is between M and M^2, but will accept any proposal for any other position of the status quo that must be changed to preserve the majority.

Related to our discussion of agenda control are the many ways that bills can be placed on the agenda. While the United States Constitution grants the President the right to submit proposals to Congress, Article I, Section 1, of the Constitution states that "*all* legislative powers … shall be vested in a Congress." Thus, only the House of Representatives and Senate possess the power to determine whether proposals are considered in their own chambers.[12] Within the House, committees with a particular jurisdiction and specialized task forces have the power to initiate policy change in their policy area. But simply proposing legislation hardly implies that it will be considered by the full legislative body.

Mirroring the fractionalization of power in the Constitution and the divisions in American politics, something of a dual system of agenda power has developed in the House and Senate, in which the legislature divides power among individual committees and the leaders of the majority parties (on the mirroring principle, see Ferejohn, 1987; McCubbins, Noll, and Weingast, 1987, 1989). With the exception of some "privileged" bills,[13] most scheduling in the House is controlled by the Rules Committee (Lapham, 1954; Jones, 1968; Fox and Clapp, 1970; Oppenheimer, 1977; Dion and Huber, 1996; Sinclair, 2002; Oleszek, 2004; Cox and McCubbins, 2005), which in turn is controlled by the Speaker, who is elected by the majority party. Party leaders also determine the membership of other committees.

[12] In the past, Congress delegated the ability to place items on its agenda to executive branch agencies through the one-house legislative veto, by which decisions by an agency could be overturned by a majority in either the House or Senate. In *Immigration and Nationalization Service v. Chadha*, 462 U.S. 919 (1983), the Supreme Court ruled that the Constitution prohibits Congress from writing laws containing a legislative veto provision on the grounds that Article I, Section 1, requires that all legislation must be written by Congress.

[13] The U.S. House Standing Rules grants the privilege to five committees to have direct access to the floor on select legislation.

Committees act as filters, shaping nearly all proposals in their particular policy jurisdiction. They exercise positive agenda control in the sense that they write the bills that are submitted to the Rules Committee to be placed on the agenda of the legislature. They also have negative agenda (or gatekeeping) power in that they can decide simply not to pass legislative proposals on to the Rules Committee. This power is limited in that the floor can pass a "discharge petition" that forces the committee to report a bill, but such a petition is costly to undertake and so is rarely undertaken.

The delegation of the legislature's agenda-setting authority to party leaders and committees creates the potential for mischief, i.e., agency loss, and is the reason why the discretion of each agent of the majority party is limited. At issue is how members assure that the people to whom the agenda-setting authority has been delegated do not take advantage of this authority to use it for personal gain. Legislatures use both checks and balances to accomplish these tasks. These checks and balances provide others with a veto over the actions of agenda setters.

The agenda power of the majority leadership provides an incentive for the majority party's representatives on a committee to take actions that are responsive to the interests of the whole party (Cox and McCubbins, 1993; Kiewiet and McCubbins, 1991). To the extent that the party exercises control over committee assignments and that some assignments are more valued by members than others, committee members have an incentive to be responsive to the party's collective interests (Cox and McCubbins, 1993) in order to be rewarded with subjectively desirable assignments. The shortage of plenary time on the floor of the legislature creates another incentive for substantive committees to compete against each other, in something of a tournament, where the reward for satisfying the party's interest is time for floor consideration of their bills (c.f. Cox and McCubbins, 1993, 2005).

A similar relationship holds between the party and its leadership with regard to the leadership's scheduling activities. The leadership of the majority party has an incentive to pursue the majority party's collective interests to the extent that the party can discipline its leaders (Cox and McCubbins, 1993, 2005). Party leaders are selected by a majority vote of party members. Moreover, disaffected members of the majority party can vote against the wishes of party leaders on the floor or even defect to a minority party if they are dissatisfied with the leadership's exercise of agenda control. The constraints on party leaders imply that party leaders act on behalf of the collective interest of the party, not just themselves.

An important element of agenda control is veto power. Any person or group with the power to block or significantly to delay policy is referred to as a veto gate or a gatekeeper. Nations differ substantially in the number of veto gates that inhabit the legislative process. The United States, for example, represents the end of the spectrum with a large number of veto gates because it has a bicameral legislature that is decentralized into numerous committees plus a President with veto power. In the House of Representatives alone, the substantive committees and their subcommittees, the Rules Committee, the Speaker, and the Committee of the Whole each constitute veto gates through which legislation normally must pass, and the Senate has even more veto gates

due to their lax restrictions on debate. The United Kingdom occupies the other end of the spectrum with parliamentary government and a relatively weak upper legislative chamber.

Whenever legislators consider a bill, they must consider its effect relative to what would occur if no law were passed. In virtually every legislature the final vote pits the final bill against the reversionary policy. The reversionary policy is not necessarily the extant policy. For example, some laws contain "sunset provisions," which mandate that a program be dissolved or an appropriation be terminated by some specified date. A law being considered for renewal under a sunset provision faces a reversionary policy of no law even though the status quo is that the law is in effect.

To understand law-making, it can be important to know whether the reversionary policy can be manipulated, and if so, who possesses the power to do so (c.f. Romer and Rosenthal, 1978). This requires understanding the relationships between the reversion policy, the proposed policy and the preferences of policy makers. Reversionary policies can be defined formally by the Constitution and/or statutes, or through informal solutions to immediate problems. The U.S. Constitution defines the reversion point for some budgetary items (a zero budget), but statutes typically define the reversionary policy for entitlements, such as Social Security, as adjusted annually to account for inflation.

The effectiveness of agenda control is contingent on the reversionary outcome. Whether those who possess positive agenda control will be able to make "take-it-or-leave-it" offers (also known as ultimatum bargaining) to the legislature depends largely on the attractiveness of the reversionary outcome. Positive agenda control confers much greater power if the reversionary policy is no policy at all, as with budgets and sunset provisions, than if the reversionary policy is a continuation of the status quo, as with entitlements and laws without sunset provisions.

Most legislatures possess rules that structure the handling of proposed legislation. Rules define voting procedures, what amendments (if any) that will be considered, the procedures under which amendments will be considered, provisions for debate, the public's access to the proceedings, and so forth. Because of the instability of majority rule voting as exemplified by the Condorcet paradox, the sequence in which amendments are considered determines the composition of the final bill.

As a proposal approaches the floor, the party's influence grows. The majority party's members delegate to their leadership the authority to represent their interests on a broad variety of matters. In the U.S. House of Representatives, the Rules Committee, the Speaker and (if expenditure of funds are required) the appropriations and budget committees all hold power that checks the ability of substantive committees to exploit their agenda control. If a committee's proposal conflicts with the party's collective interest and if the issue is important to the party, either the Speaker or the Rules Committee can kill or amend the proposal, or the budget committees can refuse to supply the necessary funds to implement it. This system of multiple veto points, each controlled by a partially non-overlapping subset of the members of the majority party, constitutes a system of checks and balances to constrain the ability of a substantive committee or the party leaders to pursue policies that are not in the interests of other members of the majority

party. Legal scholars have long recognized that the legislative process has implications for policy (c.f. Farber and Frickey, 1991) and for statutory interpretation (c.f. Eskridge, 1994). PPT gives a new understanding of how the elements of these key processes fit together.

4.3. Policy consequences of legislative structure

The American system of government is defined by deliberate separation of powers, which creates an institutional structure rife with veto players. As the number of effective veto players increases, the government's ability to be resolute (to commit to policy) increases while its ability to be responsive (to change policy) decreases (Cox and McCubbins, 2001). While numerous veto points reduce policy instability, the cost is that government action tends to be more responsive to particularistic interests rather than to broad policy goals than would be the case if the Constitution made a different tradeoff between resoluteness and responsiveness. This Constitutional structure does not imply an absence of collective goods or public-regarding legislation. Rather, the tradeoff created by the Constitution shapes the terrain of policy tendencies that pervade law-making.

Both political parties in the U.S. have created relatively stable reputations for the type of policies they support. Since the Civil War both parties have shown consistent differences on tax and monetary policy (Studenski and Krooss, 1963; Berglof and Rosenthal, 2004). The parties also express consistent differences over agricultural policy, domestic and foreign spending, energy policy, education and health policy (Sullivan and O'Connor, 1972; Bresnick, 1979; CQ Farm Policy, 1984; Browning, 1986; Kiewiet and McCubbins, 1991; Peterson, 1990; Den Hartog and Monroe, 2004; Monroe, 2004). In sum, the obstacles to policy-making in the U.S. legislature have not prevented political parties from presenting differentiated but consistent visions of the role of government.

The Constitutional separation of powers in the U.S. encourages some forms of privatization of public policy, although less than would arise if the only policy-making entity was the House, with its fragmented constituencies. Because the President and to a lesser extent Senators from large states are held accountable for the broad performance of government, while House members and other Senators primarily are held responsible for the effect of policy on relatively small constituencies, policy outcomes represent a compromise of the preferences that representatives derive from different systems of representation. In order to overcome indecisiveness and to forge coalitions among legislators with heterogeneous preferences, private goods are sometimes used as the basis for legislative bargaining, with the consequence that broad public policy goals are packaged with distributive politics that are dominated by special interests, characterized by fiscal pork and rent-seeking, and morselized—all of which contribute to inefficiency.

Perhaps the most widely discussed form of policy inefficiency is *fiscal pork*, which refers to geographically targeted public expenditures for which the incidence and location of projects follow a political rather than an economic logic (Ferejohn, 1974; Weingast, Shepsle, and Johnson, 1981; Cox and McCubbins, 2001). This form of pol-

icy includes classic pork-barrel projects such as dams and levies as well as projects involving water quality (Pisani, 2002), transportation (Baron, 1990; Hamman, 1993), technology (Cohen and Noll, 1991), energy (Stewart, 1975; Davis, 1982; Vietor, 1984; Arnold, 1990, Chapter 9), and defense (Fox, 1971; Kanter, 1972).

The institutional features of the Senate exacerbate the problems associated with fiscal pork. Because senators' districts are geographically defined (as opposed to the population-based boundaries of members of the House), Senate policy-making tends to favor rural interests (Lee, 1998), especially agriculture (McConnell, 1966; Congressional Quarterly, 1984) and other resource-based industries (e.g., coal—Ackerman and Hassler, 1981).[14] Furthermore, because the Senate is less majoritarian than the House (due to the filibuster and the need to rely on unanimous consent agreements—see Krehbiel, 1986 and Binder and Smith, 1997), the distribution of pork by the Senate tends towards universalism (Weingast, 1979; Bickers and Stein, 1994a, 1994b, 1996; Weingast, 1994), whereas the House is more partisan (Cox and McCubbins, 2005).

Another source of inefficiency is *rent-seeking*—a term that refers to an array of subsidies, tax provisions, and regulatory exceptions that special interests extract from government (c.f. Tullock, 1965; Krueger, 1974; Buchanan, Tollison, and Tullock, 1980). Many scholars lament the ways in which rent-seeking perverts democratic accountability and distort economic incentives (Buchanan and Tullock, 1962; Stigler, 1971), while others focus on the related, yet distinct, problem of how the representation of special interests within the legislature causes fragmented, incoherent policy (Shepsle and Weingast, 1987; Weingast and Marshall, 1988). In the extreme, this "balkanization" of politics can lead to the dominance of sub-governments (such as subcommittees) that agree to let each control a particular area of policy for their private benefit (Dahl, 1956; Schelling, 1960; McConnell, 1966; Lowi, 1969; Shepsle and Weingast, 1987).

An example of the ways in which special interests cause inefficient policy outcomes is provided by Banks' (1991) discussion of the roots of the *Challenger* disaster. As Banks documents, once the research and development for the shuttle was underway, the program "picked up political steam" as core political constituencies—shuttle contractors and manned space-flight advocates in NASA—grew in number as expenditures on the project increased. This example shows how programs create support constituencies as they are implemented (Noll and Owen, 1981). Indeed, political support for the project became powerful enough to overcome growing evidence of the severe economic and technical shortcomings of the project.

The core political constituencies for the shuttle program placed great emphasis on the timing of the first launch, leading Congress to push NASA for a quick launch despite misgivings about the operational readiness of the technology among those responsible for implementing the program. Thus, distributive politics conflicted with and overcame

[14] To a large extent, this was also true of the House prior to redistricting. See McCubbins and Schwartz (1988) for further discussion.

the economically efficient courses of action, which, in this case would have been to extend the R&D period and to emphasize capability and safety over timing. But extending R&D would have entailed delaying the transition to the more expensive—and hence politically more beneficial—operational phase, which was opposed by contractors and advocates of manned space flight. Thus, distributive politics led to declaring operational a vehicle that was regarded as unsafe and economically unsound by the managers of the program.

Distributive politics also causes policy in the U.S. to be *morselized*—that is, divided into subcomponents (morsels). Morselization allows elements of a program to be dispersed among politicians as "goodies" for constituents. For example, the broad policy goal of reducing water pollution is divided into many grants to cities for sewage treatment plants (Arnold, 1979; Weingast, 1994), for which members of Congress then can claim credit. While such morsels still aggregate to a public good, the morselization process is inefficient in that the means of production are politically determined, and so do not constitute the least costly means of providing the public good to society.

The separation of powers makes other branches of government more distant from distributive politics. The sources of resistance to excessive responsiveness to special interests that are favored by the legislature are the President, the civil service bureaucracy and the judiciary. The following sections discuss the role of each in making law and policy, and the extent to which they can constrain the tendency of the legislature to favor inefficient policies.

5. The President

In the U.S. and most European democracies, the legislature is the dominant institution for making law. Nevertheless, despite the unequivocal wording of Article I, Section 1, the U.S. Constitution grants some law-making authority to the President. And, in many democracies throughout the world, the Chief Executive possesses the power to issue decrees that have statutory status (Shugart and Carey, 1992, Shugart and Haggard, 2001). Thus, in the U.S. and many other democracies, the Chief Executive plays a significant role in forging legislative bargains that yield new laws, and so the content of law in part reflects the Chief Executive's policy preferences. As with the legislature, if the President's decisions are corrupt, then so, too, is the law that emanates from the President's participation in the law-making process.

In the analysis of citizen voting we noted that the President and the legislature face distinctly different political incentives in making policy because they are elected from different constituencies. Another important factor influencing presidential behavior is the career time-horizon inherent in the office. Unlike other Constitutional positions in the government, the President is limited to two terms. This provision not only shortens the time horizon of the President, but also attenuates the responsiveness of the President to citizen preferences, especially in the second term. In addition, because of the importance, visibility, historical significance and clear accountability of the office, the

President's personal reputation hinges much more on the broad performance of the government than is the case for legislators. To the extent that the desire for respect and status motivate human behavior, the President's behavior is likely to be influenced more by these concerns than is the behavior of a legislator. All of these factors together imply that the decisions of the President are likely to be more responsive to the overall efficiency and effectiveness of government than are the decisions of the legislature.

Of course, for these differences in incentives and policy orientation to influence the law, the President must have the power to affect the law, which is the issue to which we now turn.

5.1. Presidential law-making powers

The Constitution grants the President two types of powers: 1) legislative as defined in Article I, Section 7 (veto power), Article II, Section 2 (treaties), and Article II, Section 3 (statutory proposals and the power to convene special sessions of Congress); and 2) executive, as described elsewhere in Article II. In addition, nothing in the Constitution prevents the President from using the visibility of the office and the information that the executive branch collects to organize public support for policies.

5.1.1. Veto power

The ability to veto legislation is the most powerful presidential legislative tool, especially when the President faces a Congress that is controlled by the opposition. The veto power confers more wide-reading influence than simply the authority to prevent the enactment of legislation that is not overwhelmingly popular in both branches of the legislature. Both actual vetoes (Cameron, 2000) and veto threats, either implicit (see Matthews, 1989; Cameron and Elmes, 1995; McCarty, 1997) or explicit (Ingberman and Yao, 1991), provide the President with significant leverage in shaping the final contours of legislation. This leverage is particularly useful in constraining and influencing congressional policy initiatives during periods when the President is not a member of the party that controls Congress (Kiewiet and McCubbins, 1988; Kernell, 1991). Even during periods of unified partisan control, the veto stabilizes policy (Hammond and Miller, 1987; Brady and Volden, 1998; Krehbiel, 1998). If the President prefers policies that are closer to the status quo (or reversion) than Congress, the veto is a very powerful tool.

Nevertheless, the efficacy of the Presidential veto is limited. Although the veto enables the President to limit the departure of new laws from the reversion point, it does not give the President leverage to pull policy further from the reversion point than Congress prefers. Kiewiet and McCubbins (1988) demonstrate the limits of this "asymmetric" veto power on appropriations decisions. They show that the President, while able to use the veto to limit congressional appropriations, cannot use the veto to extract appropriations that exceed the amount preferred by Congress. Furthermore, the President's ability to use the veto successfully is tied to the President's "resources" (presidential popularity and the seat share of the President's party in Congress) and the

"political environment" (when a bill is enacted in relation to the election cycle) (Rohde and Simon, 1985; Wooley, 1991).

The President's veto power is also limited because the President may face electoral punishment for vetoing legislation. Groseclose and McCarty (2001) show that, on average, presidential approval drops significantly following vetoes, particularly during periods of divided government. This argument implies that Congress may be able to use the veto power against the President by passing legislation that harms a key constituency of the President's party, thereby forcing the President to lose support regardless of whether the bill is signed or vetoed (Gilmour, 1995, Ch. 4, 1999).

5.1.2. Treaty power

Article II, Section 2, of the Constitution grants the power to negotiate treaties to the President, but it also states that two thirds of the Senate must concur for a treaty to be enforceable. The Senate has, on occasion, rejected treaties that the President negotiated (see Helbich, 1967). In other cases, the Senate has adopted only part of a treaty or ratified only an amended version. Thus, the constitutional requirement of Senate approval limits the President's authority in foreign affairs (Glennon, 1983).

Congress sometimes passes so-called "fast-track" legislation that commits it to vote on the treaty as proposed by the President without amendment or condition, but this legislation, because it requires passage in both the House and Senate, requires that the House as well as the Senate be given the opportunity to ratify the treaty. Thus, the power granted by fast track is, to some degree, offset by making the House a second veto player. Furthermore, some treaties require further legislation and appropriations to be effectively implemented, and Congress can effectively veto a treaty by failing to pass these bills.

Finally, presidents have increasingly used executive agreements with other countries as a means of skirting the treaty process entirely (Moe and Howell, 1999a). We discuss this topic below in Section 5.2.2.

5.1.3. Legislative proposal power

Though not a member of the legislature, the President frequently drafts legislation and proposes it to Congress.[15] In doing so, and most notably in formulating yearly budget proposals, the President can make use of many bureaucratic resources (the OMB, for example; see Heclo 1975, 1977). Yet, this proposal power is weak.

The proposal power of the President is conditional upon congressional consent. Some of the most important proposal powers of the President, such as budget and tax proposals, are requested by statutes that specify the matters to be addressed in the proposals.

[15] Though Presidents cannot formally introduce a bill in Congress, they routinely introduce legislation by way of a member of Congress of the President's party, who is the official sponsor of the legislation.

Moreover, all executive legislative proposals must pass through the standard gauntlet of congressional veto gates, starting with substantive committees. These proposals always receive extensive scrutiny and revision by Congress (Kiewiet and McCubbins, 1991).

Presidential proposal power depends on the partisan composition of the legislature and the presidency. In the post-war era, the raw number of "important" laws does not vary significantly between Democratic and Republican presidential administrations (Mayhew, 1991), yet Democratic Presidents have proposed considerably more legislation than Republican Presidents. Moreover, Presidents of both parties have proposed more legislation under unified government (Browning, 1986, p. 80). The willingness to propose legislation apparently is influenced by its anticipated success, and scholars have long noted that Presidents are much more successful in the legislature if their party controls Congress (Edwards, 1980, 1989; Bond and Fleisher, 1990; Peterson, 1990).[16] Ronald Reagan, for example, faced a Democratic House during his terms as President. While Reagan was successful in proposing increases in defense programs, he failed to reduce spending on domestic social programs. Reagan's differential success in domestic and defense spending contributed to the rapidly increasing budget deficits of the 1980s (McCubbins, 1991).[17]

5.1.4. Coalition building power

Beyond vetoes and treaties, the President's most effective law-making tools are the informal resources that aid him in building coalitions. In the modern age of media, the President's visibility enables the President to pressure members of Congress to support administration proposals by "going public" (Kernell, 1986; Edwards, 1983; Canes-Wrone, 2001). That is, the President can appeal to the public, playing on the electoral concerns of members of Congress, to force legislative action on a bill.

Certainly, there are instances where public appeals are effective, notably during the budget battles of the 1980's and 1990's; however, public appeals also have limits. First, the President's position must enjoy sufficient popular support to cause Congress to be concerned. Second, even with public support, the President can only go public so many times before the public stops paying attention (Popkin, 1991). Finally, the President is not the only player who can go public. The President must also consider the electoral consequences of a dispute with Congress, and if congressional leaders can capture the media's attention, Congress can parry the president's moves by also going public (Groseclose and McCarty, 2001).

The President also can build coalitions by dolling out Presidential patronage, in the form of fundraising assistance and campaign support (Cohen, Krassa, and Hamman, 1991; Davidson and Oleszek, 2000), well-publicized visits to the White House

[16] For rejoinders to Mayhew (1991), which point to the difference in the content of legislation between periods of unified and divided government, see Sundquist (1992), Lohmann and O'Halloran (1994), Epstein and O'Halloran (1996, 1999), Edwards, Barrett, and Peake (1997), Binder (1999), and Cameron (2000).

[17] On this point, see also Cox and McCubbins (1991) on tax policy since the New Deal.

(Neustadt, 1960; Covington, 1987), rides on Air Force One (Walsh, 2003), placement of federal construction projects, and the geographic distribution of other federal programs (Edwards, 1980). Similarly, the President is able to facilitate log rolls across bills, promising not to veto (or to offer support for) one bill for support on another (Cameron, 2000).

The President's coalition building power is limited by its partisan element. That is, much like proposal powers, the President's ability to build successful coalitions depends on which party controls each branch of Congress. Presidential support scores tend to be very strong among members of Congress from the Presidents' party, and very weak among members of the opposite party (Kiewiet and McCubbins, 1991). Further, what is sometimes mistaken for Presidential patronage—such as the change to Rural Free Delivery in the late 19th century—is actually partisan pork distributed by Congress to its members (Kernell and McDonald, 1999).

5.2. *Executive powers*

As chief executive, the President's authority over executive branch agencies confers indirect law-making power. Statutory law requires implementation and enforcement by agencies, which inevitably implies some power to make law.

5.2.1. *Executive orders*

In some cases the President can bypass the coalition building process and make policy directly. In many nations, the chief executive has the constitutional power to issue decrees, which usually cannot directly and permanently override a statute but otherwise have the same legal standing as a statute. For example, the agencies that regulate telecommunications in India and Mexico initially were established by decrees, not statutes. The U.S. Constitution does not grant the President the power to issue decrees, but it does give the President the authority to implement policy and to manage the executive branch. To exercise this power, the President issues Executive Orders. Like decrees, they can not explicitly contradict statutes or create authority where none has been granted by Congress or the Constitution, but otherwise these orders can influence law by setting forth procedures and standards for decision-making by agencies.

In response to the common notion that the President lacks the ability to act unilaterally in making law (e.g., Peterson, 1990), some have argued that the power to issue executive orders confers the ability under some circumstances to end-run a hostile Congress and unilaterally to make policy (Moe and Howell, 1999a, 1999b; Mayer, 1999, 2001; Deering and Maltzman, 1999; Howell, 2003). An impressive list of government actions have occurred through executive order (the Emancipation Proclamation and the creation of several important agencies, including the Environmental Protection Agency, the Food and Drug Administration, and the Office of Management and Budget). Evidence regarding the frequency of and success against court challenges to executive orders indicates that they almost always remain in force.

Nevertheless, Executive Orders as a source of law have important limitations. Most executive orders have little importance, so that their overall success rate is not a particularly revealing statistic. Moreover, the President is constrained by the *Youngstown Steel* decision, which, among other things, states that Presidential actions that directly violate the will of Congress are illegal. Furthermore, in issuing executive orders the President is subject to limitations in dealing with the bureaucracy in the form of legislated administrative structures and procedures, which are designed to protect the influence of Congress over agency decision-making (McCubbins, Noll, and Weingast, 1987, 1989).

Some executive orders arise from statutory authority that has been delegated by Congress. In these cases, executive orders either fulfill a statutory obligation or implement a statutory authority, and so are simply the result of effective policy-making delegations by Congress, rather than the President's means of end-running the opposition.

Once a President issues an executive order, in most cases expenditures are likely to be necessary to carry it out. Congress can undermine the order by simply writing into the relevant appropriations bill that funds cannot be spent for the purpose of carrying out the order. The President can veto the appropriations bill, but the President's veto threat is usually not an effective means for increasing appropriations. Furthermore, a President who uses delegated authority to issue executive orders that a majority of Congress finds repugnant risks being denied such delegated authority in the future. Hence, to maximize influence over many issues, a President will think carefully about departing from the range of acceptable outcomes according to the preferences of congressional majorities.

From the preceding discussion, the value of executive orders as a source of presidential policy control can be summarized. First, executive orders can be an important source of presidential policy authority in areas where Congress itself is unable to act either initially to produce a statute or reactively to prohibit implementation. These cases enable the President to take advantage of a circumstance in which the diversity of preferences in Congress causes policy gridlock. Second, for a variety of reasons Congress may prefer to let the President control the details of policy implementation. In areas where the outcome of policy actions is uncertain, Congress may regard the executive as being more flexible to respond to new information (Bawn, 1995), and if the policy is highly controversial, Congress may use delegation to increase the political accountability of the President (and lessen the accountability of Congress) for the ultimate policy outcome (Fiorina, 1982). In these cases Congress regards delegation of authority to be in its collective interest, and can subsequently use the appropriations process to overturn presidential decisions that are unacceptable to a majority of legislators.

5.2.2. Executive agreements

In order to overcome the constraints that the Senate imposes on treaty ratification, Presidents often opt to negotiate executive agreements instead. Executive agreements allow Presidents to enter into arrangements with other countries without Senate approval, thereby enabling Presidents to sidestep treaty rejections. As Cronin (1980) empha-

sizes in discussing the "imperial presidency," Presidents used executive agreements throughout the 1960's and 1970's to arrange important mutual-aid and military-base agreements with other countries. As O'Halloran (1994) points out, executive agreements often require implementing legislation and so constitute a form of legislative delegation. Accordingly, executive agreements are therefore subject to amendment and authorization by both the House and the Senate. In essence, to obtain an executive agreement, the President trades a 2/3 voting requirement in the Senate for a simple majority in both chambers. As a result, during periods of divided control, Congress places tighter reins on the President's authority to negotiate executive agreements (Lohmann and O'Halloran, 1994).[18]

5.2.3. *Executive oversight*

As the Chief Executive Officer, the President controls hundreds of agencies and seemingly unlimited resources. Among the executive powers granted to the President in Article II of the Constitution are the position of commander in chief of the army and the authority to require written reports from heads of executive agencies. Furthermore, as a practical matter, almost all appointment powers are also vested in the President.

As Chief Executive, the President seems to have a powerful advantage in policymaking. In reality, the President's control over agencies is far less extensive than a CEO's control of a corporation. Because Congress controls the budget, the President lacks funds to pay for programs and authority to sanction agencies by withholding appropriations; legislation controls expenditures. Furthermore, legislation determines administrative structure and process (McCubbins, Noll, and Weingast, 1987, 1989). Bureaucratic structure and process is created with the aim of making bureaucratic agencies responsive to the will of the legislature, not just the President (Weingast and Moran, 1983; McCubbins and Schwartz, 1984; McCubbins, 1985; McCubbins and Page, 1987; Calvert, Moran, and Weingast, 1987; Epstein and O'Halloran, 1999).

The disparity between Congress and the President is exemplified by comparing the Office of Management and Budget (OMB) in the Executive Office of the President (Pfiffner, 1979; Moe, 1985), the General Accountability Office (GAO) and Congressional Budget Office (CBO), which work directly for Congress. GAO investigates federal programs and audits expenditures, while the CBO estimates the budgetary impact of proposed legislation and provides economic expertise about anticipated revenues and expenditures of cyclically sensitive policies. Both CBO and GAO also provide economic evaluations of specific policies. OMB performs the same functions and prepares the President's annual budget, but despite its formal location in the Executive Office of the President, it exists at the pleasure of Congress and is much smaller than the GAO. As much as anything else, OMB aids Congress in formulating the budget—if it did not serve this purpose, it would not exist (Heclo, 1984; Kiewiet and McCubbins, 1991).

[18] See Cronin (1980) on Congress limiting the President's use of executive agreements.

5.2.4. Appointments

The President controls the top administrators in the executive branch. Senate confirmation is required of nominees for most top posts. Because Senate approval is virtually always granted, even in times of partisan division between the President and the Senate, many conclude that the President determines the policy preferences of political appointees (Moe, 1985), but others conclude that the Senate has considerable influence (Snyder and Weingast, 2000). The fact that Presidential nominees are rarely rejected does not necessarily mean that the Senate does not influence appointments. In some cases, at least, rejection of an appointment is costly to the President. For example, the process of rejecting a nominee gives the President's opponents a very public forum for criticizing the policies and judgment of the President. Moreover, a rejected nominee's embarrassment makes potential nominees more reluctant to let their names be sent forward. Hence presidents usually succeed in appointments, but the reason may be that they allow the Senate to have influence.[19]

The President can fire most high-level officials, the exceptions being political appointees to independent agencies and positions that are reserved for the civil service. The President can influence the civil servants who can not be fired by controlling the allocation of bonuses among the Senior Executive Service and their promotion to the Senior Executive Service, which are the jobs with the greatest prestige and highest pay.

The President's authority to make appointments is an important source of policy influence, but is nevertheless subject to the same limitations that apply to executive orders and agreements. While a President can pick executives who prefer particular policies and fire those who do not, the decisions of Presidential appointees are constrained by statutory mandates, the appropriations process and administrative procedures. Here the power of the President is more negative than positive: an agency can slow down implementation of a program or fail to spend all of its appropriations, but it is unlikely to succeed in carrying out policies that are not supported by its statutory mandate or the requirements of its statutory decision-making procedures.

5.3. Assessing the role of the president

PPT of the role of the President provides good news and bad news. The good news is that the Constitutional separation of powers achieves two useful ends: the system of checks and balances grants considerable power to influence the law, and the incentives created by the method of electing the President counteract the excessive responsiveness to particularistic interests in the legislature. The bad news, of course, is that the law-making powers of the President are not as strong as those of the legislature. In essence,

[19] As McCarty (2004) shows, the willingness of Congress to delegate authority to an agency depends on the harmony of preferences between the agency and Congress. If an appointee reflects only the President's interest, Congress will delegate less authority to the agency. In some cases the President can gain greater authority by appointing someone who is more compatible with congressional interests.

the President has the authority to use the office to influence Congress and even to make policy within a range of discretion that Congress will tolerate, but, in the end, the role of Congress in making law is dominant. Presidents can constrain pork, rent-seeking and morselization, and within limits can push policy in the direction of economic efficiency and universalistic goals, but they cannot prevent these inefficient activities.

An important issue with respect to the power of the President is the degree to which the President controls the bureaucracy. Because legislation frequently contains broad delegations of authority to agencies, the ability of the President to impose more universalistic objectives in policy implementation turns the responsiveness of the bureaucracy to presidential policy preferences. To this issue we now turn.

6. The bureaucracy

During the 20th Century, the debate over the nature of the bureaucracy pitted Weberians (Weber, 1946) and Progressives (see Landis, 1938; Mashaw, 1985a, 1985b, 1994, 1997; Moe, 1987, 1989), who favored giving substantial law-making power to elite civil servants, against Democrats (not the party, but a school of thought about the democratic legitimacy of delegation to the bureaucracy), who favor popular and legislative sovereignty (Shapiro, 1964; and Woll, 1977). The issue animating the original debate between Progressives and Democrats was whether professional experts without the encumbrances of political pressure should undertake administration, or whether elected officials should take as much control as possible of the details of policy, only reluctantly delegating authority to bureaucrats and then only with detailed instructions and safeguards to prevent the bureaucracy from seizing control of policy.

Shapiro (1964, p. 45) summarizes the point of contention between Progressives and Democrats: "Somewhere in the examination of every agency of American government, we may wish to ask to what extent the structure and function of this agency accords with whatever theory of democracy we have." Woll (1977) offers the typical worry about the expanded role of the bureaucracy advocated by the Progressives. "In this respect the development of a bureaucracy that is not elected and that exercises broad political functions has apparently resulted in the breakdown of a primary constitutional check on arbitrary governmental power" (p. 29).

The debate between Progressives and Democrats, though framed in normative terms, is rooted in a disagreement about the positive theory of relationships among citizens, elected officials and bureaucrats. Whether elected officials *should* delegate authority to an expert bureaucracy hinges on another question. As a matter of positive analysis, *can* elected officials control the bureaucracy in the sense that the policy preferences of the bureaucracy reflect the policy agreement among legislators that gave rise to the statutory law that empowered the agency? If the answer to this question is yes, then elected officials can enlist the technical expertise of civil servants without ceding to them control of policy.

Whether the control of the bureaucracy by elected officials is desirable in turn hinges on two questions addressed in preceding sections. First, as a matter of positive analysis, do elections and the legislative process cause statutes to reflect the preferences of citizens? If the answer to this question is yes, then the decisions of bureaucrats are responsive to citizens. Second, as a matter of normative analysis, are the preferences of citizens as reflected in elections normatively compelling? If the answer to this question is yes, then bureaucratic delegation is normatively compelling as well since the chain of arguments implies that delegation marshals the skills of analysts to advance the normatively attractive goals of citizens.

The schools of thought about bureaucracy differ according to how they answer each of these questions. Explicating these differences and the insights that PPT brings to this debate are the subjects of this section.

6.1. Schools of thought on bureaucratic autonomy

There are five distinct modern schools of thought with respect to the debate about the desirability of delegating policy-making (hence law-making) authority to a bureaucracy. These five schools are linked to the eight general schools of legal thought discussed in Section 2. Progressives and their New Deal successors argued that the bureaucracy is the only forum in which technocratic, scientific, apolitical policy-making is feasible. Landis (1938), a leading Realist, argued: "The administrative process is, in essence, our generation's answer to the inadequacy of the judicial and the legislative processes." This school favors broad, vague delegations to a bureaucracy populated by civil servants who are hired and promoted on the basis of merit.

In the early 1950s, Pluralists, exemplified by Truman (1951) and building on the work of Bentley (1908), replaced Progressives as the dominant school of thought about the role of bureaucracy. The premises of Pluralism are similar to the Sociological Jurisprudential School and Realists in that they believe that law is policy and that making policy is necessarily political. Pluralists also believe that bureaucrats (and the courts) are competent to make political decisions that serve a general public interest. To Pluralists, the bureaucracy is just another arena where groups compete and communicate their interests so that bureaucracy is a mechanism for forging deals among competing social interests. Pluralists view this competition as taking place in a political environment in which power is distributed among antagonistic interests, so that the outcome is likely to be a compromise of interests that serves the social good.

Four newer schools of thought have responded to different components of the optimistic picture painted by the Progressives and Pluralists. These are Public Choice, Civic Republicans, New Progressives and Neodemocrats.

The view of Public Choice about bureaucracy is derived from their skeptical view about democracy. Public Choice scholars argue that special interests, not public interests, capture the benefits of government intervention (Buchanan and Tullock, 1962; Kolko, 1965; MacAvoy, 1965; McConnell, 1966; Lowi, 1969; Stigler, 1971). These scholars extend this argument to bureaucrats by regarding them as another device for

creating and allocating rents, either for their own benefit (Niskanen, 1971) or for the benefit of elected officials (Stigler, 1971) to the detriment of economic efficiency and society as a whole. Mashaw (1997, p. 4) has characterized this view of American politics "somewhat hyperbolically, as a world of greed and chaos, of private self-interest and public incoherence. It is this vision that provides the primary challenge for today's designers of public institutions; for it is a vision that makes all public action deeply suspect."

Public Choice scholars level two main criticisms against bureaucracy. First, bureaucrats, in maximizing their personal welfare, have bargaining power over elected officials and use this power to extract budgets that are in excess of the amounts necessary to provide services (Niskanen, 1971). Second, special interests dominate agencies, because either the bureaucrats or their elected over-seers sell policy to the highest bidder. The inevitability of bureaucratic implementation costs leads these scholars to advocate strict limits to the size and scope of government and "undelegation" of legislative authority to avoid selfish misuse of discretion (Lowi, 1969).

While Public Choice is by no means the dominant school of thought about either bureaucracy or democracy, its critiques have been taken seriously by scholars of other schools. The other responses to Progressives and Pluralists actually accept some aspect of the Public Choice critique, but place it in a broader context that softens or even reverses its harsh conclusions about the efficacy of democratic government.

Two new forms of Progressivism resurrect the social desirability of bureaucracy while incorporating at least the possibility of democratic pathologies as put forth by Public choice scholars. The first is called Civic Republicanism and the second is the New Progressivism.

In a twist on Jacksonian Republicanism, Sunstein (1990) and Seidenfeld (1992) argue that the bureaucracy can lead citizens in policy deliberation and, moreover, in doing so can instill "civic republicanism." Civic Republicans see democracy as coming in two flavors. Day-to-day politics is not carefully followed by most citizens, and as a result is capable of the pathologies noted by the critics of democracy, whether Public Choice or the others that were summarized in Section 3. But "deliberative democracy" arises when citizens think seriously about policy and engage in public investigation and discussion about the consequences of alternative policy actions. Civic Republicans argue that in deliberative processes, citizens are not as likely simply to pursue narrow, short-sighted personal interests, and more likely to take into account the general welfare of society. Thus, one task of society is to maximize the extent to which policy is the outcome of deliberation.

Civic Republicans view delegation to properly designed agencies as a mechanism for creating deliberative democracy. Specifically, Seidenfeld argues that "given the current ethic that approves of the private pursuit of self-interest as a means of making social policy, reliance on a more politically isolated administrative state may be necessary to implement something approaching the civic republican ideal."

Two positive theoretical hypotheses underpin the normative prescription of Civic Republicans. First, elections and the process of law-making by elected officials do a poor

job of transmitting citizen preferences into statutes. In this regard Civic Republicans resemble Public Choice in rejecting the optimism of Pluralists. Second, a largely independent bureaucracy that must satisfy procedural requirements to interact with citizens produces decisions that are more responsive to citizen preferences. In this case, Civic Republicans reject Public Choice and resemble Pluralists in that they emphasize the representation of citizen interests within the bureaucratic process rather than the technical expertise of a well-selected civil service.

New Progressivism, most completely explicated by Mashaw (1985a, 1985b, 1994, 1997), also rejects pessimism about bureaucracy as a necessary feature of delegation. Mashaw (1997, p. 206) argues that agencies can be constructed to be competent and responsive to public desires, but not because the process is deliberative. A distinctive feature of New Progressivism is that it recognizes that not all bureaucratic delegations do lead to policy implementation that serves a plausible definition of the public interest. But New Progressives tend to see these examples as exceptions that can be avoided.

One cause of bureaucratic failure is simply mistakes—errors by elected officials in setting up the procedures and powers of agencies (Breyer, 1982, 1993; Mashaw, 1983). Here the solution is not unlike the prescription advocated by Civic Republicans: elected officials should take greater care (engage in more deliberation) in designing policies. The other source of bureaucratic failure is invisible day-to-day involvement of elected officials in the affairs of agencies, typically in responding to a demand for service from an unhappy supporter (Mayhew, 1974; Mashaw, 1994). This problem can be mitigated by ensuring that oversight is rare and politically visible, such as by enacting sunset provisions and making multi-year appropriations and authorizations. Thus, New Progressives propose that agencies should have broad and vague mandates and that Congress should exercise more care in designing policies and methods for oversight.

Neodemocrats (not the contemporary branch of the party, but a reference to a school of thought) agree with Pluralists and Democrats that the bureaucracy is political.[20] Unlike Progressives and Pluralists, Neodemocrats agree that excessive or uncontrolled delegation undermines democratic legitimacy (Shapiro, 1964; Melnick, 1983). But unlike Democrats and like Progressives and Pluralists, Neodemocrats see delegation as a potentially valuable way to negotiate political compromises and to bring technical expertise to making law, and therefore as a necessary part of modern government. In short, Neodemocrats see delegation as having costs (as emphasized by Democrats and Public Choice) and benefits (as emphasized by Progressives and Pluralists).

The distinctive feature of Neodemocrats is that they also argue that elected officials can and do control the policies that are pursued by agencies (early examples are Wilmerding, 1943 and Fenno, 1973). These scholars focus their attention on studying how democratic, principally legislative, control of the bureaucracy comes about.

[20] The term "Neodemocrat" emphasizes that this group favors popular control of administration. These scholars adopt the Progressives' premise that political problems can be mitigated through the design of political institutions, so they could be called Neoprogressives. We eschew the latter to avoid confusion with the self-proclaimed New Progressive School, discussed below.

The recent work uses PPT to analyze the structure and process of legislative delegation (see, for example, Fiorina, 1977a; Cohen, 1979; Wilson, 1980; Fisher, 1981; Weingast and Moran, 1983; McCubbins and Schwartz, 1984; Weingast, 1984; McCubbins, 1985; McCubbins and Page, 1987; Moe, 1987; Noll, 1983; McCubbins, Noll, and Weingast, 1987, 1989; Kiewiet and McCubbins, 1991; Bawn, 1995, 1997; Epstein and O'Halloran, 1996).

The debate amongst scholars of delegation and the bureaucracy revolves around the efficacy of democratic institutions. Progressives and their recent offshoots see elections and elected officials as at best capable of providing only general directions about policies, but unable to do well in specifying the details (c.f. Abramson, 1994 and Posner, 1995). Citizens and elected officials lack the information available to administrative agencies. Elected officials, once they get past the general goals of policy, are susceptible to capture by a special interest when they tackle the largely invisible tasks of designing policy details and engaging in day-to-day oversight of agencies. Due to these limitations of the democratic system, old and new Progressives argue that the details of policy implementation should be delegated to apolitical bureaucratic experts.

This conclusion is directly at odds with that of Public Choice, which sees bureaucracy as a seeker of rents and server of special interests. This conclusion also is at odds with the analysis of Neodemocrats, who agree that broad bureaucratic discretion represents an abdication of a legislative responsibility and allows the usurpation of popular sovereignty. But Neodemocrats also argue that elected officials design agencies so that generally their objectives are served. Whether this political control of bureaucratic decisions works for good or ill depends on whether the goals that the legislature pursues and embeds in agencies are responsive to citizens.

The normative and positive debates regarding the role of the bureaucracy in policy-making closely parallel each other. Weber and the Progressives argue that policy-making is best left to apolitical, appointed administrators, because politicians lack the expertise, patience and public spirit necessary to make good public policy. In line with this normative argument is positive analysis claiming that much of the modern American bureaucracy is independent of legislative and executive oversight and control. Much of Public Choice scholarship accepts the positive argument that bureaucrats have great autonomy, but then claims that bureaucrats use their unbridled leeway in policy-making to allow themselves to be captured by special interests, to shirk their duties, and to engage in corruption (Tullock, 1965; Niskanen, 1971).

PPT seeks to develop a theory of bureaucratic behavior that takes into account both the objectives of elected officials in delegating policy-making authority and the instruments available to elected officials to solve the agency problem that accompanies delegation. In this sense, PPT is most closely aligned with the view of delegation put forth by Neodemocrats. The resulting theory describes how the Congress and the President influence bureaucratic law-making, which has led to a new view of administrative law. We now turn to a review of this work.

6.2. PPT of administrative law

Why would elected representatives allow bureaucrats to act autonomously, especially if they implement policy in a corrupt manner? Or, for that matter, why would legislators, who want to deliver particularistic benefits to selected constituents, delegate power to a scientific bureaucracy that will ignore these preferences in pursuit of economic efficiency and distributive justice?

Many scholars argue that Congress and the President are either incapable or unwilling to oversee and control bureaucratic policy-making (Ogul, 1976; Fiorina, 1977a; Dodd and Schott, 1979). Congressional incentives and capabilities are poorly matched, such that the management resources available to the elected branches of government are woefully inadequate relative to the size of the task of overseeing the bureaucracy (Aberbach, 1990). Fiorina (1979) provides a valuable insight about this perspective. He argues that Congress is clearly capable of controlling the bureaucracy, but that it may have no incentive to do so. Indeed, Fiorina emphasizes that for some policies the reelection goals of legislators give them no incentive to work for coordinated control of the bureaucracy. Why should Congress take political chances by setting detailed regulations that are sure to antagonize some political actor or constituent? When it comes to controlling the bureaucracy, electoral incentives lead representatives to "let the agency take the blame and the Congressmen the credit" (Fiorina, 1979, p. 136).

Democrats favor representative policy-making. Because they believe that the bureaucracy cannot be bridled, they also believe that legislative delegation should be avoided (Lowi, 1969; Stewart, 1975; Aranson, Gellhorn, and Robinson, 1982). Neodemocrats model bureaucratic policy-making as part of a game between Congress, the President, the courts, the bureaucracy, and the public. In this policy game, the bureaucracy's discretion is *conditional* (Fiorina, 1981a, 1981b; Weingast and Moran, 1983; Calvert, Moran, and Weingast, 1987; Moe, 1987). Under some conditions bureaucratic decisions will align closely with Congress' or the President's wishes (or both), while under other conditions they will not, depending on the incentives of legislators.

Delegation of legislative authority to the executive thus presents something of a dilemma. To capture the benefits of specialization and the division of labor as explained by Weberians and the benefits of bargains among interests as discussed by Pluralists, members of Congress delegate, therefore sacrificing some control. In so doing, they may in turn sacrifice the public interest as the agency empowered through delegation may be both unaccountable to elected officials and either captured by special interests or its own selfish objectives, as argued by the more pessimistic Realists. Alternatively, as New Progressives see it, a corrupt Congress sacrifices the opportunity to sell policy to special interests by delegating to scientific elites in pursuit of the public interest. In either case, the goals of the legislature are sacrificed through delegation. Yet despite the potential problems, elected officials have opted to delegate on a massive scale.

6.2.1. Why elected officials delegate

The basic question that PPT seeks to answer is why elected officials choose to engage in extensive delegation. In a sense, the answer is obvious: *elected officials are not as afraid of the potential gap between the goals of elected officials and the outcome of bureaucratic decisions as the scholars who emphasize the depth of the agency problems arising from delegation*. Subsidiary questions that PPT has recognized as important to understanding why elected officials delegate are when (i.e., under what conditions) do bureaucrats enjoy some degree of discretion in policy-making, how much leeway are they be able to exercise, when and how does Congress, the President, or the courts, singly or jointly, influence the decision-making of bureaucrats, and how do the delegation strategies of Congress, the President, and the courts change under conditions of divided government, unified government, and partisan realignment?

In answering these questions, research has looked beyond the overt methods of managing bureaucratic behavior, such as appointments, salaries and oversight, which many would agree are not sufficient by themselves to control delegations to the bureaucracy. Instead, scholars have emphasized budgetary control (Wildavsky, 1964; Kiewiet and McCubbins, 1991), appropriations riders (Kirst, 1969), Presidential and OMB leadership and oversight (Moe, 1987; Sundquist, 1988; Moe and Wilson, 1994; Wood and Waterman, 1994), judicial review and deck-stacking procedures (Noll, 1971, 1985; McCubbins, 1985; McCubbins and Page, 1987; McCubbins, Noll, and Weingast, 1987, 1989), and even external pressures, such as competition from other agencies. These devices include ex post reward-and-sanction mechanisms, which operate through what Weingast (1984) calls "the law of anticipated reactions," as well as *ex ante* institutional mechanisms that change the costs and benefits of taking various actions, thereby channeling agency decision-making (McCubbins, Noll, and Weingast, 1987).

6.2.2. Delegation and agency theory

PPT introduced the analogy of agency theory to thinking about legislative delegation to bureaucrats (e.g. Weingast, 1984). Abstractly, delegation is a "principal-agent problem." The principal is the person who requires a task to be performed, and the agent is the person to whom the principal delegates authority to complete that task. In all delegations, a necessary condition is that the principal must gain some advantage from delegating, such as involving more people in executing a demanding task or taking advantage of an agent's specialization or expertise. Delegation always brings disadvantages in the form of agency losses and agency costs. Agency losses are the principal's welfare losses when the agent's choices are sub-optimal from the principal's perspective. Agency costs are the costs of managing and overseeing agents' actions (including the agent's salary, and so on).

Three conditions give rise to agency losses, and thus the delegation dilemma. The first condition is that the agent must have *agenda control*. That is, the principal delegates to the agent the authority to take action without requiring the principal's informed consent

in advance.[21] This puts the principal in the position of having to respond to the action ex post after its consequences are observed, rather than being able to veto it *ex ante* on the basis of accurate expectations about its likely effects. The second condition is a *conflict of interest* between the principal and the agent. If the two have the same interests, or if they share common goals, then the agent will likely choose an outcome that the principal finds satisfactory. The third condition is that the principal *lacks a fully effective means of correction*, in the sense that the principal cannot overturn the decision after the agent makes it without incurring cost. Conventionally, the lack of an effective means of reversing the agent's decisions frequently is said to be due to the agent's expertise—the agent is chosen because of expertise, so the principal must acquire expertise or hire another expert to evaluate and then alter the agent's choice.

Members of Congress may lack an effective check over agency decisions because of the separation of legislative powers (held jointly by Congress and President) and executive power (held by the President, but supervised by the Congress). This sets up the so-called "multiple principal" problem. The legislative process in the United States ensures that the consent of at least majority coalitions in the House and Senate, plus either the President or additional members of both chambers, is given before a proposal becomes law. Because these principals must all agree to legislation—even legislation to check an agency's actions—the agency may be unconstrained within some sphere of activity. Broad agency discretion may exist even if all principals match the agency's expertise. The breadth of agency discretion depends on the extent of conflicting interests among the many principals. An agency needs only to make a single "veto player" (someone who can block legislation) sufficiently happy to sustain the agency's policy against an override or other form of punishment (McCubbins, Noll, and Weingast, 1989; Ferejohn and Shipan, 1990; Gely and Spiller, 1990, 1992; Ferejohn and Spiller, 1992).[22]

Agencies take many types of actions, such as proposing rules and adjudicating cases. Often these actions are taken without the appearance of congressional oversight, and therefore many deem these bureaucrats as unaccountable. Of course, when agencies make decisions, their actions are not necessarily final. Congress can overturn their decision by passing new legislation, which can be as simple as a brief rider on an appropriations bill that orders the agency not to spend any funds enforcing a particular rule. Even though Congress does not frequently override agencies, the possibility that they can do so creates an incentive for the agency to take the preferences of members of Congress into account. In a similar fashion, the threat of rewarding or sanctioning an agency for its actions may also create incentives for the agent to respect the wishes of members of Congress. These factors constitute an *ex post* form of control, by which is meant possible actions that can be taken after the agency has made a decision. The next section explores how ex post controls resolve some aspects of the delegation dilemma.

[21] Informed consent means that the principle possesses at least as much information as the agent about the consequences of the action.

[22] Of course, the President, courts and often individual House and Senate committees have the ability to unilaterally reject a proposal or punish an agent.

6.2.3. Solving the agency problem: ex post corrections and sanctions

The first major source of the delegation problem is the fact that agencies often possess an "institutional" advantage, in that the agencies collectively make voluminous decisions, and Congress must pay potentially large costs to respond legislatively. The agency's "first-mover" advantage potentially puts Congress in the position of facing a *fait accompli* from an agency. One important countermeasure by the legislature to mitigate bureaucratic agenda control is institutional checks. Operationally, institutional checks require that when authority has been delegated to the bureaucracy, at least one other actor has the authority to veto or block the actions of the bureaucracy. Before *Chadha* undid the process, Congress used the *ex post* legislative veto to check agency discretion. The legislative veto allowed the House and Senate, and in some instances either one alone, to veto bureaucratic policy proposals before they were implemented (Fisher, 1981).

Other *ex post* mechanisms add up to what has been referred to as "the big club behind the door" (Weingast, 1984). In addition to the threat to eliminate an agency altogether, Congress can make use of numerous checks on agency implementation. Congress can also make life miserable for an agency by endless hearings and questionnaires. For political appointees with short time horizons, this harassment can defeat their purpose for coming to Washington. In sum, ex post sanctions provide ex ante incentives for bureaucrats to avoid those actions that trigger them; and the best way to avoid them is to further congressional interests. Congress can also reduce the agency's budget or prohibit the use of funds for particular purposes or policies.

Similarly, enabling legislation (describing the nature of the delegation to the agency) can establish Presidential vetoes over proposed rules, or can grant only the authority to propose legislation to Congress. Another form of veto threat is an appropriations rider that prevents implementation of the agency's decision, whereby Congress can undermine a decision without rejecting it outright (Kirst, 1969).

In making proposals and engaging in rule-making, bureaucratic agents must anticipate the reaction of political leaders and accommodate their demands and interests. In discussing Congress, Weingast (1984, p. 156) notes: "Ex post sanctions ... create ex ante incentives for bureaucrats to serve congressmen." That is, Congress's big club engenders the well-known *law of anticipated reactions*, whereby bureaucrats are aware of the limits to acceptable behavior and know that they run the risk of having their agency's programs curtailed or careers ended if they push those limits too far.

Bureaucratic expertise relative to Congress often cited as the reason that delegation leads to loss of political control and accountability. But the problem is not that legislators lack information or that bureaucrats monopolize it. Legislators have access to sources of information and expertise on technical subjects from sources outside of the bureaucracy, such as legislative staff, constituents, interest groups, and private citizens, as well as from their own expert agencies, CBO and GAO. Rather, the problem is that gathering and evaluating information is costly, and the presence of costs to discover non-complying behavior inevitably causes Congress to regard some potential

non-complying behavior to be not worth the cost of detecting and correcting. The proper response to this problem by Congress is to find cost-minimizing methods to monitor agencies, to which we now turn.

6.2.4. Solving the agency problem: oversight

The information requirement for evaluating policy implementation is sometimes interpreted to mean that in order to ascertain whether an agency is doing its job, political leaders must engage in proactive oversight: they must gather enough information to assess whether an agency is producing good solutions to the problems that it confronts. This idea is false, however. Legislators do *not* need to master the technical details of policies in order to oversee effectively an agency's actions. Legislators need only to be capable of collecting and using enough information to reach reasonable conclusions about whether an agency is serving their interests. Moreover, if legislators *can* engage in effective oversight, they need not always *actually* engage in oversight to cause agencies to take their preferences into account in making decisions. The probability of detecting noncompliance with legislative purpose need not be 100 percent to cause agencies to ponder the risk of noncompliance.

Congressional oversight takes two forms: "police patrol" and "fire alarm" (McCubbins and Schwartz, 1984). In the former, members of Congress actively seek evidence of misbehavior by agencies, looking for trouble much like a prowling police car. In the latter, members wait for signs that agencies are improperly executing policy: members use complaints to trigger concern that an agency is misbehaving, just as a fire department waits for citizens to pull a fire alarm before looking for a fire.[23] Conventional wisdom nearly exclusively assumes that oversight is of the police patrol form.

Fire-alarm oversight has several characteristics that are valuable to political leaders. To begin, leaders do not have to spend a great deal of time looking for trouble. Waiting for trouble to be brought to their attention ensures that if it exists, it is important enough to cause complaints. In addition, responding to the complaints of constituents allows political leaders to advertise their problem-solving role and to claim credit for fixing problems (Fiorina and Noll, 1978). In contrast, trouble discovered by patrolling might not concern any constituents and thus would yield no electoral benefit. Thus, political leaders are likely to prefer the low-risk, high-reward strategy of fire-alarm oversight to the more risky and costly police-patrol system.

The logic of fire-alarm oversight can be incorporated into the one-dimensional model of policy choice, and shown in Figure 6.1. Let M represent the policy goal of the legislation that gives an agency its mandate, and A represent the preferred policy of the agency. Also assume that the enabling legislation grants standing in the process of the agency to a group that has a most-preferred policy of M. Thus group, at some cost K,

[23] The intuition behind fire alarm oversight has also been formally modeled in the context of the judiciary's appeals process (Shavell, 2004).

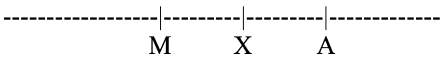


Figure 6.1. Controlling agencies with fire alarms.

can report the agency's policy deviation for the purpose of having it restored to M. Let X be defined so that the difference to this group in the value of a deviation from M to X equals K. Thus, if the agency attempts to adopt its preferred position, the group has more to gain by challenging the decision than the cost of doing so, and so will pull the fire alarm. If the agency adopts any policy between M and X, the group will not find a challenge worthwhile. Hence, the agency can move policy to X, but not all the way to A. Whether fire-alarm oversight is preferred by political actors to police patrol oversight depends on the relative magnitude of the cost saving from the former compared to the loss of ability to detect the smaller deviations that the watchdog group does not regard as significant enough to be worth challenging.

This theoretical model has two implications. First, the oversight process induces decision-makers to make decisions that are close to the democratically legitimate outcome M (at least within the range governed by the cost of an appeal). Second, unless an agency makes a serious error in estimating the stakes of the group that can pull the fire alarm, inducing compliance is costless because the fire alarm does not actually need to be pulled. If the agency accurately anticipates the response of the fire-alarm group, it will pick an outcome that does not generate an incentive to mount a challenge.

6.2.5. Solving the agency problem: administrative procedures

The mechanics of fire-alarm oversight are embedded in the administrative procedures of agencies that are within the jurisdiction of the Administrative Procedure Act (APA) of 1946 (McCubbins, Noll, and Weingast, 1987, 1989), as amended by further legislation and as extended and interpreted by the courts. First, an agency cannot announce a new policy without warning, but must instead give "notice" that it will consider an issue, and do so without prejudice or bias in favor of any particular action. Second, agencies must solicit "comments" and allow all interested parties to communicate their views. Third, agencies must allow "participation" in the decision-making process, the extent of which is often mandated by the statute creating the program. When investigative proceedings are held, parties can bring forth testimony and evidence and often may cross-examine other witnesses. Fourth, agencies must deal explicitly with the evidence presented to them and provide a "rationalizable" link between the evidence and their decisions. Fifth, agencies must make available a record of the final vote of each member in every proceeding. Failure to follow any of these procedures creates a cause of action to appeal the agency's decision to the courts.

As legal scholars have long observed, these requirements have obvious rationales in procedural due process, but beyond rights of due process, they also have profound political implications (McNollgast, 1999). These requirements force agencies to collect

information to guide its decisions, but this goal could be achieved in much less elaborate ways—including judicial review on the basis of the balance of evidence supporting the agency's position. The important additional insight about these procedures is that they facilitate the political control of agencies in five ways.

(1) Procedures ensure that agencies cannot secretly conspire against elected officials to present them with a *fait accompli*, that is, a new policy with already mobilized supporters. Rather, the agency must announce its intentions to consider an issue well in advance of any decision.

(2) Agencies must solicit valuable political information. The notice and comment provisions assure that the agency learns which relevant political interests are affected by its proposed decision and something about the political costs and benefits associated with various actions. That participation is not universal (and may even be stacked) does not entail political costs to members of Congress. Diffuse groups that do not participate, even when their interests are at stake, are much less likely to become an electoral force in comparison with those that do participate.

(3) The proceeding is public, thereby enabling political principals to identify not just the potential winners and losers of the policy but their views. Rules against *ex parte* contact protect against secret deals between the agency and some constituency it might seek to mobilize against Congress or the President.

(4) The sequence of decision-making—notice, comment, deliberation, collection of evidence, and construction of a record to support an action—creates opportunities for political leaders to respond when an agency seeks to move in a direction that officials do not like. By delaying the process, Congress has time to decide whether to intervene before a decision becomes a *fait accompli*.

(5) Because participation in the administrative process is expensive, it serves to indicate the stakes of a group in an administrative proceeding. These stakes are indicators of the resources the group can bring to bear in taking out political reprisals against congressional principals whom they hold accountable for policy outcomes.

These features of the APA reduce an agency's information advantage and facilitate fire alarm oversight. By granting rights of participation and information to interest groups, administrative procedures reduce an agency's information advantage. Congress uses the APA to delegate some monitoring responsibility to those who have standing before an agency and who have a sufficient stake in its decisions to participate in its decision-making process, and, when necessary, to trigger oversight by pulling the fire alarm. In addition, administrative procedures create a basis for judicial review that can restore the status quo without requiring legislative correction. As a result, administrative procedures cope with the first-mover advantage of agencies.

6.2.6. Solving the agency problem: ex ante controls

While *ex post* methods of controlling agencies are always present, utilizing them to respond to an agency decision requires legislative action. Some legislative action, such has oversight hearings (including those designed to harass administrators) can be done unilaterally. So too can single-chamber legislative vetoes, but unfortunately this approach has been severely curtailed by the Supreme Court in *Chada*. Fast-track treaties are now the only important sources of policy change which makes use of the one-house veto by either chamber.

When legislation is required to correct an agency, action must be taken by both chambers of Congress (and their committees) and the legislation must survive the possibility of a Presidential veto. Because multiple actors must assent in order to undertake successful legislative action, the agency will face looser constraints on its actions if the principals—i.e., majorities in the House and the Senate, and President—disagree among themselves. The agency needs only to make a majority in a single chamber happy with a policy choice to protect against a legislative *ex post* reversal.

The problem of the absence of the ability to engage in effective *ex post* correction of an agency decisions is shown in Figure 6.2. Here H, P and S represent the policy ideal points of the House, President and Senate, respectively, where H and S are the positions of the median voters in those bodies. Let SQ represent the status quo policy as contained in statutes, and let A represent the ideal point of an agency that is charged with implementing policy. The issue to be examined is the discretion available to the agency if it can adopt a policy without being detected by any of its political principals. If the agency adopts policy A, the Congress and the President agree that policy should be changed; however, the old outcome is not likely to be restored. Let A^* be the policy that the President regards as equally valuable as A. If Congress adopts any policy to the left of A^*, the President will veto the bill. If Congress can not muster a $2/3$ majority in both legislative branches to override the veto, then A will stand. Hence, the best that Congress can do in response to the non-complying adoption of A is to propose legislation at A^*.

If the agency rationally anticipates the response of Congress, it can do better than the ultimate result A^*. If the agency adopts the President's ideal point, P, the President will veto any attempt at correction, and P will then stand as the policy unless Congress overrides the President's veto of correcting legislation. Suppose Congress can override the President's veto for any bill that is to the right of V. In that case, the agency can guarantee V by either adopting it or adopting some policy to the right of V and letting Congress pass a veto-proof correction.

Figure 6.2. Agency power without ex ante oversight.

A variant of all of these results holds regardless of the relative positions of the four major players. All that is required for ex post correction to be inadequate is that the status quo legislative bargain differs from the preferred policy of the agency and that the three branches have different ideal points. The agency always has some discretionary power to move policy within the range of outcomes between the two extreme ideal points without fear of legislative correction or punishment.

This analysis explains why most administrative procedures that have been adopted are for the purpose of preventing non-complying behavior before it happens or through the courts, rather than correcting it after it occurs through legislation.

In creating and funding bureaucratic agencies, the legislature anticipates the problems just discussed. When delegating, legislators decide consciously whether to take steps to mitigate these problems. This section examines ways that members of Congress and the President can structure an agency's decision-making process so that it is more responsive to their preferences.

One important countermeasure that Congress and the President may take to mitigate the power of bureaucratic agenda control is the aforementioned strategy of employing institutional checks. Checks on agency agenda power can also be created so that they affect the agency's choice *ex ante*, that is, before it makes a proposal. In our earlier work (McCubbins, Noll, and Weingast, 1987, 1989) we argued that tools available to political actors for controlling administrative outcomes through process, rather than substantive guidance in legislation, are the procedural details, the relationship of the staff resources of an agency to its domain of authority, the amount of subsidy available to finance participation by underrepresented interests, and resources devoted to participation by one agency in the processes of another.

By structuring who gets to make what decisions when, as well as by establishing the process by which those decisions are made, the details of enabling legislation can stack the deck in an agency's decision-making. In effect, this is the same problem that we discussed earlier in terms of a legislative majority delegating to its agents the discretion to determine a policy agenda. We have argued that the winning coalition in Congress will use its ability to establish the structure and process of agency decision-making to fix the range of feasible policies. This, in turn, implies a definition for the agency's range of policy discretion.

For example, elaborate procedures, with high evidentiary burdens for decisions and numerous opportunities for seeking judicial review before the final policy decision is made, benefit groups having considerable resources for representation. When combined with the absence of a budget for subsidizing other representation or a professional staff for undertaking independent analysis in the agency, cumbersome procedure works to *stack the deck* in favor of well-organized, well-financed interests (Noll, 1983).

Congress and the President can use procedural deck-stacking for many purposes. A prominent example of how procedures were used to create a "captured" agency was the original method for regulation of consumer product hazards by the U.S. Consumer Product Safety Commission (CPSC). Although the CPSC was responsible for both identifying problems and proposing regulations, it was required to use an "offeror" process,

whereby the actual rule writing was contracted out. Usually the budget available to the CPSC for creating a regulation was substantially less than the cost of preparing it. Consequently, only groups willing to bear the cost of writing regulations became offerors, and these were the groups most interested in consumer safety: testing organizations sponsored by manufacturers or consumer organizations. Thus, this process effectively removed agenda control from the CPSC and gave considerable power to the entities most affected by its regulations (Cornell, Noll, and Weingast, 1976).

In 1981, Congress amended this process by requiring that trade associations be given the opportunity to develop voluntary standards in response to all identified problems, assuring that agenda control was never granted to consumer testing organizations. The 1981 legislation illustrates how procedures can make policy more responsive to a politically relevant constituency by enhancing that special interest's role in agency procedures.

The U.S. National Environmental Policy Act (NEPA) of 1969 provides another example of how this works. In the 1960s, environmental and conservation groups in the United States became substantially better organized and more relevant politically. By enacting NEPA, Congress imposed procedures that required all agencies to file environmental impact statements on proposed projects. This requirement forced agencies to assess the environmental costs of their proposed activities. NEPA gave environmental actors a new, effective avenue of participation in agency decisions and enabled participation at a much earlier junction than previously had been possible.

An example of the policy consequences of NEPA is its effects on decisions by the Nuclear Regulatory Commission (NRC) in licensing nuclear power plants (Cohen, 1979; McNollgast, 1990). NEPA gave environmentalists an entry point into the proceedings before the NRC for approving new projects. Initially the Atomic Energy Commission (the predecessor to NRC) asserted that it was exempt from NEPA, but the 1971 decision in *Calvert Cliffs* required the agency to follow NEPA's requirements, and thereafter environmental impact reports were a necessary part of the approval process.

Environmentalists used this entering wedge to raise numerous technical issues about the risks of components of nuclear reactors, thereby dramatically slowing down the licensing process. Although the interveners rarely won their contentions, their interventions succeeded in raising the costs of nuclear power plants so dramatically that no new plants were actually built. Between 1978 and 1995, no new nuclear plants were ordered, and moreover, every single project planned after 1974 was cancelled (as were a third of those ordered before 1974).

The 1972 California Coastal Zone Conservation Act required similar institutional checks. The statute's objectives were to protect scenic and environmental resources along California's coastline and to preserve public access to the beach. In this case, the key procedure was to grant numerous bodies a veto over proposed changes in land use in the coastal zone. Local governments were the first in line to approve or deny any proposed project, then one of the six regional coastal commissions, and then the statewide coastal commission reviewed all permits passed by the local governments. The commis-

sioners were also given the power to levy substantial monetary fines against violators, which helped induce compliance.

The creation of a permit review procedure with diffused power automatically biased the regulatory process against approving new coastal projects. By carefully choosing the *procedure* of the California coastal initiative, the state legislature was able to achieve its statutory goals to curtail further development even thought the statute contained a broadly-stated, seemingly balanced substantive mandate.

Perhaps the most important tool that legislatures use to stack the deck in bureaucratic decision-making is the establishment of the burden of proof. In all agency decisions, proof must be offered to support a proposal. The burden of proof affects agency decisions most apparently when the problem that is before the agency is fraught with uncertainty. In such a circumstance, proving anything—either that a regulation is needed to solve a problem, or that it is unnecessary—is difficult, if not impossible. Hence, assigning either advocates or opponents of regulation a rigorous burden of proof essentially guarantees that they cannot obtain their preferred policy outcome.

For example, the U.S. Federal Food, Drug, and Cosmetics Act of 1938, as amended, requires that before a pharmaceutical company can market a new drug, it must first prove that the drug is both safe and efficacious. By contrast, in the Toxic Substances Control Act of 1976, Congress required that the Environmental Protection Agency (EPA), before regulating a new chemical, must prove that the chemical is hazardous to human health or the environment. The reversionary outcome is that companies may market new chemicals, but not new drugs. The results of the differences in these two burdens of proof are stark: very few new drugs are brought to market in the United States each year (relative to the rates in other countries), while the EPA, by contrast, has managed to regulate none of the 50,000 chemicals in commerce under these provisions in the Toxic Substances Control Act.

Congress has successfully used modifications in the burden of proof to change the outcome of regulation. By requiring a certain actor to prove some fact in order to take a regulatory action, Congress can stack the deck against that particular actor's most preferred outcome.

The Airline Deregulation Act of 1978, amending the Civil Aeronautics Act from the 1930s, provides one example. Under the original act, in order to enter a new market by offering flights between a pair of cities, the prospective entrant bore the burden of proof to demonstrate to the Civil Aeronautics Board (CAB) that without entry, service would be inadequate. Thus, a potential entrant in a market that already was being served had the virtually impossible task of showing that someone who wanted service was being denied. In the Kennedy Amendments, Congress changed the procedure used by the CAB, shifting the burden of proof to the existing carriers to show that new entry would lead to less service. This modification now biased the process in favor of allowing entry, and against the old protections that had profited carriers for so long. As a result, airline entry was essentially deregulated.

More recently, when stories of abuses of power by the Internal Revenue Service came to national attention, Congress again responded by shifting the burden of proof. In this

case the burden shifted from taxpayers, who had been required to prove that they had not violated tax law, to the IRS, which now must prove that a taxpayer has violated a tax law. The shift in the burden of proof raises the cost of tax enforcement, and therefore reduces the number of tax claims that the IRS can file. The effect is to benefit taxpayers by forcing the IRS to abandon some enforcement actions that it would have filed under the old system. Again, this change in the administrative process stacks the deck in favor of one group of actors' preferred outcome.

Using administrative procedures as instruments to control the bureaucracy is part of a broader concept called *the mirroring principle* (McCubbins, Noll, and Weingast, 1987, p. 262). Political officials can use deck-stacking to create a decision-making environment in an agency in which the distribution of influence among constituencies reflects the political forces that gave rise to the agency's legislative mandate. As argued above, the enacting coalition faces large impediments to reforming and passing corrective legislation when an agency deviates from their intended policy. This coalition therefore has an incentive to use structure and process to enfranchise in the agency's procedures the constituencies it originally sought to benefit. This environment endures long after the coalition behind the legislation has disbanded. As a result, policy is more durable—therefore raising the credit due to legislators for enacting a statute that is more valuable to its proponents. Without policy durability, legislative victories would not be long lasting, and constituents would not be willing to reward legislators for policy change.

The point of mirroring and deck-stacking is not to pre-select policy, but to cope with uncertainty about the most desirable policy action in the future. Procedures seek to ensure that the winners in the political battle over legislation will also be the winners in the process of implementing the program. By enfranchising interests that are represented in the legislative majority, a legislature need not closely supervise the agency to insure that it serves its interests, but can allow an agency to operate on "autopilot" (McCubbins, Noll, and Weingast, 1987, p. 271). Policy then can evolve without requiring new legislation to reflect future changes in the preferences of the enacting coalition's constituents, and political principals can be more secure in using fire-alarm oversight of the agency.

Legislatures can further limit the potential mischief of agency agenda control by carefully setting the reversionary policy in the enabling statute that established the agency. For example, consider some entitlement programs specified by statute, such as Social Security and Medicare, in which the agency has no discretion over either who qualifies for assistance or how much they will receive. Another example is the widespread use of "sunset" provisions, whereby an agency's legal authority expires unless the legislature passes a new law to renew the agency's mandate.

Courts also play a role in the political control of the bureaucracy. Administrative procedures affect an agency's policy agenda only if they are enforced. The legislature can delegate enforcement to the courts, in which case procedure can affect policy with minimal oversight by politicians. For supervision by the courts to serve this function, judicial remedy must be highly likely when the agency violates its rules. If so, the courts and the constituents who bring suit guarantee compliance with procedural constraints, which in turn guarantees that the agency choice will mirror political preferences with-

out any need for active "police-patrol" oversight. Of course, for this process to work, the courts must be willing to ensure that agencies adhere to the requirements of their underlying statutory mandates and procedural requirements, which is the topic that is explored in Section 7.

PPT analysis of the political control of the bureaucracy does not provide protection against the most insidious potential problem with delegation, interest-group capture. PPT only argues that agencies are not a source of capture that is independent of the actions and goals of elected officials. If elected officials are a willing co-conspirator in agency capture, evidence that they influence policy will not assuage fears that the public interest is subverted. In this case, the structure of Congress provides some additional checks. The control committees in Congress, especially the appropriations and budget committees, serve to check capture by reducing any substantive committee's ability to act unilaterally (Kiewiet and McCubbins, 1991). That is, by requiring committee proposals to pass through the appropriations process, substantive committees can be disciplined by the appropriations committees' ability to reject their proposals. Recall that substantive committees are more likely to represent specific constituencies, but control committees are representative of the entire legislature and so protect each party's brand name to voters. Hence, capture of the latter is less likely.

Nevertheless, some policies do not require budget authority (such as regulations). If party leaders do not possess sufficient information or incentives to detect and to constrain capture when it emerges in legislation, the "iron triangle" among a constituency, an agency, and its oversight committees can emerge and be difficult to undo. In this case, understanding how capture arises is still useful, because it provides information about the likely performance of an agency while legislation is pending without requiring expertise about the substance of the policy.

Essentially PPT identifies the political conditions under which Congress is likely to create a captured agency (i.e., a policy that is of primary interest to a small constituency and of minor interest to others), and PPT of administrative law provides a check-list of procedures that facilitate capture of an agency. This check-list can be considered by party leaders, control committees (such as Rules), and other interests considering the implicit deck-stacking in proposed legislation. If members of Congress and their leaders choose to ignore this information, and thereby to let a small subset of their peers create a captured agency, delegation becomes abdication, but the condition under which it happens is that no one other than the favored interest groups cares very much that a captured agency is being created.

6.3. PPT of political control of the bureaucracy: summary

Delegation can succeed when one of two conditions is satisfied (Lupia and McCubbins, 1998). The first is the *knowledge condition*, which is that the principal, through personal experience or knowledge gained from others, can distinguish beneficial from detrimental agency actions. The second is the *incentive condition*, which is satisfied when the agent has sufficient incentive to take account of the principal's welfare. These condi-

tions are intertwined in that a principal who becomes enlightened with respect to the consequences of delegation can motivate the agent to take actions that enhance welfare.

The institutions that govern the administrative process often enable legislators both to learn about agents' actions and to create incentives for bureaucratic compliance, so that one or both of the conditions for successful delegation are satisfied. Legislators' implementation and reliance on these institutions is the keystone of successful delegation. These institutions imply that their day-to-day operation often goes unseen, but bureaucratic output still is affected (Weingast, 1984).

We are not arguing that all necessarily is well in the Washington establishment. Delegation produces agency losses and entails agency costs, and the sum of these can exceed the benefits gained from delegating. The interesting questions are when do the costs exceed the benefits and how can we tell? In any case delegation, while entailing some loss of control, is not equivalent to abdication of law-making authority by elected officials. Instead, delegation is just a cost.

Taken together, the findings of PPT suggest a new view of administrative law. Unlike the Civic Republicans, who see administrative law as bureaucratic-led democratic deliberation, unlike the Progressives, who see it as ensuring political and scientific quality, and unlike the New Progressives, who see it as creating procedural justice, PPT sees administrative law as a political choice that channels the direction of policy outcomes in a manner favored by those who write the laws. In this sense, administrative procedures, in general, are normatively neutral in that they can be used to create agencies that behave in any way the political principals desire, from enlightened experts seeking to benefit society through the provision of pgs to captured hacks doing the bidding of a particular interest as part of an iron triangle. In short, delegation is neither inherently good nor bad for democracy; its net effect depends on the details.

7. The courts

Most modern schools of legal thought turn to the courts to check and redress wrongs created by electoral and legislative processes. The preponderance of modern legal theory holds that the centuries old tradeoff between popular sovereignty and elite control weighs heavily in favor of elite—i.e., *judicial*—control of dispute resolution.

Should judges play a bigger or smaller role in creating and implementing governmental policy? What are the tradeoffs? Precisely what role for the judiciary produces the best policy outcomes? These questions—whether judges should be passive or active, or modest or aggressive—ought to be confronted head-on rather than obscured by endless talk about legitimacy (Posner, 1990).

PPT of the courts seeks to identify the conditions under which the court can exercise independent discretion, and when its authority is final and supreme. By necessity, PPT addresses when the court is not supreme, and instead acquiesces to or acts as an agent of the legislature and President. In addition, given the similarities between PPT scholarship

on the judiciary and the bureaucracy, PPT examines the issues of judicial independence and discretion in the same fashion.

As observed by Posner (1990) and Shapiro (1964), the conclusions of a positive analysis of the courts have implications for the normative debates between democrats and elitists. For example, our theory of the legislative process, and the evidence we presented, provides a means of assessing the premises of the various schools of legal and bureaucratic thought. These results push us to accept some and to reject other arguments about the role of the courts in statutory interpretation and judicial review of agency procedures. Further, by addressing when courts are supreme, and how supremacy is conditioned on institutional structure and procedure, we can demarcate limits to normative arguments about how law ought to be made. That is, we can comment on the plausibility of premises, and we may be able to address when these premises, and the theories built on them, are reasonable bases for judicial and bureaucratic reforms.

PPT provides a different view of the courts than the treatments in either legal scholarship or political science. Because PPT embeds courts in a political process, it shows how the courts interact with Congress, the President, and the bureaucracy.

7.1. PPT and statutory interpretation

Scholars of law and politics typically regard judicial decisions as subsequent to legislation. From this perspective courts are omnipotent actors, imposing any outcome they wish. This perspective also allows theories and recommendations concerning how judges ought to decide cases to be unconstrained. In statutory interpretation, for example, courts are free to make any interpretation they wish, perhaps based on normative principles of law, moral philosophy, policy preferences or ideology. A court that decides wrongly is at fault, and the corrective is to exhort the court to mend its ways.

PPT provides a different framework for analyzing courts by observing that statutory interpretation is an on-going process. Legal scholars are right to observe that a necessary condition for a statutory interpretation case to come before the courts is that Congress must pass a law. But they are wrong to assume that courts have the last word. Congress can act in response to judicial decisions, which implies that statutory interpretation is not a two step-process that ends with the judiciary, but an on-going process in which Congress and the courts interact repeatedly. PPT demonstrates that this change in perspective provides a very different way of understanding judicial decisionmaking in general, statutory interpretation in particular.

7.1.1. The strategic judiciary in PPT

To see the logic of the approach consider a one dimension spatial model and three actors, the President (P), the Congress (C), and the courts (J), and with status quo Q.[24] Consider the political configuration depicted in Figure 7.1.

[24] Marks (1988) initiated this form of analysis. See also Epstein and Knight (1998), Eskridge and Ferejohn (1992a, 1992b), Eskridge (1991) and Levy and Spiller (1994).

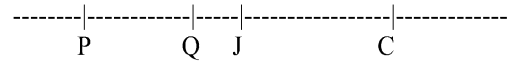


Figure 7.1. The power of courts in statutory interpretation.

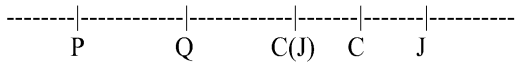


Figure 7.2. Constraints on an extremist court.

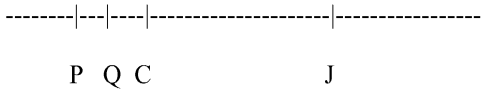


Figure 7.3. Courts facing political officials with closely aligned preferences.

Notice that every policy between P and C is a legislative equilibrium in that if any of these points is the status quo, any bill that is preferred by one makes the other worse off, so no legislation can pass. The point Q, therefore, is a stable legislative equilibrium.

Not all points between P and C are necessarily a policy equilibrium in a larger game on which the judiciary interprets the meaning of the law. Given their preferences and the latitude afforded them by their role as interpreters, the court will move policy from Q to its ideal point, J, which is a new stable equilibrium in the legislative process.

The court's ability to influence legislation depends on the configuration of preferences. If the court's ideal is outside the interval between P and C, as shown in Figure 7.2, the court is constrained by politics.

If the court attempts to implement its ideal policy, J, it will fail because J is not between P and C. If the court adopts J, both Congress and the President are better off moving policy back between their ideals—specifically, to any point between C(J) and C.[25]

These examples illustrate a general result. In a system of separation of powers, the range of discretion and hence independence afforded the courts is a function of the differences between the elected branches. If the branches exhibit little disagreement about the ideal policy, judicial discretion is low. Figure 7.3 demonstrates this point.

In this political setting, J stands the same relation to P as in the previous figures, but C is located much closer to P. Figures 7.1 and 7.2 might correspond to a case of divided government (different parties hold the two branches), while Figure 7.3 might represent united government (a united party holds the presidency and a majority in Congress). If the courts attempt to implement their ideal policy, J, under the conditions of Figure 7.3,

[25] C(J) is the policy that makes the median voter in Congress indifferent with J, imply that the median prefers all policies between C(J) and J to either C(J) or J.

they will fail. Both P and C prefer all points between their ideal policies to J. The best the court can do is to implement policy C. This configuration shows that courts freedom of action is highly constrained when it faces a relative united set of elected officials.

More generally, these results show how judicial independence depends on the political environment. Some judicial decisions located between P and C will stand in the sense that elected officials cannot reverse them. But other decisions will be reversed—those outside of P and C. To the extent that judges want avoid being overturned by Congress, they have an incentive to make strategic decisions; namely, decisions that take into account the political configuration so that their decisions will not be overturned.

7.1.2. Application to affirmative action

The above discussion left policy abstract. To show the power of these models to yield new conclusions, consider the evolution of an important policy area in the United States, expanding the meaning of civil rights legislation (see Weingast, 2002).

The landmark Civil Rights Act of 1964 forced Southerners to end their system of apartheid that suppressed African-Americans. Beginning in the early 1970s, a series of court cases expanded the meaning this act.[26] In brief, the civil rights act was an anti-discrimination law, requiring equal opportunity for all individuals regardless of race, creed, or gender. In a series of decisions in the 1970s, the Supreme Court expanded the meaning of the act to include a degree of affirmative action.

This phenomenon raises the political and legal question: Why did a conservative Supreme Court under Chief Justice Warren Burger expand civil rights? Undoubtedly the conservative majority on the Court preferred the status quo to this outcome. To solve this puzzle, Eskridge (1991) uses PPT models to explain that the answer lies in the interaction of Congress and the courts. Eskridge argues that the conservative Court expanded civil rights strategically. By taking modest steps to expand rights, the Court forestalled an even larger change in the scope of the law by Congress. The argument draws on the idea of the "filibuster pivot" (Brady and Volden, 1998; Krehbiel, 1998). Senate rules allow a minority of senators to defeat a bill by "filibustering," continuing the debate to prevent a measure from coming up for a vote. The Senate can end a filibuster only by a successful motion to end debate (cloture), which requires a super-majority of 60 positive votes.

To pass the 1964 bill required defeating a filibuster by southern Democrats. At that time cloture required obtaining support from two-thirds of the Senate. The policy setting depicted in Figure 7.4 reveals the effect of the filibuster.

In Figure 7.4, Q is the status quo, f is the ideal policy of the filibuster pivot (that is, the last Senator who must be brought on board to end debate), and M is the median legislator's ideal. Without a filibuster rule, policy would move from the status quo to M, the median voter's ideal policy. The filibuster pivot prefers all policies between Q and f(Q) to Q, but Q to all policies outside this region. Any policy outside of the interval

[26] Notably, *Duke Power* (1971) and *United States Steelworkers v Weber* (1979).

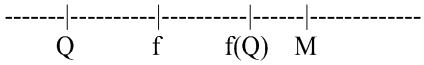


Figure 7.4. The effect of the filibuster.

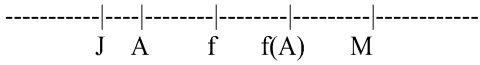


Figure 7.5. Civil rights policy.

[Q, f(Q)] makes the pivot worse off. If the Senate tries to pass the median senator's ideal, M, Senators who prefer Q to M will filibuster, and the majority favoring M will not be able to end debate. Thus, the biggest policy change that the Senate can pass is f(Q), which is the point nearest M that is filibuster-proof.

Much of the drama in the passage of the 1964 act concerned the parliamentary maneuvers to defeat the filibuster (see, e.g., Eskridge and Frickey, 1990; Graham, 1990; Whalen and Whalen, 1985; and Rodriguez and Weingast, 2003). To understand the transformation of civil rights by a conservative Supreme Court in the 1970s, consider the policy setting in Figure 7.5, where the set of policy alternatives represents the degree of federal support for civil rights, J represents the ideal policy of the conservative Supreme Court majority, A represents the policy enacted by the 1964 Act, f is the ideal policy of the filibuster pivot (a conservative Republican), and M is the median senator's ideal policy. As before, f prefers all policies between A and f(A).

The critical feature of the new political environment of the 1970s is that the median in Congress was far more liberal than the median in the 1964 Congress that passed the Civil Rights Act. Eskridge argues that the more liberal Congress would have passed new civil rights legislation, moving policy to the maximum that is feasible in the Senate, that is, from A to f(A).

In this setting, the Supreme Court acted first to preserve as much of the status quo as possible. By acting first, the Supreme Court moved policy from A to f. This move precluded any further move by Congress because any policy change from f toward M would make the filibuster pivot worse off.

This model has several implications. First, it shows the power of the model in specific policy settings to give new answers to important political puzzles.[27] More broadly, it shows the strategic role of the courts in the United States policymaking process. Courts are not the end of the process of policymaking and implementation; they interact with Congress and the president. This forces them to be strategic; failing to do so implies less influence and hence less force of their decisions.

[27] PPT scholars have used analyses of this type in many contexts. In addition to the references in the text, see Brady and Volden (1998) on the minimum wage, Ferejohn and Shipan (1989) on telecommunications policy, Riker (1982) on federal aid to education, Weingast and Moran (1983) on the FTC and Weingast (1984) on the SEC.

The political logic of PPT models implies that judicial decisions cannot solely be based on normative principles. Following normative principles alone requires that the courts ignore the political situation, implying that political officials will sometimes overturn their decisions. This political reality forces the courts to face a choice: either act strategically, and hence compromise their normative principles, or act according to principle but then have Congress overturn both the court's decision and the normative logic underlying it.

7.2. The courts and legal doctrine in a system of separated powers

The normative and positive debates regarding the role of the courts in policy-making closely parallel each other. Both debates rely on assumptions about congressional decisions and the efficacy of congressional oversight and control over the judiciary. The overwhelming consensus on the latter issue is that, except under rare circumstances, Congress and the President are unable, in the short run, to exert much influence over the Supreme Court's choice of legal doctrine. Others, such as Murphy (1962) and Rosenberg (1991) argue that management tools such as appointment power, budgets and selection of jurisdiction, which work effectively against the bureaucracy, have only limited effect on judicial incentives.

Missing from this debate is the approach implied by PPT of Law. The question pioneered by Shapiro (1964) and pursued in depth by Cohen and Spitzer (1994, 1996) is how the structure and process of the judiciary affect the Supreme Court's ability to influence legal doctrine. When will the Supreme Court have the ability to set legal doctrine and to have its doctrine implemented, and when will its influence be restrained? Under what conditions can Congress and the President affect legal doctrine through manipulating judicial structure and process?

Congress and the President have access to substantial mechanisms of control over the judiciary, which become apparent when one considers seriously the judicial system as a whole, and not just the Supreme Court in isolation. To see this, we consider our earlier discussion (McNollgast, 1995) showing an indirect route of political influence: by changing the structure of the federal judiciary, Congress and the President can bring potent influence to bear on the Court.

In response to political and partisan considerations and constraints, the elected branches manipulate the size and jurisdiction of the federal judiciary, which Congress and the President have determined since 1789 in a series of Judiciary Acts. Congress and the President have expanded the federal judiciary when: (1) the branches of government are under unified control of a single party; (2) unified control arose after a period of control by the now "out" party or after a prolonged period of divided government; and (3) the Supreme Court's policy preferences are out of alignment with the preferences of the new governing party. Under these circumstances, Congress and the President create new judicial slots to make partisan appointments to the lower bench. These appointments change the political orientation of the lower federal bench. This political change, in turn, forces the Supreme Court to adjust its doctrine in favor of elected officials.

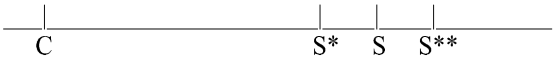


Figure 7.6. Supreme court doctrine.

The following model illustrates how a change in the composition of the lower courts alters judicial doctrine, and thereby is a means by which Congress and the President can influence the Supreme Court without changing its membership. For simplicity, assume that feasible judicial decisions can be arrayed on one dimension, that the Supreme Court and every lower court each has an ideal decision, and that each court prefers decisions closer to its ideal to those further away. The Supreme Court's doctrine is represented as the set of decisions around the Supreme Court's ideal policy that will not lead to a reversal. This circumstance is depicted in Figure 7.6, where C is the ideal Position of the lower court, S is the ideal point of the Supreme Court, and [S*, S**] is the Supreme Court's doctrine, or the interval of decisions that will not be reversed. Note that if C were inside the interval [S*, S**], the lower court faces no dilemma: the ideal decision can not be successfully appealed. But for the preference configuration shown, the lower court must consider the likelihood of successful appeal in picking a decision.

If the Supreme Court can not review all decisions by the lower courts, decisions face some probability $p < 1$ that they will be reviewed. For simplicity, assume that at the time of appeal the Supreme Court does not know the position of decision on the continuum—it must hear the case to figure out whether it complies. If there are N decisions by lower courts and the capacity of the Supreme Court to hear cases is K, then the probability a case will be reviewed is K/N.[28] In this case, if a lower court picks its own ideal point, it will have its decision reset to S with probability K/N but will obtain its most preferred outcome C with probability K/N. Or, the lower court can pick S* and have no fear of reversal. If the lower court maximizes expected value, it will pick its ideal decision if $V(S^*) < (K/N)V(S) + [1 - (K/N)]V(C)$, where $V(^*)$ is the value the lower court places on each outcome. In this setting, if the Supreme Court's doctrine is repugnant to the lower court, it will pick C and risk reversal, whereas if the lower court does not see much of a difference in the values of C and S*, it will comply by picking S*.

In this setting, the Supreme Court chooses doctrine strategically as a means of influencing lower courts. The optimal strategy for the Supreme Court is to set S* and S** so as to minimize the average distance between its ideal point and the decisions of the lower courts. As the size of the interval [S*, S**] grows larger, lower courts have less to gain by picking their ideal points rather than either S* or S**, whichever is nearer to C. In the example above, as S* approaches C, V(S*) increases, while the value of defying the Supreme Court remains the same. Hence, the lower court is more likely to pick S* rather than C, so that a wider set of acceptable decisions induces more compliance.

[28] This assumption is clearly unrealistic, but it also is not necessary for the general results to follow. We use it because it vastly increases the transparency of the model.

Now consider the effect if Congress and the President expand the lower courts and appoint new judges of a different ideology than Supreme Court, causing more lower court judges to be threats to defy the Supreme Court. In response to expansion of the lower courts, the Supreme Court will expand its doctrine. Since the Supreme Court cannot review all lower court decisions, it has an incentive to expand the set of acceptable decisions so that some lower courts that would otherwise defy it now choose to comply. The Supreme Court's doctrinal expansion favors the preferences of elected officials who were responsible for appointing the defiant lower court judges.

In this model, doctrine is a function of both normative principles and strategic aspects of the judicial hierarchy, namely, the need of the Supreme Court to police the lower courts. Elected officials can take advantage of the logic of that system to alter the Supreme Court's doctrine. McNollgast (1995, 2006) show that this is most likely to occur under the conditions noted above: when a new party takes united control of the government after a period in opposition or of divided government, and when the ideology of a new government is at variance with the Supreme Court's doctrine. Historically, these situations correspond to the largest expansions of the lower courts: under Abraham Lincoln, Franklin Roosevelt, and Ronald Reagan.

7.3. Interpreting statutes in a system of separated and shared powers

The positive and normative debates on statutory interpretation also parallel each other. Again, at the heart of these debates is disagreement about the role of Congress in American democracy and about the efficacy of Congress and the Presidency as representative institutions.

On one side of the debate are the intentionalists, who argue that courts should, and in fact do, follow legislative intent in their decisions. For some scholars and jurists, intent is defined solely by the plain language of the text (Easterbrook, 1984). Others argue that language is often ambiguous, especially when it comes to applying general statutory rules to the facts of a particular case, and thus jurists and bureaucrats must look beyond the text to discover its meaning (Posner, 1990; Eskridge, 1994).

On the other side of the debate are non-intentionalists: scholars who are not interested in the intent of the authors of a statute. These scholars believe that Congress is not representative, including scholars in Critical Legal Studies, Public Choice, and Political Science *cum* Realist schools discussed earlier, as well as those who argue, somewhat nihilistically, that collective intent is an impossible standard (see, e.g., Riker and Weingast, 1988; Shepsle, 1992). This work argues that courts should and/or do ignore the text of statutes and other legislative signals in favor of other commands.

The critics are correct in arguing that *only* if Congress is a representative body, and *only* if collective intent is a useful concept do courts have an obligation to follow the sovereign commands of the legislature. That is, if Congress is corrupt, captured by interest groups, or otherwise seriously unrepresentative of all citizens, if majorities within Congress act without regard to the will of minorities, or if legislative actions are a ran-

dom result from chaos and agenda manipulation so that collective intent is an impossible anthropomorphism, then jurists have no good reason not to ignore statutes.

As previous sections of this essay argue, PPT provides reason to believe that Congress *is* representative, that legislative intent *is* a meaningful concept, and, further, that legislative intent *is* discoverable (McNollgast 1992, 1994; Rodriguez and Weingast, 2003). We make this intentionalist argument on the basis of modern research that rejects the most extreme views about the failure of democratically elected government as a representative institution.[29] Congress chooses collectively between relatively clear alternatives, and thus the intent of those voting can be discerned, if viewed through the proper lens. Through deliberation, members of Congress and the President reach an agreement about the intent of legislative language, such that it is not a fool's errand to discover intent. Understanding the legislative process provides us with a set of criteria by which to judge which statements and documents are credible with respect to revealing collectively agreed upon intentions, and which are likely to be strategic or merely political grandstanding.[30] While this approach may not always yield unique interpretation, it yields fewer mistakes than other approaches to interpretation.

PPT of statutory interpretation begins by considering the incentives of legislatures when building a record for agencies and courts to consider when deciding the meaning of a statute (see Eskridge and Ferejohn 1992a, 1992b; McNollgast 1992, 1994; and Rodriguez and Weingast, 2003). PPT begins with the same assumptions that are used by those who argue that the legislative process is chaotic, namely that legislators have divergent preferences and that all legislators seek to advance their own interpretation of the statue, rather than the compromise that arises from the legislative bargain. In the course of consideration of a controversial bill, through its many committee versions and through the floor amendment process, in some circumstances legislators are likely to reveal where their preferences lie in the policy space at issue in the bill. Some are ardent supporters or opponents, while others are pivotal, i.e. those with centrist positions who actually determine whether the bill passes.

The preferences of legislators are derived from those of their supporting constituents, so that it is natural that legislators will want to make statements that show constituents that they are being faithfully and energetically represented. This, ardent supporters have an incentive to make statements for the benefit of their constituents that imply a more expansive version of the statutes than was actually adopted. But ardent supporters face a conflicting incentive: supporters want the bill to pass, and so will seek to make statements that convince less ardent, pivotal legislators to vote for the bill. Likewise, opponents of the statute have an incentive to please their constituents by claiming that the

[29] This literature is large and rapidly expanding in recent years, and includes Fenno (1978), Brady (1988), Jacobson (1990), Kiewiet and McCubbins (1991), Rohde (1991), Snyder (1992), Cox and McCubbins (1993), Aldrich (1995), Sinclair (1995), and Lupia and McCubbins (1998).

[30] McNollgast (1994) and Rodriguez and Weingast (2003) explore the implications and importance of credibility for statutory interpretation. Generally, see Lupia and McCubbins (1994, 1998) for more on the topic of credibility.

proposed bill is a catastrophe, but also to convince pivotal members that the statute does not really move existing policy in hopes that others will interpret the statute as narrow and limited. As a result, the legislative history is likely to contain conflicting statements by the same legislators, depending on which incentive is motivating their behavior.

To make sense of inconsistent statements, one must take account of the context of statements in the legislative record. Statements that represent joint agreements by all supporters, such as committee reports when committees include both ardent and pivotal supporters, or statements as part of a colloquy between ardent and pivotal supporters, reflect communications of mutual agreement among the coalition that enacted the statute. By contrast, statements such as speeches outside the context of negotiating the content of the bill, such as personal memoirs or *ex post* statements to "revise and extend the record," have no role in forging the bargain that gave rise to the statute, and so have no credibility as indicators of legislative intent. Indeed, legislators share a desire to have an opportunity to play to the home constituency by making statements that reflect the preferences of constituents. Likewise, statements by opponents, whether aimed at constituents or future interpreters, have no credibility because opponents are not part of the coalition that enacted the statute. The only credible statements by opponents are those than are made in the context of convincing pivotal members to vote against the statute.

The preceding discussion PPT offers simultaneously an explanation for why the legislative history is conflicting and yet a useful guide for determining which statements are useful indicators of the agreement among supporters of a statute. Whereas the legislative history is rarely so complete that all ambiguities in a statute can be clearly resolved, PPT does support the value of some "canons" of statutory interpretation that have broad validity. An obvious canon is that any interpretation that is more consistent with language that was rejected anywhere in the process—in committee or on the floor—does not reflect legislative intent. Another obvious canon is that floor leaders of a bill, when discussing the interpretation of the statute with members who are pivotal, are the most likely source of accurate statements of intent because their statements are made in their role as a representative of the entire enacting coalition, not as an individual member.

8. PPT of law: concluding observations

One of PPT's principal objectives is to broaden the study of law to include elections, elected officials and the bureaucracy as well as the courts in the system of making law. Congress, the president, and the bureaucracy all produce law directly, and these branches as well as citizens indirectly influence law-making by the courts because of the interactions and interdependencies among them. Put simply, studying judicial sources of law is too restrictive to provide a complete understanding of law.

An essential feature of PPT of Law is the contention that law *is* structure and process. That is, in writing and passing statutes Congress and the President state (often vaguely) not just the aims of a law, but also the structure and process that determine how decisions will be made to embellish and implement those aims, including with some precision

when, how and on what grounds the courts can play a role in this process. When elected officials pass statutes, they establish the institutions, procedures and rules by which policy will be made by agencies and courts. The policy itself is not law, but the product of the law. Structure and process direct policy toward some outcomes and away from others, and these outcomes entail winners and losers. Law is the set of instructions to bureaucrats and judges about what they should do and how they should do it. Policy emerges from the strategic choices by all of the relevant actors, as mediated and channeled by law in the form of structure and process.

PPT of Law is further distinguished by the fact that it assumes that the purpose of law—i.e., the purpose of a structure and process for making policy—is political. That is, law is designed to advance the political agenda of a winning political coalition. By designing the structure and process of policy-making, political actors allocate rights, determine the relative importance of different costs and benefits, and ultimately affect the level and distribution of wealth in society. The key normative assumption within this framework is that the choice of structure and process, like the choice of substantive purpose, is governed by the democratic procedures proscribed by the Constitution, to which the citizens give their consent.

How we view the democratic sources of law—those from Congress, the President, and the bureaucracy, as opposed to the courts—depends on how effective democratic institutions are at representing citizen interests. This survey reviewed the literature on each element in the chain from citizens to Congress and the President, to the bureaucracy, to the courts. Although public failures, legislative pathologies, and interest group capture are all a source of problems in democratic system, these elements are not the only factors influencing democratic law-making. PPT of law provides a coherent framework for understanding how each component of the government operates, and how each shapes the behavior of the others.

As a positive theory that focuses on how institutions affect behavior by shaping incentives, PPT provides understanding about the relationship between democratic governance institutions and law. Regardless of the relative weights one places on various normative principles, whether democratic legitimacy, economic efficiency, individual liberty or distributive justice, a necessary first step is to connect actions to outcomes. In this sense PPT is of value to all sides of ideological disputes. Beyond this, PPT offers comfort to those who believe that government actions must have democratic legitimacy to be normatively compelling. PPT focuses on the properties of democratic institutions, shows theoretically that policies are weakly responsive to citizen preferences and empirically that these theoretical predictions are supported by the data.

References

Aberbach, J. (1990). Keeping a Watchful Eye: The Politics of Congressional Oversight. Brookings Institution Press, Washington, D.C.

Abramson, J. (1994). We, the Jury: The Jury System and the Ideal of Democracy. Basic Books, New York.

Ackerman, B.A., Hassler, W.T. (1981). Clean Coal/Dirty Air: Or How the Clean Air Act Became a Multibillion Dollar Bail-out for High Sulfur Coal Prodicers and What Should Be Done About It. Yale University Press, New Haven.

Adamany, D.W. (1973). "Legitimacy, realigning elections, and the supreme court". Wisconsin Law Review 3, 790–846.

Adamany, D.W. (1980). "The supreme court's role in critical elections". In: Campbell, B., Trilling, R. (Eds.), Realignment in American Politics. University of Texas Press, Austin.

Aldrich, J. (1995). Why Parties? The Origin and Transformation of Political Parties in America. The University of Chicago Press, Chicago.

Aldrich, J., Rohde, D. (1998). "Measuring conditional party government". Paper presented at the Annual Meeting of the Midwest Political Science Association, April 23–25, Chicago, IL.

Aldrich, J., Rohde, D. (2001). "The logic of conditional party government". In: Dodd, L.C., Oppenheimer, B.I. (Eds.), Congress Reconsidered, 7th Edition. Congressional Quarterly, Washington, D.C.

Aldrich, J., Cox, G., McCubbins, M., Rohde, D. (in progress). "Delegation and party homogeneity". Working Paper.

American Political Science Association (1950). Toward a More Responsible Two-Party System: A Report of the Committee on Political Parties. American Political Science Association, Washington, D.C.

Aranson, P., Gellhorn, E., Robinson, G. (1982). "A theory of legislative delegation". Cornell Law Review 68, 1–67.

Arnold, R.D. (1979). Congress and the Bureaucracy: A Theory of Influence. Yale University Press, New Haven.

Arnold, R.D. (1990). The Logic of Congressional Action. Yale University Press, New Haven.

Arrow, K.J. (1951). Social Choice and Individual Values. Yale University Press, New Haven.

Banks, J. (1991). "The Space Shuttle". In: Cohen, L.R., Noll, R.G. (Eds.), The Technology Pork. Brookings Institution, Washington.

Barnum, D. (1985). "The supreme court and public opinion: judicial decision-making in the post-new deal period". Journal of Politics 47, 652–666.

Baron, D.P. (1990). "Distributive politics and the persistence of amtrak". The Journal of Politics 52, 883–913.

Baum, L. (1980). "Responses of federal district judges to court of appeals policies: an exploration". Western Political Quarterly 33, 217–224.

Baum, L. (1988). "Measuring policy change in the U.S. supreme court". American Political Science Review 82, 905–912.

Bawn, K. (1995). "Political control versus expertise: congressional choices about administrative procedures". American Political Science Review 89, 62–73.

Bawn, K. (1997). "Choosing strategies to control the bureaucracy: statutory constraints, oversight, and the committee system". Journal of Law, Economics and Organization 13, 101–126.

Becker, G. (1983). "A theory of competition among pressure groups for political influence". Quarterly Journal of Economics 98, 371–400.

Bentley, A.F. (1908). The Process of Government: A Study of Social Pressures. The University of Chicago Press, Chicago.

Bentley, A.F. (1949). The Process of Government. Principia Press, Bloomington.

Berelson, B.R., Lazarsfeld, P.F., McPhee, W.N. (1954). Voting. The University of Chicago Press, Chicago.

Berglof, E., Rosenthal, H. (2004). "Congress and the history of bankruptcy law in the United States: from the federalists to the whigs". Paper presented at the History of Congress Conference, Stanford California, April 9–10.

Besley, T., Burgess, R. (2002). "The political economy of government responsiveness: theory and evidence from India". Quarterly Journal of Economics 117, 1415–1452.

Besley, T., Coate, S. (1997). "An economic model of representative democracy". Quarterly Journal of Economics 103, 903–937.

Besley, T., Coate, S. (1998). "Sources of inefficiency in representative democracy: a dynamic analysis". American Economic Review 88, 139–156.

Besley, T., Pande, R., Rahman, L., Rao, V. (2004). "The politics of public goods provision: evidence from Indian local elections". Journal of the European Economic Association 2 (2-3), 416–426.
Bickel, A.M. (1962). The Least Dangerous Branch: The Supreme Court at the Bar of Politics. Bobbs-Merrill, Indianapolis.
Bickers, K., Stein, R. (1994a). "Universalism and the electoral connection: a test and some doubts". Political Research Quarterly 47, 295–317.
Bickers, K., Stein, R. (1994b). "Response to Barry Weingast's reflections". Political Research Quarterly 47, 329–333.
Bickers, K., Stein, R. (1996). "The electoral dynamics of the federal pork barrel". American Journal of Political Science 40, 1300–1326.
Binder, S.A. (1997). Minority Rights, Majority Rule: Partisanship and the Development of Congress. Cambridge University Press, New York.
Binder, S.A. (1999). "The dynamics of legislative gridlock, 1947–1996". American Political Science Review 93 (September), 519–533.
Binder, S.A., Smith, S.S. (1997). Politics or Principle? Filibustering in the United States Senate. Brookings Institution, Washington, D.C.
Black, D. (1958). The Theory of Committee and Elections. Cambridge University Press, Cambridge.
Blackstone, W. (1765–1769). Commentaries on the Laws of England. Clarendon Press, Oxford.
Bond, J.R., Fleisher, R. (1990). The President in the Legislative Arena. University of Chicago Press, Chicago.
Brady, D. (1988). Critical Elections and Congressional Policy-making. Stanford University Press, Stanford.
Brady, D., Volden, C. (1998). Revolving Gridlock: Politics and Policy from Carter to Clinton. Westview Press, Boulder.
Bresnick, D. (1979). "The federal educational policy system: enacting and revising Title I". The Western Political Quarterly 32, 189–202.
Breyer, S. (1982). Regulation and Its Reform. Harvard University Press, Cambridge.
Breyer, S. (1993). Breaking the Vicious Circle: Toward Effective Risk Regulation. Harvard University Press, Cambridge.
Browning, R.X. (1986). Politics and Social Welfare Policy in the United States. University of Tennessee Press, Knoxville.
Buchanan, J. (1968). The Demand and Supply of Public Goods. Rand McNally, Chicago.
Buchanan, J.M., Tullock, G. (1962). The Calculus of Consent. University of Michigan Press, Ann Arbor.
Buchanan, J.M., Tollison, R., Tullock, G. (1980). Toward a Theory of the Rent Seeking Society. Texas A & M University, College Station.
Caldeira, G.A. (1987). "Public opinion and the U.S. supreme court: FDR's court packing plan". American Political Science Review 81, 1139–1153.
Caldeira, G.A. (1991). "Courts and public opinion". In: Gates, J.B., Johnson, C.A. (Eds.), The American Courts: A Critical Assessment. CQ Press, Washington, D.C.
Caldeira, G.A., Wright, J. (1988). "Organized interests and agenda-setting in the U.S. supreme court". American Political Science Review 82, 1109–1127.
Calvert, R.L., Moran, M.J., Weingast, B.R. (1987). "Congressional Influence over policymaking: the case of the FTC". In: McCubbins, M.D., Sullivan, T. (Eds.), Congress: Structure and Policy. Cambridge University Press.
Cameron, C.M. (1994). "New avenues for modeling judicial politics". Paper presented at the Conference on the Political Economy of Public Law, Rochester, NY.
Cameron, C.M. (2000). Veto Bargaining: Presidents and the Politics of Negative Power. Cambridge University Press, Cambridge.
Cameron, C.M., Elmes, S. (1995). "A theory of sequential veto bargaining". Working Paper. Departments of Political Science and Economics, Columbia University.
Canes-Wrone, B. (2001). "The president's legislative influence from public appeals". American Journal of Political Science 45, 313–329.
Cardozo, B.N. (1922). The Nature of the Judicial Process. Yale University Press, New Haven.

Carp, R.A., Rowland, C.K. (1983). Policymaking and Politics in the Federal District Courts. University of Tennessee Press, Knoxville.
Carter, S. (1988). "The confirmation mess". Harvard Law Review 101, 1185–1201.
Chipman, J.S., Moore, J.C. (1976). "The scope of consumer's surplus arguments". In: Tang, A.M., Westfield, F.M., Worley, J.S. (Eds.), Evolution, Welfare and Time in Economics. D.C. Health, Lexington, Mass., pp. 69–123.
Cohen, A.S., Taylor, E. (2000). American Pharaoh: Mayor Richard J. Daley: His Battle for Chicago and the Nation. Little, Brown, Boston.
Cohen, L.R. (1979). "Innovation and atomic energy: nuclear power regulation, 1966 present". Law and Contemporary Problems 43, 67–97.
Cohen, L., Matthews, S.A. (1980). "Constrained plott equilibrium, directional equilibrium and global cycling sets". Review of Economic Studies 47 (5), 975–986. No. 150; #166.
Cohen, L., Noll, R. (1991). The Technology Pork Barrel. Brookings Institution, Washington, D.C.
Cohen, L., Spitzer, M. (1994). "Solving the *Chevron* puzzle". Law and Contemporary Problems 57, 65–110.
Cohen, L., Spitzer, M. (1996). "Judicial deference to agency action: a rational choice theory and an empirical test". Southern California Law Review 69, 431–476.
Cohen, J.E., Krassa, M.A., Hamman, J.A. (1991). "The impact of presidential campaigning on midterm U.S. senate elections". American Political Science Review 85, 165–178.
Coke, E. (1608). Prohibitions del Roy, 12 Co. Rep. 63, 65, 77 Eng. Rep. 342.
Condorcet, M. (1785). "An essay on the application of probability theory to plurality decision making: an election between three candidates". In: Sommerlad, F., McLean, I. (Eds.), The Political Theory of Condorcet. Oxford University Press, Oxford, UK, pp. 69–80. (1989).
Congressional Quarterly (1984). Farm Policy: The Politics of Soil, Surpluses and Subsidies. Congressional Quarterly, Washington, D.C.
Cook, B.B. (1977). "Public opinion and federal judicial policy". American Journal of Political Science 21, 567–600.
Cooper, J., Brady, D.W. (1981). "Institutional context and leadership style: the house from cannon to rayburn". American Political Science Review 75, 411–425.
Cooter, R., Ulen, T. (1988). Law and Economics. Scott, Foresman, Glenview.
Cornell, N., Noll, R., Weingast, B. (1976). "Safety regulation". In: Owen, H., Schultze, C. (Eds.), Setting National Priorities The Next Ten Years. Brookings Institution, Washington.
Covington, C.R. (1987). "Mobilizing congressional support for the president: insights from the 1960's". Legislative Studies Quarterly 12, 77–95.
Cox, G.W. (1987). The Efficient Secret: The Cabinet and the Development of Political Parties in Victorian England. Cambridge University Press, Cambridge.
Cox, G.W. (1997). Making Votes Count: Strategic Coordination in the World's Electoral Systems. Cambridge University Press, Cambridge.
Cox, G.W. (1999). "The empirical content of rational choice theory". Journal of Theoretical Politics 11 (2), 147–169.
Cox, G.W., McCubbins, M.D. (1991). "Divided control of fiscal policy". In: Cox, G.W., Kernell, S. (Eds.), The Politics of Divided Government. Westview Press, Boulder, CO.
Cox, G.W., McCubbins, M.D. (1993). Legislative Leviathan: Party Government in the House. University of California Press, Berkeley.
Cox, G.W., McCubbins, M.D. (1994). "Bonding, structure, and the stability of political parties: party government in the house". Legislative Studies Quarterly 19, 215–231.
Cox, G.W., McCubbins, M.D. (2000). "The institutional determinants of economic policy outcomes". In: Haggard, S., McCubbins, M.D. (Eds.), Presidents Parliaments and Policy. Cambridge University Press, Cambridge.
Cox, G.W., McCubbins, M.D. (2001). "The institutional determinants of economic policy outcomes". In: McCubbins, M.D., Haggard, S. (Eds.), Presidents, Parliaments and Policy. Cambridge University Press, Cambridge.

Cox, G.W., McCubbins, M.D. (2002). "Agenda power in the U.S. house of representatives, 1877 to 1986". In: Brady, D., McCubbins, M.D. (Eds.), Party, Process, and Political Change in Congress: New Perspectives on the History of Congress. Stanford University Press, Palo Alto.

Cox, G.W., McCubbins, M.D. (2005). Setting the Agenda: Responsible Party Government in the U.S. House of Representatives.

Cox, G.W., Rosenbluth, F. (1996). "Factional competition for the party endorsement: the case of Japan's Liberal Democratic Party". British Journal of Political Science 259–269.

CQ Farm Policy (1984) Farm Policy: The Politics of Soil, Surpluses, and Subsides. By Nancy A. Blanpied. Congressional Quartely Inc., Washington, D.C.

Craswell, R., Schwartz, A. (1994). Foundations of Contract Law. Oxford University Press, New York.

Critical Review (1995). Refers to the entire issue of Critical Review, vol. 9 of that year.

Cronin, T.E. (1980). "A resurgent congress and the imperial presidency". Political Science Quarterly 95, 209–237.

Dahl, R. (1956). A Preface to Democratic Theory. The University of Chicago Press, Chicago.

Dahl, R. (1957). "Decision-making in a democracy: the supreme court as a national policy maker". Journal of Public Law 6, 279–295.

Dahl, R. (1967). Pluralist Democracy in the United States: Conflict and Consent. Rand McNally, Chicago.

Davidson, R.H., Oleszek, W.J. (2000). Congress and Its Members, 7th edn. CQ Press, Washington, D.C.

Davis, D.H. (1982). Energy Politics. St. Martin's Press, New York.

Davis, O.A., Hinich, M.J., Ordeshook, P.C. (1970). "An expository development of a mathematical model of the electoral process". American Political Science Review 64 (2), 426–448. #1095.

Deering, C., Maltzman, F. (1999). "The politics of executive orders: legislative constraints on presidential power". Political Research Quarterly 52, 767–783.

Den Hartog, C.F., Monroe, N.W. (2004). "The value of majority status: the effect of Jeffords's switch on asset prices of republican and democratic firms". Paper presented at the Annual Meeting of the Midwest Political Science Association, April 15–18, Chicago, IL.

Dion, D., Huber, J. (1996). "Procedural choice and the house committee on rules". Journal of Politics 58, 25–53.

Dodd, L., Schott, R. (1979). Congress and the Administrative State. Wiley, New York.

Downs, A. (1957). An Economic Theory of Democracy. Harper, New York.

Duverger, M. (1954). Political Parties. Wiley, New York.

Easterbrook, F.H. (1984). "Legal interpretation and the power of the judiciary". Harvard Journal of Law & Public Policy 7, 87.

Edwards, G.C. III (1980). Presidential Influence in Congress. W.H. Freeman, San Francisco.

Edwards, G.C. III (1983). The Public Presidency. St. Martin's Press, New York.

Edwards, G.C. III (1989). At the Margins: Presidential Leadership of Congress. Yale University Press, New Haven.

Edwards, G.C. III, Barrett, A., Peake, J. (1997). "The legislative impact of divided government". American Journal of Political Science 41, 545–563.

Epstein, D. (1998). "Partisan and bipartisan signaling in Congress". Journal of Law, Economics, and Organization 14 (2), 183–204.

Epstein, D., O'Halloran, S. (1996). "A theory of strategic oversight: congress, lobbyists, and the bureaucracy". Journal of Law, Economics and Organization 11, 227–255.

Epstein, D., O'Halloran, S. (1999). Delegating Powers: A Transaction Cost Politics Approach to Policy-making Under Separate Powers. Cambridge University Press, Cambridge.

Epstein, L. (1985). Conservatives in Court. University of Tennessee Press, Knoxville.

Epstein, L., Knight, J. (1996). "On the struggle for judicial supremacy". Law and Society Review 30, 87–120.

Epstein, L., Knight, J. (1998). The Choices Justices Make. CQ Press, Washington, D.C.

Epstein, L., Kobylka, J.F. (1992). The Supreme Court and Legal Change: Abortion and the Death Penalty. University of North Carolina Press, Chapel Hill.

Epstein, L., Walker, T.G. (1995). "The role of the supreme court in American Society: playing the reconstruction game". In: Epstein, L. (Ed.), Contemplating Courts. CQ Press, Washington, D.C.

Epstein, L., Walker, T.G., Dixon, W.J. (1989). "A neo-institutional perspective of supreme court decision-making". American Journal of Political Science 33, 825–841.
Eskridge, W. (1991). "Overriding supreme court statutory interpretation decisions". Yale Law Journal 101, 331–424.
Eskridge, W. (1994). Dynamic Statutory Interpretation. Harvard University Press, Cambridge.
Eskridge, W., Ferejohn, J. (1992a). "Making the deal stick: enforcing the original constitutional structure of lawmaking in the modern regulatory state". Journal of Law, Economics, and Organization 8, 165–189.
Eskridge, W., Ferejohn, J. (1992b). "The Article I, Section 7 Game". Georgetown Law Journal 80, 523–564.
Eskridge, W., Frickey, P. (1990). "Statutory interpretation as practical reasoning". Stanford Law Review 42, 321–384.
Eskridge, W., Frickey, P. (1994). "Editors' introduction". In: Hart, H.M. Jr., Sacks, A.M. (Eds.), The Legal Process: Basic Problems in the Making and Application of Law. Foundation Press, New York, NY.
Farber, D., Frickey, P. (1991). Law and Public Choice. The University of Chicago Press, Chicago.
Fenno, R. (1973). Congressmen in Committees. Little, Brown, and Company, Boston.
Fenno, R. (1978). Home Style: House Members in their Districts. Little, Brown, and Company, Boston.
Ferejohn, J. (1974). Pork Barrel Politics: Rivers and Harbors Legislation, 1947–1968. Stanford University Press, Stanford.
Ferejohn, J. (1987). "The structure of agency decision processes". In: McCubbins, M.D., Sullivan, T. (Eds.), Congress: Structure and Policy. Cambridge University Press, Cambridge.
Ferejohn, J., Satz, D. (1994). "Rational choice and social theory". Journal of Philosophy 91, 71–87.
Ferejohn, J., Shipan, C. (1989). "Congressional influence on administrative agencies: a case study of telecommunications policy". In: Dodd, L., Oppenheimer, B. (Eds.), Congress Reconsidered, 4th edn. Congressional Quarterly Press, Washington, D.C.
Ferejohn, J., Shipan, C. (1990). "Congressional influence on bureaucracy". Journal of Law, Economics and Organization 6.
Ferejohn, J., Spiller, P. (1992). "The economics and politics of administrative law and procedures: an introduction". Journal of Law, Economics, and Organization 8 (1).
Ferejohn, J., Weingast, B.R. (1992a). "Limitations of statutes: strategic statutory interpretation". Georgetown Law Review 80, 565–582.
Ferejohn, J., Weingast, B.R. (1992b). "A positive theory of statutory interpretation". International Review of Law and Economics 12, 263–279.
Fiorina, M.P. (1974). Representatives, Roll Calls, and Constituencies. Lexington Books, Lexington, MA.
Fiorina, M.P. (1977a). Congress: Keystone of the Washington Establishment. Yale University Press, New Haven.
Fiorina, M.P. (1977b). "The case of the vanishing marginals: the bureaucracy did it". American Political Science Review 71, 177–181.
Fiorina, M.P. (1979). "Control of the bureaucracy: a mismatch of incentives and capabilities". In: Livingston, W., Dodd, L., Schott, R. (Eds.), The Presidency and the Congress: A Shifting Balance of Powers? The University of Texas Press, Austin.
Fiorina, M.P. (1981a). "Some problems in studying the effects of resource allocation in congressional elections". American Journal of Political Science 25, 543–567.
Fiorina, M.P. (1981b). Retrospective Voting in American National Elections. Yale University Press, New Haven.
Fiorina, M.P. (1982). "Legislative choice of regulatory forms: legal process or administrative process". Public Choice 39, 33–66.
Fiorina, M.P., Noll, R.G. (1978). "Voters, bureaucrats, and legislators: a rational choice perspective on the growth of bureaucracy". Journal of Public Economics 9, 239–254.
Fisher, L. (1975). Presidential Spending Power. Princeton University Press, Princeton.
Fisher, L. (1981). The Politics of Shared Power: Congress and the Executive. CQ Press, Washington, D.C.
Fox, D.M. (1971). "Congress and U.S. military service budgets in the post-war period: a research note". Midwest Journal of Political Science 15, 382–393.

Fox, D.M., Clapp, C. (1970). "The house rules committee and the programs of Kennedy and Johnson administrations". Midwest Journal of Political Science 14, 667–672.
Fuller, L. (1964). The Morality of Law. Yale University Press, New Haven.
Funston, R. (1975). "The supreme court and critical elections". American Political Science Review 69, 795–811.
Galanter, M. (1974). "Why the 'haves' come out ahead: speculation on the limits of legal change". Law and Society Review 9, 95–160.
Gates, J.B. (1987). "Partisan realignment, unconstitutional state policies, and the U.S. supreme court: 1837–1964". American Journal of Political Science 31, 259–280.
Gates, J.B. (1992). The Supreme Court and Partisan Realignment: A Macro- and Micro Level Analysis. Westview Press, Boulder.
Gely, R., Spiller, P.T. (1990). "A rational choice theory of supreme court statutory decisions, with applications to the state farm and grove city cases". Journal of Law, Economics, and Organization 6, 263–300.
Gely, R., Spiller, P.T. (1992). "The political economy of supreme court constitutional decisions: the case of Roosevelt's court packing plan". International Review of Law and Economics 12, 45–67.
George, T.E., Epstein, L. (1992). "On the nature of supreme court decision making". American Political Science Review 86, 323–337.
Giles, M.W., Walker, T.G. (1975). "Judicial policy-making and southern school segregation". Journal of Politics 37, 917–936.
Gilligan, T.W., Krehbiel, K.K. (1990). "Organization of informative committees by a rational legislature". American Journal of Political Science 34, 531–564.
Gilmour, J.B. (1995). Strategic Disagreement: Stalemate in American Politics. University of Pittsburgh Press, Pittsburgh.
Gilmour, J.B. (1999). "Objectionable legislation and the presidential propensity to veto". Presented at the 1999 Annual Meeting of the Western Political Science Association, Seattle, WA.
Glennon, M. (1983). "The senate role in treaty ratification". The American Journal of International Law 77, 257–280.
Gosnell, H.F. (1937). Machine Politics: Chicago Model. University of Chicago Press, Chicago.
Graham, R. (1990). Patronage and Politics in Nineteenth-Century Brazil. Stanford University Press, Palo Alto, CA.
Green, D., Shapiro, I. (1994). Pathologies of Rational Choice Theory: A Critique of Applications in Political Science. Yale University Press, New Haven.
Groseclose, T., McCarty, N. (2001). "The politics of blame: bargaining before an audience". American Journal of Political Science 45, 100–119.
Hamman, J.A. (1993). "Bureaucratic accommodation of congress and the president: elections and the distribution of federal assistance". Political Research Quarterly 46, 863–879.
Hammond, T.H., Miller, G.J. (1987). "The core of the constitution". The American Political Science Review 81, 1155–1174.
Hart, H.L.A. (1961). The Concept of Law. Clarendon Press, Oxford.
Hart, H.M., Sacks, A.M. (1958). The Legal Process: An Introduction to Decision-Making By Judicial, Legislative, Executive and Administrative Agencies. Cambridge.
Hart, H., Sacks, A. (1994). The Legal Process: Basic Problems in the Making and Application of Law, Revised edition of 1958 tentative edition. Foundation Press, New York, NY.
Heclo, H. (1975). "The office of management and budget and the presidency: the problem of neutral competence". The Public Interest 38, 80–98.
Heclo, H. (1977). A Government of Strangers: Executive Politics in Washington. Brookings Institution, Washington, D.C.
Heclo, H. (1984). "Executive budget making". In: Mills, G.B., Palmer, J.L. (Eds.), Federal Budget Policy in the 1980s. The Urban Institute, Washington, D.C.
Helbich, W. (1967). "American liberals in the league of nations controversy". The Public Opinion Quarterly 31, 568–596.

Hobbes, T. (1651). Leviathan, pt. II. Printed for Andrew Crooke, London.
Hobbes, T. (1971). In: Cropsey (Ed.), A Dialogue between a Philosopher and a Student of the Common Laws of England. The University of Chicago Press, Chicago.
Holmes, O.W. (1881). The Common Law. Little, Brown, and Company, Boston.
Holmes, O.W. (1897). "The path of law". Harvard Law Review 10, 457–478.
Horwitz, M. (1992). The Transformation of American Law, 1870–1960: The Crisis of Legal Orthodoxy. Oxford University Press, New York.
Howell, W. (2003). Power without Persuasion: The Politics of Direct Presidential Action. Princeton University Press, Princeton.
Ingberman, D., Yao, D. (1991). "Presidential commitment and the veto". American Journal of Political Science 35, 357–389.
INS v. Chadha (1983) 462 U.S. 919.
Jacobson, G.C. (1990). "The effects of campaign spending in house elections: new evidence for old arguments". American Journal of Political Science 34, 334–362.
Jones, C.O. (1968). "Joseph G. Cannon and Howard W. Smith: an essay on the limits of leadership in the house of representatives". Journal of Politics 30, 617–646.
Kanter, A. (1972). "Congress and the defense budget: 1960–1970". The American Political Science Review 66, 129–143.
Kernell, S. (1986). "The early nationalization of political news in America". American Political Development 1, 255–278.
Kernell, S. (1991). "Facing an opposition congress". In: Cox, G., Kernell, S. (Eds.), The Politics of Divided Government. Westview Press, Boulder, CO.
Kernell, S., McDonald, M.P. (1999). "Congress and America's political development: the transformation of the post office from patronage to service". American Journal of Political Science 43, 792–811.
Kiewiet, D.R., McCubbins, M.D. (1988). "Presidential influence on congressional appropriations decisions". American Journal of Political Science 32, 713–736.
Kiewiet, D.R., McCubbins, M.D. (1991). The Logic of Delegation: Congressional Parties and the Appropriations Process. The University of Chicago Press, Chicago.
Kirst, M.W. (1969). Government Without Passing Laws: Congress' Nonstatutory Techniques for Appropriations Control. University of North Carolina Press, Chapel Hill.
Knight, J., Epstein, L. (1996). "The norm of stare decisis". American Journal of Political Science 40, 1018–1035.
Kobylka, J.F. (1991). The Politics of Obscenity: Group Litigation in a Time of Legal Change. Greenwood Press, New York.
Kobylka, J.F. (1995). "The mysterious case of establishment clause litigation: how organized litigants foiled legal change". In: Epstein, L. (Ed.), Contemplating Courts. CQ Press, Washington, D.C.
Kolko, G. (1965). "Railroads and Regulation, 1877–1916". Princeton University Press, Princeton.
Krehbiel, K. (1986). "Unanimous consent agreements: going along in the senate". Journal of Politics 48, 541–564.
Krehbiel, K. (1991). Information and Legislative Organization. University of Michigan Press, Ann Arbor.
Krehbiel, K. (1998). Pivotal Politics: A Theory of U.S. Lawmaking. University of Chicago Press, Chicago.
Krehbiel, K. (2000). "Party discipline and measures of partisanship". American Journal of Political Science 44, 212–227.
Krueger, A.O. (1974). "The political economy of the rent-seeking society". The American Economic Review 64, 291–303.
Kuklinski, J., Stanga, J.E. (1979). "Political participation and government responsiveness: the behavior of California superior courts". American Political Science Review 73, 1090–1099.
Landis, J. (1938). The Administrative Process. Yale University Press, New Haven.
Landis, W.M., Posner, R.A. (1975). "The independent judiciary in an interest group perspective". Journal of Law and Economics 18, 875–901.
Langdell, C.C. (1871). A Selection of Cases on the Law of Contracts: With References and Citations. Little, Brown, and Company, Boston.

Langdell, C.C. (1880). A Summary of the Law of Contracts, 2nd edn. Little, Brown, and Company, Boston.
Lapham, L.J. (1954). Party Leadership and the House Committee on Rules. Ph.D. Dissertation, Harvard University.
Laver, M., Shepsle, K.A. (1996). Making and Breaking Governments: Cabinets and Legislatures in Parliamentary Democracies. Cambridge University Press, Cambridge, New York.
Lee, F. (1998). "Representation and public policy: the consequences of senate apportionment for the geographic distribution of federal funds". Journal of Politics 60, 34–62.
Levy, B., Spiller, P.T. (1994). "The institutional foundations of regulatory commitment: a comparative analysis of telecommunications regulation". Journal of Law, Economics, and Organization 10 (2), 201–246.
Lijphart, A. (1977). Democracy in Plural Societies: A Comparative Explanation. Yale University Press, New Haven, CT.
Lijphart, A. (1996). "The puzzle of Indian democracy: a consociational interpretation". Am. Polit. Sci. Rev. 90, 258–268.
Lijphart, A. (1999). Patterns of Democracy: Government Forms and Performance in Thirty-Six Countries. Yale University Press, New Haven, CT.
Llewellyn, K. (1930). "A realistic jurisprudence—the next step". Columbia Law Review 30, 431–464.
Llewellyn, K. (1931). "Some realism about realism—responding to dean pound". Harvard Law Review 44, 1222–1264.
Llewellyn, K. (1960). The Common Law Tradition: Deciding Appeals. Little, Brown, and Company, Boston.
Lohmann, S., O'Halloran, S. (1994). "Divided government and U.S. trade policy: theory and evidence". International Organization 48, 595–632.
Lowi, T. (1969). The End of Liberalism: The Second Republic of the United States. Norton, New York.
Lupia, A., McCubbins, M.D. (1994). "Designing bureaucratic accountability". Law and Contemporary Problems 57, 91–126.
Lupia, A., McCubbins, M.D. (1998). The Democratic Dilemma: Can Citizens Learn What they Need to Know? Cambridge University Press, Cambridge.
Lupia, A., McCubbins, M.D., Popkin, S.L. (2000). Elements of Reason. Cambridge University Press, Cambridge.
MacAvoy, P.W. (1965). The Economic Effects of Regulation: The Trunk-Line Railroad Cartels and the Interstate Commerce Commission before 1900. MIT Press, Cambridge.
Macey, J.R. (1986). "Promoting public-regarding legislation through statutory interpretation: an interest group model". Columbia Law Review 86, 223–313.
Maltzman, F., Smith, S.S. (1994). "Principals, goals, dimensionality, and congressional committees". Legislative Studies Quarterly 19 (4), 457–476.
Marks, B. (1988). "A model of judicial influence on congressional policymaking: Grove City College v. Bell". Ph.D. Dissertation, Department of Economics, Washington University.
Marshall, T.R. (1989). Public Opinion and the Supreme Court. Unwin Hyman, London.
Mashaw, J. (1983). Bureaicratic Justice: Managing Social Security Disability Claims. Yale University Press, New Haven.
Mashaw, J. (1985a). Due Process in the Administrative State. Yale University Press, New Haven.
Mashaw, J. (1985b). "Prodelegation: why administrators should make political decisions". Journal of Law, Economics, and Organization 1, 81–100.
Mashaw, J. (1994). "Improving the environment of agency rulemaking: an essay on management, games, and accountability". Law and Contemporary Problems 57 (2), 185–257.
Mashaw, J. (1997). Greed, Chaos, and Governance. Yale University Press, New Haven.
Matthews, S.A. (1989). "Veto threats: rhetoric in a bargaining game". The Quarterly Journal of Economics 104, 347–369.
Mayer, K. (1999). "Executive orders and presidential power". Journal of Politics 61, 445–466.
Mayer, K. (2001). With the Stroke of a Pen: Executive Orders and Presidential Power. Princeton University Press, Princeton.
Mayhew, D. (1974). Congress: The Electoral Connection. Yale University Press, New Haven, CT.

Mayhew, D. (1991). Divided We Govern: Party Control, Lawmaking, and Investigations, 1946–1990. Yale University Press, New Haven.
McCarty, N. (1997). "Presidential reputation and the veto". Economics and Politics 9, 1–26.
McCarty, N. (2004). "The appointments dilemma". American Journal of Political Science 48 (3), 413–428.
McConnell, G. (1966). Private Power and American Democracy. Knopf, New York.
McCubbins, M.D. (1985). "Legislative design of regulatory structure". American Journal of Political Science 29, 721–748.
McCubbins, M.D. (1991). "Party politics, divided government, and budget deficits". In: Kernell, S. (Ed.), Parallel Politics: The Politics of Economic Policy in Japan and the United States. Brookings Institution, Washington, D.C.
McCubbins, M.D., Page, T. (1987). "A theory of congressional delegation". In: McCubbins, M.D., Sullivan, T. (Ed.), Congress: Structure and Policy. Cambridge University Press, Cambridge.
McCubbins, M., Rosenbluth, F. (1995). Structure & Policy in Japan and the U.S. Cambridge University Press, New York.
McCubbins, M.D., Schwartz, T. (1984). "Congressional oversight overlooked: police patrols and fire alarms". American Journal of Political Science 28, 165–179.
McCubbins, M.D., Schwartz, T. (1988). "Congress, the courts, and public policy: consequences of the one man, one vote rule". American Journal of Political Science 32, 388–415.
McCubbins, M.D., Noll, R.G., Weingast, B.R. (1987). "Administrative procedures as instruments of political control". Journal of Law, Economics and Organization 3, 243–277.
McCubbins, M.D., Noll, R.G., Weingast, B.R. (1989). "Structure and process, politics and policy: administrative arrangements and the political control of agencies". Virginia Law Review 75, 431–482.
McKelvey, R. (1979). "General conditions for global intransitives in formal voting models". Econometrica 7 (5), 1085–1112.
McKelvey, R. (1976). "Intransitivities in multidimensional voting models and some implications for agenda control". Journal of Economic Theory 12, 472–482.
McNollgast (1990). "Positive and normative models of due process: an integrative approach to administrative procedures". Journal of Law, Economics and Organization 6, 307–332.
McNollgast (1992). "Positive canons: the role of legislative bargains in statutory interpretations". Georgetown Law Journal 80, 705.
McNollgast (1994). "Legislative intent: the use of positive political theory in statutory interpretation". Law and Contemporary Problems 57, 3.
McNollgast (1995). "Politics and the courts: a positive theory of judicial doctrine and the rule of law". Southern California Law Review 68, 1631–1683.
McNollgast (1999). "The political origins of the administrative procedure act". Journal of Law, Economics, and Organization 15 (1), 180–217.
McNollgast (2006). "Conditions for judicial independence". Journal Contemporary Legal Issues.
Melnick, R.S. (1983). Regulation and the Courts: The Case of the Clean Air Act. Brookings Institution, Washington, D.C.
Moe, T. (1985). "The politicized presidency". In: Chubb, J.E., Peterson, P.E. (Eds.), The New Direction in American Politics. Brookings Institution, Washington, D.C.
Moe, T. (1987). "An assessment of the positive theory of congressional dominance". Legislative Studies Quarterly 12 (4), 475–520.
Moe, T. (1989). "The politics of bureaucratic structure". In: Chubb, J.E., Peterson, P.E. (Eds.), Can the Government Govern? Brookings Institution Press, Washington, D.C.
Moe, T., Howell, W. (1999a). "The presidential power of unilateral action". Journal of Law, Economics and Organization 15, 132–179.
Moe, T., Howell, W. (1999b). "Unilateral action and presidential power: a theory". Presidential Studies Quarterly 29, 850–873.
Moe, T., Wilson, S. (1994). "Presidents and the politics of structure". Law and Contemporary Problems 57 (2), 1–44.

Monroe, N.W. (2004). "The policy impact of unified government: evidence from the 2000 presidential election". Paper Presented at the Annual Meeting of the American Political Science Association, September 2–5, Chicago, IL.
Murphy, W. (1962). Elements of Judicial Strategy. The University of Chicago Press, Chicago.
Nagel, S.S. (1961). "Political party affiliation and judges' decisions". American Political Science Review 55, 843–850.
Nagel, S.S. (1969). The Legal Process from a Behavioral Perspective. Dorsey Press, Homewood.
Neustadt, R.E. (1960). Presidential Power: The Politics of Leadership. John Wiley & Sons, Inc., New York.
Niskanen, W. (1971). Bureaucracy and Representative Government. Aldine-Atherton, Chicago.
Noll, R.G. (1971). Reforming Regulation: An Evaluation of the Ash Council Proposals. Brookings Institution, Washington, D.C.
Noll, R.G. (1976). "Breaking out of the regulatory dilemma: alternatives to the sterile choice". Indiana Law Journal 51 (3), 686–699.
Noll, R.G. (1983). "The political foundations of regulatory policy". Journal of Institutional and Theoretical Economics 139 (3), 377–404.
Noll, R.G. (1985). "Government administrative behavior: a multidisciplinary survey and synthesis". In: Noll, R.G. (Ed.), Regulatory Policy and the Social Sciences. University of California Press, Berkeley.
Noll, R.G. (1989). "Economic perspectives on the politics of regulation". In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. II. North Holland Publishing Co., New York.
Noll, R.G., Owen, B.M. (1981). The Political Economy of Deregulation. American Enterprise Institute, Washington, D.C.
Noll, T.G., Weingast, B.R. (1991). "Rational actor theory, social norms, and policy implementation: applications to administrative processes and bureaucratic culture", co-author Barry R. Weingast. In: Monroe, K.R. (Ed.), The Economic Approach to Politics. Harper Collins, New York.
O'Connor, K., Epstein, L. (1983). "Beyond legislative lobbying: women's rights groups and the supreme court". Judicature 67, 134–143.
O'Connor, K. (1980). Women's Organizations' Use of the Court. Lexington Books, Lexington.
O'Brien, D. (1986). Storm Center: The Supreme Court in American Politics. Norton, New York.
Ogul, M. (1976). Congress Oversees the Bureaucracy. University of Pittsburgh Press, Pittsburgh.
O'Halloran, S. (1994). Politics, Process, and American Trade Policy. University of Michigan Press, Ann Arbor.
Oleszek, W. (2004). Congressional Procedures and the Policy Process, 6th edn. CQ Press, Washington, D.C.
Olson, M. (1965). The Logic of Collective Action: Public Goods and the Theory of Groups. Harvard University Press, Cambridge.
Oppenheimer, B.I. (1977). "The rules committee: new arm of leadership in a decentralized house". In: Dodd, L.C., Oppenheimer, B.I. (Eds.), Congress Reconsidered. Praeger, New York.
Peltason, J. (1955). The Federal Courts in the Political Process. Doubleday Press, Garden City.
Peltzman, S. (1976). "Toward a more general theory of regulation". Journal of Law and Economics 19, 211–240.
Penn, M.E. (2006). "The banks set in infinite spaces". Social Choice and Welfare 27 (3), 531–843.
Perry, H.W. Jr. (1991). Deciding to Decide: Agenda Setting in the United States Supreme Court. Harvard University Press, Cambridge.
Peterson, M.A. (1990). Legislating Together: The White House and Capitol Hill from Eisenhower to Reagan. Harvard University Press, Cambridge.
Petrocik, J. (1981). Party Coalitions: Realignments and the Decline of The New Deal Party System. University of Chicago Press, Chicago.
Pfiffner, J.P. (1979). The President, the Budget, and Congress: Impoundment and the 1974 Budget Act. Westview Press, Boulder.
Pisani, D.J. (2002). "A tale of two commissioners: Frederick H. Newell and Floyd Dominy". Presented at History of the Bureau of Reclamation: A Symposium, Las Vegas, NV, June 18.
Polinsky, A.M. (1989). An Introduction to Law and Economics. Little, Brown, and Company, Boston.

Popkin, S.L. (1991). The Reasoning Voter: Communication and Persuasion in Presidential Campaigns. The University of Chicago Press, Chicago.
Posner, R.A. (1974). "Theories of economic regulation". Bell Journal of Economics 5, 335–358.
Posner, R.A. (1986). Economic Analysis of Law. Little, Brown, and Company, Boston.
Posner, R.A. (1990). The Problems of Jurisprudence. Harvard University Press, Cambridge, MA.
Posner, R.A. (1995). Overcoming Law. Harvard University Press, Cambridge, MA.
Pound, R. (1931). "The call for a realist jurisprudence". Harvard Law Review 44, 697.
Pritchett, C.H. (1948). The Roosevelt Court: A Study in Judicial Politics and Values, 1937–1947. Macmillan Company, New York.
Ramseyer, J.M., Rasmussen, E. (1994). "Cheap bribes and the corruption ban: a coordination game among rational legislators". Public Choice 78, 305–327.
Riker, W.H. (1962). The Theory of Political Coalitions. Yale University Press, New Haven.
Riker, W. (1982). Liberalism against Populism: A Confrontation Between the Theory of Democracy and the Theory of Social Choice. W.H. Freeman, San Francisco.
Riker, W., Ordeshook, P. (1973). An Introduction to Positive Political Theory. Prentice-Hall, Englewood Cliffs.
Riker, W., Weingast, B. (1988). "Constitutional regulation of legislative choice: the political consequences of judicial deference to legislatures". Virginia Law Review 74 (2), 372–402.
Rodriguez, D.B., Weingast, B.R. (2003). "The positive political theory of legislative history: new perspectives on the 1964 Civil Rights Act and its interpretation". University of Pennsylvania Law Review 151, 1417.
Rohde, D. (1991). Parties and Leaders in the Postreform House. The University of Chicago Press, Chicago.
Rohde, D., Simon, D. (1985). "Presidential vetoes and congressional response: a study of institutional conflict". American Journal of Political Science 29, 397–427.
Rohde, D., Spaeth, H. (1976). Supreme Court Decision-making. W.H. Freeman, San Francisco.
Romano, R. (1993). The Genius of American Corporate Law. AEI Press, Washington, D.C.
Romano, R. (1994). "Comment on presidents and the politics of structure". Law and Contemporary Problems 57 (2), 59–64.
Romer, T., Rosenthal, H. (1978). "Political resource allocation, controlled agendas, and the status quo". Public Choice 33, 27–44.
Rosenberg, G. (1991). The Hollow Hope: Can Courts Bring About Social Change? The University of Chicago Press, Chicago.
Saalfeld, T. (1997). "Professionalization of parliamentary roles in Germany: an aggregate level analysis 1949–1994". In: Müler, W., Saalfeld (Eds.), Members of Parliament in Western Europe: Roles and Behavior. Frank Cass, Portland.
Schelling, T.C. (1960). The Strategy of Conflict. Harward University Press, Cambridge, MA.
Schubert, G. (1959). Quantitative Analysis of Judicial Behavior. Free Press, Glencoe.
Schubert, G. (1965). The Judicial Mind: The Attitudes and Ideologies of Supreme Court Justices, 1946–1963. Northwestern University Press, Evanston.
Schuck, P. (1994). Foundations of Administrative Law. Oxford University Press, New York.
Schwartz, E. (1992). "Policy, precedent, and power: a positive theory of supreme court decision-making". Journal of Law, Economics, and Organization 8, 219–252.
Schwartz, E., Spiller, P.T., Urbiztondo, S. (1994). "A positive theory of legislative intent". Law and Contemporary Problems 57, 51–76.
Segal, J.A. (1984). "Predicting supreme court cases probablistically: the search and seizure cases, 1962-81". American Political Science Review 78, 891–900.
Segal, J., Spaeth, H. (1993). The Supreme Court and the Attitudinal Model. Cambridge University Press, Cambridge.
Seidenfeld, M. (1992). "A civic republican justification for the bureaucratic state". Harvard Law Review 105, 1511.
Shapiro, M. (1964). Law and Politics in the Supreme Court: New Approaches to Political Jurisprudence. Free Press of Glencoe, New York.

Shavell, S. (1987). Economic Analysis of Accident Law. Harvard University Press, Cambridge.
Shavell, S. (2004). "The appeals process and adjudicator incentives". Working Paper.
Shepsle, K. (1979). "Institutional arrangements and equilibrium in multidimensional voting models". American Journal of Political Science 23, 27–59.
Shepsle, K. (1992). "Congress is a 'they' not an 'it': legislative intent as an oxymoron". International Review of Law and Economics 12, 239–256.
Shepsle, K., Weingast, B. (1981). "Structure-induced equilibrium and legislative choice". Public Choice 37, 503–519.
Shepsle, K., Weingast, B. (1987). "The institutional foundations of committee power". American Political Science Review 81, 85–104.
Shugart, M.S., Carey, J.M. (1992). Presidents and Assemblies: Constitutional Design and Electoral Dynamics. Cambridge University Press, New York.
Shugart, M.S., Haggard, S. (2001). "Institutions and public policy in presidential system". In: McCubbins, M.D., Haggard, S. (Eds.), Structure and Policy in Presidential Democracies. Cambridge University Press, New York.
Sinclair, B. (1983). Majority Leadership in the U.S. House. Johns Hopkins Press, Baltimore.
Sinclair, B. (1995). Legislators, Leaders, and Lawmaking: The House of Representatives in the Postreform Era. Johns Hopkins University Press, Baltimore.
Sinclair, B. (2002). "Do parties matter?" In: Brady, D., McCubbins, M.D. (Eds.), Party, Process, and Political Change: New Perspectives on the History of Congress. Stanford University Press, Stanford.
Smith, A. (1776). An Inquiry into the Nature and Causes of the Wealth of Nation.
Smith, S.S. (1989). Call to Order: Floor Politics in the House and Senate. Brookings Institution, Washington, D.C.
Smith, S.S., Deering, C.J. (1984). Committees in Congress. CQ Press, Washington, D.C.
Snyder, J. (1992). "Committee power, structure-induced equilibrium, and roll call votes". American Journal of Political Science 36, 1–30.
Snyder, S.K., Weingast, B.R. (2000). "The American system of shared powers: The President, Congress, and the NLRB". Journal of Law, Economics, and Organization 16 (2), 269–305.
Songer, D.R., Lindquist, S.A. (1996). "Not the whole story: the impact of justices' values on supreme court decision-making". American Journal of Political Science 40, 1049–1063.
Songer, D.R., Reid, S. (1989). "Policy change on the U.S. courts of appeals: exploring the contribution of the legal and democratic subcultures". Typescript.
Songer, D.R., Segal, J.A., Cameron, C.M. (1994). "The hierarchy of justice: testing a principal-agent model of supreme court-circuit court interactions". American Political Science Review 38, 673–696.
Spaeth, H. (1963). Supreme Court Policy-making: Explanation and Prediction. W.H. Freeman, San Francisco.
Spiller, P.T. (1992). "Rationality, decision rules, and collegial courts". International Review of Law and Economics 12, 186–190.
Spiller, P.T., Spitzer, M. (1992). "Judicial choice of legal doctrines". Journal of Law, Economics and Organization 8, 8–46.
Staudt, N. (2005). "On congress". Working Paper, School of Law, Washington University.
Stewart, C.H. III (1989). Budget Reform Politics: The Design of the Appropriations Process in the House of Representatives, 1865–1921. Cambridge University Press, New York.
Stewart, R.B. (1975). "The reformation of American Administrative Law". Harvard Law Review 88, 1667–1814.
Stigler, G.J. (1971). "The theory of economic regulation". Bell Journal of Economics 2, 3–21.
Studenski, P., Krooss, H.E. (1963). Financial History of the United States: Fiscal, Monetary, Banking, and Tariff, Including Financial Administration and State and Local Finance. McGraw-Hill Book Company, New York.
Sullivan, J.L., O'Connor, R.E. (1972). "Electoral choice and popular control of public policy: the case of the 1966 house elections". The American Political Science Review 66, 1256–1268.
Sundquist, J.L. (1981). The Decline and Resurgence of Congress. Brookings Institution, Washington, D.C.

Sundquist, J.L. (1988). "Needed: a political theory for the new era of coalition government in the United States". Political Science Quarterly 103, 613–635.
Sundquist, J.L. (1992). Constitutional Reform and Effective Government, revised edition. Brookings Institution, Washington, D.C.
Sunstein, C. (1985). "Interest groups in American Public Law". Stanford Law Review 38, 29–87.
Sunstein, C. (1990). After the Rights Revolution: Reconceiving the Regulatory State. Harvard University Press, Cambridge.
Symposium on Law and Public Choice (1988). Virginia Law Review 74, 2.
Taagepera, R., Shugart, M.S. (1989). Seats and Votes: The Effects and Determinants of Electoral Systems. Yale University Press, New Haven.
Tanenhaus, J., Schick, M., Muraskin, M., Rosen, D. (1963). "The supreme court's certiorari jurisdiction: cue theory". In: Schubert (Ed.), Judicial Decision-making. Free Press of Glencoe, New York.
Truman, D. (1951). The Governmental Process: Political Interests and Public Opinion. Knopf, New York.
Tullock, G. (1965). The Politics of Bureaucracy. Public Affairs Press, Washington.
Vietor, R. (1984). Energy Policy in America Since 1945: A Study of Business-Government Relations. Cambridge University Press, Cambridge.
Walsh, K.T. (2003). Air Force One: A History of Presidents and their Planes. Hyperion, New York.
Weber, M. (1946). "Bureauacracy". In: Gerth, H., Mills, C.W. (Eds.), From Max Weber. Oxford University Press, Oxford.
Weber, M. (1968). Economy and Society: An Outline of Interpretive Sociology. Roth, Wittich, C. (Eds.), Translators: Fischoff, E., et al. Bedminster Press, New York.
Wechsler, H. (1959). "Toward neutral principles of constitutional law". Harvard Law Review 73 (1), 31.
Weingast, B.R. (1979). "A rational choice perspective on congressional norms". American Journal of Political Science 23, 245–262.
Weingast, B.R. (1984). "The congressional bureaucratic system: a principal–agent perspective". Public Choice 44, 147–192.
Weingast, B. (1994). "Reflections on distributive politics and universalism". Political Research Quarterly 47, 319–327.
Weingast, B.R. (2002). "Rational choice institutionalism". In: Katznelson, I., Milnor, H. (Eds.), Political Science: State of the Discipline. Norton, New York.
Weingast, B., Marshall, W. (1988). "The industrial organization of congress". Journal of Political Economy 96, 132–163.
Weingast, B., Shepsle, K., Johnson, C. (1981). "The political economy of benefits and costs: a neoclassical approach to distributive politics". Journal of Political Economy 89, 642–664.
Weingast, B.R., Moran, M.J. (1983). "Bureaucracy discretion or congressional control? Regulatory policymaking by the Federal Trade Commission". Journal of Political Economy 91, 765–800.
Whalen, Ch., Whalen, B. (1985). The Longest Debate: A Legislative History of the 1964 Civil Rights Act. Seven Locks Press, Washington, D.C.
Wildavsky, A. (1964). The Politics of the Budgetary Process. Little, Brown, and Company, Boston.
Wilmerding, L. (1943). The Spending Power: A History of the Efforts of Congress to Control Expenditures. Yale University Press, New Haven.
Wilson, J.Q. (1980). The Politics of Regulation. Basic Books, New York.
Wilson, W. (1885). Congressional Government. Houghton, Mifflin, Boston.
Woll, P. (1977). American Bureaucracy. Norton, New York.
Wood, B.D., Waterman, R.W. (1994). Bureaucratic Dynamics: The Role of Bureaucracy in a Democracy. Westview Press, Boulder.
Wooley, J. (1991). "Institutions, election cycles, and the presidential vetoes". American Journal of Political Science 35, 279–302.
Young, J.S. (1966). The Washington Community 1800–1828. Columbia University Press, New York.
Zuk, G., Gryski, G.S., Barrow, D. (1993). "Partisan transformation of the federal judiciary, 1869–1992". American Politics Quarterly 21, 439–457.

AUTHOR INDEX OF VOLUME 2

n indicates citation in a footnote.

SUBJECT INDEX OF VOLUME 2