I·R·L·E

INSTITUTE FOR REGULATORY
LAW & ECONOMICS

2012 Seminar
May 19, 2012- May 23, 2012
Aspen Meadows Resort – Home of the Aspen Institute
845 Meadows Road, Aspen, CO 81611
Phone: 1.800.452.4240

AGENDA

Saturday, May 19- Welcome to Aspen; Introduction to Modern Economic Analysis

Time	Session	Instructor(s)	Location
11:00 a.m. – 2:00 p.m.	Resort check in		
2:15 – 2:30 p.m.	Seminar Registration & Refreshments		IRLE Seminar Room
2:30 – 2:45 p.m.	Welcome and Orientation	Ray Gifford	IRLE Seminar Room
2:45 – 3:45 p.m.	The Five Prisms: Economic Analysis of Regulation	Ray Gifford	IRLE Seminar Room
3:45 – 4:00 p.m.	Break		
4:00 – 5:45 p.m.	The Double Oral Icebreaker	Lynne Kiesling	IRLE Seminar Room
5:45 – 6:45 p.m.	Optional Happy Hour		Meadows Restaurant
6:45 – 8:30 p.m.	Dinner & Talk: Antitrust 101: The Triumph of Neoclassical Analysis	Howard Shelanski	Meadows Restaurant

Reading Before Seminar

- **T**homas McCraw, *Prophets of Regulation,* chapters 2 and 7 on Charles Francis Adams and Alfred Kahn and Chapter 8, Regulation Reconsidered (1984).
- Ronald H. Coase, *Nobel Prize Lecture* (1991).
- Pierre Lemieux, *The Public Choice Revolution*, Regulation (Fall 2004), pp. 22-29.
- Richard A. Posner, *Taxation by Regulation*, 3 Bell J. Econ. 22 (1971).
- Joseph A. Schumpeter, *Capitalism, Socialism & Democracy* (1942), Chapter 7.
- F.A. Hayek, *The Use of Knowledge in Society*, The American Economic Review (1945).
- Oliver Williamson, *Why Law Economics and Organization?*, Annual Review of Law and Social Science (Dec. 2005), Vol. 1, Pages 369-396.

Assigned Reading for Saturday

- J. Bradford DeLong, *Creative Destruction's Reconstruction: Joseph Schumpeter Revisited, Chronicle of Higher Education*.
- Alfred E. Kahn, *The Deregulatory Tar Baby*, 21 J. Reg. Econ. 35 (2002).
- Timothy Muris, *Improving the Economic Foundations of Competition Policy*.
- Richard A. Posner, *Antitrust in the New Economy*, University of Chicago Working Paper (2000).

Sunday, May 20 - Core Economic Tools for Regulators and Corporate Finance

Time	Session	Instructor(s)	Location
7:00 – 8:15 a.m.	Hot Breakfast Buffet available in the Meadows Restaurant		
8:15 – 8:30 a.m.	Light Breakfast Buffet available outside the IRLE Seminar Room in the Koch Lobby		
8:30 – 10:15 a.m.	Trends in Technology Development –Communications and Electricity Networks	Doug Sicker	IRLE Seminar Room
10:15 – 10:30 a.m.	Break		
10:30 – 12:00 p.m.	Policy Implications of Modern Networks	Doug Sicker	IRLE Seminar Room
12:00 –1:30 p.m.	Lunch		Meadows Restaurant
1:30 – 3:15 p.m.	Introduction to Corporate Finance	Eric Talley	IRLE Seminar Room
3:15 – 3:30 p.m.	Break		
3:30– 5:00 p.m.	Applications of Corporate Finance To Regulated Industries	Eric Talley	IRLE Seminar Room
5:00 – 6:30 p.m.	Break		
6:30 – 8:30 p.m.	Dinner & Talk: Cybersecurity and Innovation Policy: Undiscovered County for State Utility Regulators	Phil Weiser	Meadows Restaurant
8:30 – 9:30 p.m.	Evening Seminar (Optional): The Credit Crisis and the Financing of Critical infrastructure	Eric Talley	Hefner Lounge

Assigned Readings for Sunday:

- Timothy Berners-Lee, Testimony before the United States House of Representatives Committee on Energy and Commerce, Subcommittee on Telecommunications and the Internet Hearing on *"Digital Future of the United states: Part I – The Future of the World Wide Web."*
- James DeLong, *Avoiding a Tech Train Wreck* (2008).
- David Friedman, *The Swedes Get It Right*.
- Douglass C. North, *Institutions and the Performance of Economies Over Time*.
- Ivo Welch, *A First Course in Corporate Finance* (2008) Chapters 2 & 8.
- Economist Intelligence Unit, *Regulatory Risk: Trends and Strategies for the CRO* (2005).
- Timothy A. Luehrman, *Investment Opportunities as Real Options* (1998).
- Public Utilities Fortnightly, 2009 ROE Survey (Sept. 2009).
- Ian Alexander and Timothy Irwin, *Price Caps, Rate-of-Return Regulation, and the Cost of Capital* (1996).
- M. Schmidt, *Regulation by Formula: Tools to Facilitate Changing Utility Economics*, (March 2010).
- Dierdre K. Mulligan & Fred B. Schneider, *Doctrine for Cybersecurity*, (2011).
- (Optional) Joseph Farrell & Philip J. Weiser, *Modularity, Vertical Integration, And Open Access Policies: Towards A Convergence Of Antitrust And Regulation In The Internet Age* (2003) pp. 86-90 and 97-127.

Monday, May 21 – Economic Tools, Technological Change, and Market Processes

Time	Session	Instructor(s)	Location
7:00 – 8:15 a.m.	Hot Breakfast Buffet available in the Meadows Restaurant		
8:15 – 8:30 a.m.	Light Breakfast Buffet available outside the IRLE Seminar Room in the Koch Lobby		
8:30 – 9:30 a.m.	Public Choice & Political Entrepreneurship	Phil Weiser	IRLE Seminar Room
9:30 – 10:30 a.m.	Introduction to New Institutional Economics and Vertical Integration	Howard Shelanski	IRLE Seminar Room
10:30 – 10:45 a.m.	Break		
10:45 – 12:00 p.m.	Applications of New Institutional Economics	Howard Shelanski	IRLE Seminar Room
12:00 p.m.	Lunch		Meadows Restaurant
1:30 – 3:00 p.m.	Economics of Networks, Sunk Costs, Rate Structure and Interoperability	Ray Gifford and Phil Weiser	IRLE Seminar Room
3:00 – 3:15 p.m.	Break		
3:15 p.m. – 5:00 p.m.	Regulation and Technological Change in Electric Power	Lynne Kiesling	IRLE Seminar Room
5:00 – 6:30 p.m.	Break		
6:30 – 8:30 p.m.	Dinner & Talk: Technology, Demand Response, and Consumer Property Rights in Electricity	Lynne Kiesling	Meadows Restaurant
8:30 – 9:30 p.m.	Evening Seminar (Optional): Error Costs, Stasis and Rent-Seeking: Is Our Regulatory System Worth it?	Howard Shelanski and Ray Gifford	Hefner Lounge

Assigned Reading for Monday

- Paul Joskow, *Incentive Regulation and Its Application to Electricity Networks* (RNE 2008).
- Paul Joskow, *Creating a Smarter Electricity Grid* (Winter 2012).
- L. Lynne Kiesling, *Promoting Innovation in the Electricity Industry*
- Dave Chassin and L. Lynne Kiesling, *Decentralized Coordination through Digital Technology, Dynamic Pricing, and Customer-Driven Control: The GridWise Testbed Demonstration Project* (2008).
- Phil Weiser, *Alfred Kahn as a Case Study of a Political Entrepreneur: An Essay in Honor of His 90th Birthday* (2008).
- Mark Armstrong and David E. M. Sappington, *Regulation, Competition, and Liberalization*, Journal of Economic Literature (June 2006).
- Dave Bohon, *Army Corps of Engineers Blows Mississippi River Levee* (2011).
- Daniel Gilbert & Rebecca Smith, *More Utilities Can't Prove Pipe Tests*, Wall Street Journal (Apr. 18, 2011).
- Thomas Catan, *Apple's Mobile Rules To Get FTC Scrutiny*, The Wall Street Journal (June 12, 2010).
- (optional) Ray Gifford and Eric Gunning, *The Opportunity and Peril of Smart Grid* (2010).
- (optional) Elizabeth Hartman and Kaleb Sieh, *State Regulatory Opportunities and Impediments to Smart Grid* (2010).
- (optional) Clifford Winston, *U.S. Industry Adjustment to Economic Deregulation*, 12 J. Econ. Persp. 89 (1998).
- (optional) *Town of Concord v. Boston Edison Co.*, 915 F.2d 17 (1st Cir. 1990).

Tuesday, May 22- Auctions and Experimental Economics

Time	Session	Instructor(s)	Location
7:00 – 8:15 a.m.	Hot Breakfast Buffet available in the Meadows Restaurant		
8:15 – 8:30 a.m.	Light Breakfast Buffet available outside the IRLE Seminar Room in the Koch Lobby		
8:30 – 10:00 a.m.	Electric Power Auctions	Rim Baltaduonis	IRLE Seminar Room
10:00 – 10:15 a.m.	Break		
10:15 – 12:00 p.m.	Experimental Economics 101	Rim Baltaduonis	IRLE Seminar Room
12:00 p.m.	Lunch & Break		Meadows Restaurant
5:30 p.m.	Shuttle to Dinner		Front Lobby
6:00 – 8:00 p.m.	Dinner		**LA CANTINA** **411 East Main Street** **Aspen, Colorado 81611** **970.925.3663**

Assigned Readings for Tuesday:

- Stephen Rassenti, Vernon Smith & Bart Wilson, *Using Experiments to Inform the Privatization/Deregulation Movement in Electricity*, Cato Journal (Winter 2002).
- Rimvydas Baltaduonis, *Simple-Offer versus Complex-Offer Auctions in Deregulated Electricity Markets*, Working Paper (2009).
- Rimvydas Baltaduonis, Tihomir Ancev, Tim Capon & Taylor Smart, *Flicking the Switch: Retail Demand-Side Response under Alternative Electricity Pricing Contracts*, Working Paper (2012).

Wednesday, May 23 - Putting It All Together: Back to the Real World

Time	Session	Instructor(s)	Location
7:00 – 8:15 a.m.	Hot Breakfast Buffet available in the Meadows Restaurant		
8:15 – 8:30 a.m.	Light Breakfast Buffet available outside the IRLE Seminar Room in the Koch Lobby		
8: 30 – 10:30 a.m.	Wrap-up	IRLE Faculty	IRLE Seminar Room
10:30-11:15 a.m.	Closing Comments Program Review and Feedback	All	IRLE Seminar Room
11:15 a.m.	Boxed Lunches Available at the Lobby Main Bar or hot lunch available at Meadows Restaurant		

2011 IRLE Faculty

Howard Shelanski
hshelanski@yahoo.com

Howard Shelanski received his J.D. from the University of California at Berkeley in 1992 and his Ph.D. in economics from Berkeley in 1993. After graduating from law school he clerked for Judge Stephen F. Williams of the U.S. Court of Appeals for the D.C. Circuit, Judge Louis H. Pollak of the U.S. District Court in Philadelphia, and Justice Antonin Scalia of the United States Supreme Court. Professor Shelanski practiced law with the Washington, D.C. firm of Kellogg Huber Hansen Todd & Evans before joining the Berkeley faculty in 1997, where he remained until coming to Georgetown in 2011.

Professor Shelanski has taken leave from teaching to work in government on several occasions. From 2009-2011 he was Deputy Director of the Bureau of Economics at the Federal Trade Commission. From 1999-2000 he served as Chief Economist of the Federal Communications Commission and from 1998-1999 he served as a Senior Economist for the President's Council of Economic Advisers at the White House. Professor Shelanski's teaching and research focus on antitrust, regulation, and telecommunications policy. In 2004, he received Berkeley Law's Rutter Award for Teaching Distinction.

Phil Weiser
Phil.Weiser@colorado.edu

Dean Phil Weiser is the Dean of the Law School, Thompson Professor of Law, and Executive Director and Founder of the Silicon Flatirons Center for Law, Technology, and Entrepreneurship at the University of Colorado. Dean Weiser re-joined the Colorado faculty in June, 2011. From April 2010-June 2011, he served as the Senior Advisor for Technology and Innovation to the National Economic Council Director at the White House. From July 2009-April 2010, he served as the Deputy Assistant Attorney General at the United States Department of Justice's Antitrust Division.

Since first joining the CU faculty in 1999, Dean Weiser has worked to establish a national center of excellence in telecommunications and technology law, founding the Journal on Telecommunications & High Technology Law and the Silicon Flatirons Center for Law, Technology, and Entrepreneurship as well as writing and teaching in the areas of competition policy, innovation policy, and Internet policy. Over the last ten years, Weiser has co-authored three books (The Jury and Democracy: How Jury Deliberation Promotes Civic Engagement and Political Participation (Oxford University Press 2010), Telecommunications Law and Policy (Carolina Academic Press 2006), and Digital Crossroads: American Telecommunications Policy in the Internet Age (MIT Press 2005)), written numerous articles (in both law journals and publications such as the Washington Post and Foreign Affairs), and testified before both houses of Congress. He also remained engaged in public service, arguing a number of *pro bono* cases before the Tenth Circuit Court of Appeals, co-chairing the Colorado Innovation Council, and serving as the lead agency reviewer for the Federal Trade Commission as part of the 2008 Presidential Transition.

Prior to joining the Colorado Law faculty, Dean Weiser served as senior counsel to the Assistant Attorney General in charge of the Antitrust Division at the United States Department of Justice, advising him primarily on telecommunications matters. Before his appointment at the Justice Department, Weiser served as a law clerk to Justices Byron R. White and Ruth Bader Ginsburg at the United States Supreme Court and to Judge David Ebel at the Tenth Circuit Court of Appeals. Weiser graduated with high honors from both the New York University School of Law and Swarthmore College.

Rimvydas Baltaduonis
rbaltadu@gettysburg.edu
Rim Baltaduonis is an assistant professor in the Economics Department at Gettysburg College, Pennsylvania. Rim's broad fields of specialty are industrial organization, experimental economics, and market design. His current research is focused on energy markets, especially deregulated electricity markets. Lately, Rim examined the performance of auction mechanisms in wholesale power markets when electricity suppliers act strategically, the consumer behavior under alternative retail electricity pricing contracts, the efficiency of markets for financial transmission rights. At Gettysburg College, he teaches Industrial Organization and Public Policy, Energy Economics, Health Economics, and Experimental Economics. The National Science Foundation, the International Foundation for Research in Experimental Economics and the Australian Research Council have supported his research. Rim is also an Affiliate Researcher at the Resources, Energy and Environmental Markets Lab in the University of Sydney. Prior to assuming his position at Gettysburg College Rim was an IFREE Visiting Post-doctoral Fellow in the Interdisciplinary Center for Economic Science at George Mason University and later at the Economics Science Institute at Chapman University. He holds a Ph.D. from the University of Connecticut.

Raymond L. Gifford
Ray.Gifford@Colorado.EDU
Ray Gifford is the Director of the Institute of Regulatory Law and Economics, a senior adjunct fellow at the University of Colorado's Silicon Flatirons Center, member of its advisory board, and is a Partner at Wilkinson Barker Knauer, LLP. He counsels communications, electric and gas utilities and information technology companies on state and federal aspects of regulation, administrative law, and competition policy. He is an expert in public utilities law, principally as it relates to telecommunications and energy matters. He litigates on behalf of clients in state and federal courts, serving as an expert witness on utility regulation in federal and state courts, and at state utility commissions. In addition, his law and policy work focuses on antitrust, competition policy and intellectual property law. Before returning to private law practice, Ray was Chairman of the Colorado Public Utilities Commission, serving as the chief regulator in the state of Colorado of the telecommunications, electricity, natural gas and transportation industries.

L. Lynne Kiesling
lynne@knowledgeproblem.com
Lynne Kiesling is a Distinguished Senior Lecturer in the Department of Economics at Northwestern University. At Northwestern she is also a Faculty Member in the Northwestern Institute on Complex Systems (NICO) and a Faculty Affiliate in the Center for the Study of Industrial Organization (CSIO). Her specialization is industrial organization and regulatory policy. She also teaches undergraduate courses in principles of economics, environmental economics, and history of economic thought, and writes about economics as the editor/owner at the website Knowledge Problem. As a noted expert in demand response, end-use technology, and retail competition, she has been asked to speak to various academic, industrial, and regulatory groups about regulatory policy, institutional change, and experimental economic analysis of electric power market design. She has served as a peer reviewer for the U.S. Department of Energy and the National Science Foundation, and for academic journals including Energy Journal, Public Choice, and Review of Economics and Statistics. She has provided expert testimony in proceedings before the Federal Energy Regulatory Commission, the California Public Utilities Commission, the Illinois Commerce Commission, and the New York Public Service Commission. Her full CV is available at http://faculty.wcas.northwestern.edu/~lki851/vita.pdf.

Douglas C. Sicker
Douglas.Sicker@colorado.edu

Douglas Sicker is NTIA's Chief Technology Officer and Senior Advisor for Spectrum. He is also an endowed professor in the Department of Computer Science at the University of Colorado at Boulder, with a joint appointment in the Interdisciplinary Telecommunications Program.

Dr. Sicker has held various positions in academia, industry, and government. Before joining NTIA, he was the Federal Communications Commission's Chief Technologist. Previously, he served as a senior advisor on the FCC National Broadband Plan and, before that, as Director of Global Architecture at Level 3 Communications, Inc.

Dr. Sicker has also served as Chief of the FCC's Network Technology Division. After leaving this agency, he served as Chair of the Network Reliability and Interoperability Council steering committee, an FCC federal advisory committee that focuses on network reliability, wireline spectral integrity and Internet peering and interconnection. He also served on the Technical Advisory Council of the FCC. In addition, he has also held faculty and industry positions in the field of medical sciences.

Dr. Sicker is a senior member of the IEEE, as well as a member of the ACM and the Internet Society. He has chaired and served on the program committees of numerous technical conferences including IEEE, DySPAN, ISART and TPRC. His research interests include network and wireless systems, network security, and telecommunications policy. He has research funded through the NSF, DARPA, the Internet Society and the Federal Aviation Administration. He holds a Ph.D. from the University of Pittsburgh. A full CV is available on his faculty page at the University of Colorado at Boulder.

Eric Talley
etalley@law.berkeley.edu

Eric Talley is the Rosalinde and Arthur Gilbert Professor in Law, Business and the Economy at the University of California Berkeley (Boalt Hall) School of Law. He is a leading expert in the intersection of corporate law, governance and finance. For over 15 years, he has taught courses in law schools, economics departments and business schools, in topics that include Corporate Law, Commercial Law, Microeconomics, Corporate Finance, Mergers and Acquisitions, Quantitative Methods, Game Theory, Securities Regulation, and Law and Economics. He is also affiliated with the RAND Corporation's Institute for Civil Justice. Eric holds a bachelor's degree from the University of California, San Diego, a Ph.D. in economics from Stanford University, and a J.D. also from Stanford, where he served as articles editor for the Stanford Law Review. He has taught at UC Berkeley since 2005, serving as co-director of the Berkeley Center for Law, Business and the Economy. He has held teaching appointments at Harvard, Georgetown, California Institute of Technology, Stanford, University of Sydney (Aus), U. Miami, and the University of Southern California. From 2004-2006, he was the Ivadelle & Theodore Johnson Professor of Law and Business at the University of Southern California, with joint appointments in the Gould School of Law and the Marshall School of Business (finance and business economics group). From 2001-2004, he served as director of the USC Center for Law, Economics and Organization and of USC's Olin Program in Law and Rational Choice. He is a frequent commentator on the national radio show Marketplace, and speaks frequently to corporate boards and regulators on issues pertaining to fiduciary duties, governance, and corporate finance.

Fellow & Staff

Therese Kerfoot
Research Fellow
Silicon Flatirons Center
Cell: 817.521.6903
Therese.kerfoot@gmail.com

Anna Noschese
Program Director
Silicon Flatirons Center
401 UCB
Boulder, CO 80309
Office: 303.735.5633
Cell: 303.601.4105
Anna.noschese@colorado.edu

Attendee Information: See list on next page

Saturday

May 19, 2012

HINE COMMISSION

rt of muckraking, and for his own
dvantageous in the practical art of

o suggest that Adams himself ever
riumph as muckraker and regulator.
ght of his "Erie" at all, he recalled
ld had turned the tables on him, se-
from the Union Pacific presidency,
acid portrait of 1869. But in later
e time puzzling over the decline of
he rise of coercive regulation. In ret-
trend in regulation come into much
in Adams' time. The most powerful
vision of the nation's people into de-
n of these groups drove hard not to-
soned peace and harmony Adams
the power and autonomy that each
Adams, bewildered and embittered
helplessly from the sidelines as his
e apart among dozens of quarreling
encies.

xty-five at the turn of the twentieth
trialization was proceeding at full
growing at an unprecedented rate,
re entering the United States at the
he census of 1900 disclosed the re-
n one American in two was both
ive-born parents. From his position
lass, Adams reacted to this new, di-
th downright hostility. In this re-
n subsequent understanding of the
s' bitterness at the failure of public
ine commission he had so carefully
tion from what he began to deride
olatant exaggeration, perhaps from
nse of elitism into a hateful carica-
o memorably etched of the robber
e signs, one can read the sea change
society and therefore in the condi-

CHAPTER 2

State to Federal, Railroads to Trusts

IN THE LATTER PART of the nineteenth century, Charles Francis Adams' Massachusetts Board of Railroad Commissioners took its place as the national prototype of the so-called weak regulatory commission. It became a model for many other states whose legislatures wished to regulate railroads but, for one reason or another, did not want to assume the power of setting rates and fares. An important alternative to the weak commission appeared in the midwestern states of Illinois, Wisconsin, Minnesota, and Iowa. Beginning in 1871, the legislatures of these states passed Granger Laws, which directly regulated the rates of railroads operating within each state. Three of the four states also created "strong" regulatory commissions and empowered them to prescribe maximum rates.[1]

Although significant as milestones in the history of regulation, this midwestern variety of strong commissions proved unsuccessful. Most of the Granger Laws were repealed within a few years of their enactment, for reasons that suggest a good deal about the overall nature of regulation. One stimulus to repeal was a political counterattack by the railroads. Another, more important reason derived from the severe economic depression of the 1870s, which, in both the Middle West and in Massachusetts, drove rates down faster than regulatory agencies could hope to do. Still a third reason, and the most revealing of all, lay in the very nature of the national railroad network: it simply made no sense for small state agencies to perform pricing functions for giant interstate corporations whose tracks crossed thousands of miles in a dozen different states. In the Granger cases of 1877, the United States Supreme

Court firmly upheld the power of public agencies to regulate railroads and other industries "affected with a public interest." But in practice, the Granger commissions did not achieve as much success as their misnamed weak counterparts to the east.[2]

The model weak agency, that of Massachusetts, continued to enjoy a measure of prestige even after Adams' departure in 1879. Such decline as its reputation did suffer occurred only gradually, as a result of several trends common to state railroad commissions during the late nineteenth century. For one thing, the onset of federal railroad regulation in 1887 reduced the importance of all state commissions. For another, no new commissioners as talented as Adams appeared in Massachusetts—a typical example of the difficulty nearly all regulatory agencies have experienced in consistently attracting the best people. Finally, the Massachusetts agency like other state railroad commissions went through a sharp transition in function. One by one, each agency evolved into a public utility commission.[3]

This shift in function occurred more rapidly in Massachusetts than elsewhere, but the change there was also representative of evolutions in other states. In 1885, Massachusetts created a second regulatory agency, the Board of Gas Commissioners, with supervisory authority over the new gas-lighting companies that were bringing artificial illumination to the state's cities and towns. In 1887, this agency became the Board of Gas and Electric Light Commissioners, as the state took formal notice of the rise of still another vital new industry. The gas and electric commissioners, following the pattern of the Board of Railroad Commissioners, did not wield strong authority over rates but tried instead to regulate more gently, mediating a host of difficult issues between the corporations and the public. In 1909, both the original railroad board and the newer gas and electric commission received important powers over the issuance of stocks and bonds by the companies. Four years after that, in 1913, Adams' original railroad commission was renamed the Public Utility Commission and given additional responsibility for telephone companies, street railways, and steamships. Finally, in 1919, the last institutional consolidation occurred, as the railroad and utility commissions were merged into one body, which was named the Massachusetts Department of Public Utilities. This organization still exists today, having grown over the course of a hundred years from its seed as Adams'

of public agencies to regulate rail-
ted with a public interest." But in
ons did not achieve as much suc-
nterparts to the east.[2]
t of Massachusetts, continued to
after Adams' departure in 1879.
d suffer occurred only gradually,
non to state railroad commissions
y. For one thing, the onset of fed-
educed the importance of all state
ew commissioners as talented as
s—a typical example of the diffi-
cies have experienced in consis-
ple. Finally, the Massachusetts
ommissions went through a sharp
ne, each agency evolved into a

l more rapidly in Massachusetts
there was also representative of
, Massachusetts created a second
Gas Commissioners, with supervi-
s-lighting companies that were
the state's cities and towns. In
oard of Gas and Electric Light
formal notice of the rise of still
gas and electric commissioners,
of Railroad Commissioners, did
tes but tried instead to regulate
difficult issues between the cor-
both the original railroad board
commission received important
s and bonds by the companies.
lams' original railroad commis-
ty Commission and given addi-
companies, street railways, and
last institutional consolidation
lity commissions were merged
the Massachusetts Department
tion still exists today, having
d years from its seed as Adams'

tiny agency, with its single clerk, to a small bureaucracy with more than one hundred full-time employees.[4]

Much the same pattern of institutional growth occurred in other states, as the American industrial economy began to mature. Led by New York and Wisconsin (whose commissions were reorganized in 1907), state agencies everywhere, in their efforts to do justice to both companies and consumers, focused their regulatory efforts on the question of rates. Working from the premise that gas and electric utilities were natural monopolies, nearly all commissions inaugurated a complex administrative system that came to be called rate-of-return regulation. This system, over the next several decades, dominated the entire state regulatory process, and the details of that evolution confirmed the earlier prophecies of Charles Francis Adams. As Adams had feared would happen, excessive involvement in the rate question resulted in much controversy and red tape and tended to handicap more important efforts to promote rational adjustment of the general relationship between utility corporations and the public.[5]

Many of the problems of rate-of-return regulation began in a controversial Supreme Court decision in 1898. In this important case, *Smyth v. Ames*, the Union Pacific Railroad challenged a Nebraska statute that specified the maximum rates the corporation could charge for hauling freight. The court ruled that the railroad was entitled to a "reasonable" return on "the fair value of the property being used for the convenience of the public." In setting forth this fair-value doctrine, the court inadvertently started a definitional controversy that for many years afterward plagued accountants, lawyers, utility managers, and regulatory commissioners. At one time or another, particular state and federal commissions, and the Supreme Court itself, have included in their definitions of fair value such considerations as the original cost of the property less depreciation (a view favoring the consumer in periods of inflation); the cost of reproducing the same property (a view favoring corporations in periods of inflation); the market value of the corporation's stocks and bonds; and assorted other factors. No authoritative view has ever met with unanimous acceptance by courts and commissions. Nor was definitional uncertainty the only problem. At bottom, the entire concept of "fair return on fair value" is a cost-plus formula, vulnerable to all the difficulties associated with cost-plus ways of thinking about how

a business should be conducted. Some of these problems appeared early in the history of utility regulation, and several of them still trouble state and federal commissions. Such difficulties include:

1. Temptation to pad the "rate base" (the rate base is equivalent to the value of the property being used to produce the service; it provides the arithmetic denominator for the percentage "rate of return").
2. Creation of a strong incentive to emphasize capital equipment rather than labor, whenever the one can be substituted for the other (equipment goes into the rate base and labor does not).
3. Disincentives to cut operating costs—and therefore to promote more efficient operation.
4. The problem of "boundaries," in which one part of a company is a regulated natural monopoly but another part of the same company is not.
5. "Regulatory lag," in which the drawn-out proceedings of the commission consume so much time that the situation being ruled on becomes obsolete and irrelevant to future revenue requirements.[6]

In practice, these problems promoted the rise of ingenious accounting methods by corporations, all calculated to maximize revenues in the face of regulatory limits on percentage rates of return. At worst, they made the process of rate regulation a ritualistic charade, played out in the form of full-scale "rate cases" as part of the routine operations of commissions. Such cases, conducted under elaborate procedural rules, often turned into extremely laborious hearings dominated by lawyers and engineers, incomprehensible to ordinary citizens.[7]

The Rise of Federal Regulation

The movement from state to federal jurisdiction did not occur across the entire spectrum of regulated industries. In many industries, state regulation has continued, with no abatement of importance, down to the present day. Most public utilities, for example, continue to be regulated primarily by state commissions. But for industries in which single companies could grow so large that they spread across several states, thereby making effective supervision

l. Some of these problems ap-
tility regulation, and several of
al commissions. Such difficulties

ase" (the rate base is equivalent
ing used to produce the service;
enominator for the percentage

to emphasize capital equipment
e one can be substituted for the
e rate base and labor does not).
costs—and therefore to promote

in which one part of a company
y but another part of the same

e drawn-out proceedings of the
time that the situation being
irrelevant to future revenue re-

noted the rise of ingenious ac-
all calculated to maximize rev-
ts on percentage rates of return.
rate regulation a ritualistic cha-
-scale "rate cases" as part of the
. Such cases, conducted under
urned into extremely laborious
d engineers, incomprehensible

eral jurisdiction did not occur
ated industries. In many indus-
l, with no abatement of impor-
st public utilities, for example,
by state commissions. But for
s could grow so large that they
y making effective supervision

by any one state commission impracticable, state regulation yielded to federal. The railroad industry set the pattern here. During the latter part of the nineteenth century, as the "railroad problem" became national in scope, its dimensions also became much more complex. In the East, the major regulatory issue often had to do with excessive competition: too many railroads handled too little freight. In the West and South, by contrast, the problem usually took the form of uncontrolled monopoly, as whole communities came to depend for their very existence on single railroad corporations. In all regions of the country, railroads themselves often exhibited tendencies toward questionable financial practices and discriminatory pricing. Such pricing, in which different classes of shippers were charged different rates for the same type of service, might flow logically from the economics of railroading, but it often violated deep-seated popular conceptions of fairness. And underlying the entire national railroad problem was the political incongruity of a democratic society held in apparent subordination to one of its key industries.[8]

Ultimately, railroads grew so powerful and so vital to the national economy that continued reliance on state regulation alone became futile. Federal entry into the picture was precipitated by an important Supreme Court case: *Wabash, St. Louis and Pacific Railroad Company v. Illinois* (1886). In its decision, the court ruled that commerce originating or ending outside the boundaries of a state could not be regulated by that state, even though the federal government provided no alternative means of regulation. This doctrine partially contravened the court's prior ruling in the Granger cases, and the new *Wabash* rule made some final resort to federal regulation all but certain. In 1883, Senator Shelby M. Cullom, a Republican from Illinois, had introduced a bill providing for a federal commission that would administer a set of general and flexible guidelines for the governance of the railroad industry. The *Wabash* decision attracted majority support to Cullom's bill, and in 1887 the Interstate Commerce Commission was born. By that time, a national consensus had developed. Merchants, farmers, politicians, and even many railroad managers were now convinced, after years of trial and error with other methods, that serious federal railroad regulation must be inaugurated.[9]

From our own perspective a century later, the greatest significance of the 1887 Act to Regulate Commerce lies in its creation of

the prototypical federal regulatory agency. Most of the later federal commissions were patterned on the Interstate Commerce Commission, in appointment and tenure of members and in relationships with the existing branches of government—legislative, executive, and judicial. In fact, one measure of the success that the ICC was perceived as having in its first fifty years was its imitation in the creation of later agencies. Even so, the ICC began life with several different missions, not all of which easily accorded with each other or with the economics of railroading. The act of 1887 forbade rebating, pooling, and—with some important exceptions—rate discrimination between long-haul and short-haul traffic. The statute prescribed that rates be "just and reasonable," and it provided a new arena—the ICC—in which determinations of reasonableness could occur. The five commissioners were to be appointed by the president and confirmed by the senate, for staggered terms of six years. No more than three of the five could come from any one political party, and each commissioner was to receive an annual salary of $7,500—a very large sum for the time, greater than the salary of all federal judges except those on the Supreme Court.[10]

After an auspicious beginning under its distinguished first chairman, Thomas M. Cooley, the ICC encountered severe difficulties with the courts. Here too it set the pattern for future regulatory agencies. For the federal judiciary, with the Supreme Court in the lead, sharply restricted the powers of the new commission and reduced it, by the late 1890s, to a mere collector of data. Over the next several decades, a gradual process of adjustment occurred, involving the courts, the commission, and, most important, the Congress. Responding to continual problems within the railroad industry, as well as to jurisdictional squabbles between the ICC and the courts, the national legislature steadily added to the commission's authority. The highlights in this long history included: (1) the Elkins Act of 1903, which gave teeth to the prohibition against rebating contained in the original Act to Regulate Commerce of 1887; (2) the Hepburn Act of 1906, which empowered the commission to fix maximum railroad rates, shifted the burden of proof in rate proceedings from the commission to the railroads, and made ICC decisions effective as soon as they were reached; (3) the Mann-Elkins Act of 1910, which broadened the ratemaking authority of the commission, reinforced the long-

y agency. Most of the later fed-
on the Interstate Commerce
tenure of members and in rela-
ies of government—legislative,
measure of the success that the
first fifty years was its imitation
ven so, the ICC began life with
of which easily accorded with
of railroading. The act of 1887
-with some important excep-
n long-haul and short-haul traf-
es be "just and reasonable," and
C—in which determinations of
five commissioners were to be
nfirmed by the senate, for stag-
nan three of the five could come
each commissioner was to re-
-a very large sum for the time,
eral judges except those on the

under its distinguished first
ICC encountered severe diffi-
set the pattern for future regu-
iciary, with the Supreme Court
powers of the new commission
o a mere collector of data. Over
ual process of adjustment oc-
mmission, and, most important,
inual problems within the rail-
ictional squabbles between the
egislature steadily added to the
lights in this long history in-
which gave teeth to the prohibi-
n the original Act to Regulate
burn Act of 1906, which em-
aximum railroad rates, shifted
ings from the commission to the
effective as soon as they were
of 1910, which broadened the
mission, reinforced the long-

haul/short-haul rule, and created the Court of Commerce, a short-lived experiment in specialized judicial review; (4) the Transportation (or Esch-Cummins) Act of 1920, which gave the commission power to set minimum railroad rates, to supervise the issuance of railroad securities, to approve the previously forbidden practice of pooling, and to begin comprehensive planning for a systematic national transportation network; (5) the Motor Carrier Act of 1935, which gave to the ICC regulatory authority over the important new interstate trucking industry; and (6) the Transportation Act of 1940, which added still another industry—domestic water carriers—to the roster of businesses regulated by the ICC.[11]

This litany of legislation underlines an extremely important principle: during the twentieth century—just as in Charles Francis Adams' time—the art of regulatory strategy remains in large measure a political art. In no meaningful sense did the creation of the Interstate Commerce Commission take railroad regulation "out of politics," as so many supporters of regulation thought should be done. It was true of course that Congress, by creating the ICC, relieved itself of a continuing need to attend to innumerable trivial matters peculiar to the railroad industry. Congress did not, however, permanently yield its powers to an "independent" regulatory commission. Instead, the congressional committees stayed in close touch with developments at the ICC. The commission itself often originated proposals for new legislation, and the agency took great pains to develop and nurture political support within Congress. Without this support, the ICC could not have done its job properly, especially since a jealous judiciary often threatened to undercut its authority.[12]

Throughout its history, the ICC has remained controversial. Sometimes, as during the Progressive Era, it became the target of railroad industry protests charging that it favored shippers at the expense of carriers. More often, the commission has been accused by shippers and consumer groups of trying to protect the carriers at the expense of the general public. Sometimes, critics with entirely different viewpoints have joined in blaming the ICC for the long-term decline in the nation's rail service. But the most damning line of criticism has rested on claims that the ICC has worked to impede economic efficiency. Indeed, from the historical record, it is clear that the agency's policies often prevented market forces from allocating traffic according to the individual modes of trans-

portation—whether rail, highway, or water—best suited for particular kinds of freight. The commission appears to have interfered with efficiency unintentionally, at first simply through ignorance, since in its early years the economics of transportation was not very well understood. Later on, most errors derived from the ICC's misguided attempts to be "fair" to all parties. The commission regulated transportation rates in such a way as to ensure that each industry—rail, highway, or water—received not those portions of the total traffic most appropriate to its routes, but instead a fair share. In its concern for fairness at the expense of economic efficiency, as in so many other respects, the ICC typified the practice of most regulatory agencies in American history.[13]

Eventually, the kinds of regulatory problems dealt with by state and federal railroad commissions expanded to include troubles rooted in other large enterprises. Although the profound economic movement that has become known as the "rise of big business" began with the railroads in the 1850s, it continued to move forward, in vastly expanded form, causing revolutions in manufacturing and distribution. These changes occurred between about 1880 and 1920. Prior to this period, no single manufacturing enterprise, indeed no entire manufacturing industry, had attained sufficient size to affect masses of people. Before the 1880s, even major factories customarily employed no more than a few hundred workers. The largest manufacturing companies were usually capitalized at less than one million dollars. Yet, within a single generation after 1880, all this changed. By 1890, each of several large railroads employed more than 100,000 workers. By 1900, John D. Rockefeller's Standard Oil Company had grown into a huge multinational corporation capitalized at $122 million. By 1904, James B. Duke's American Tobacco Company completed a series of mergers and internal expansions that took it to a capitalization of $500 million (up from $25 million in 1890). And in 1901, the creation of the United States Steel Corporation climaxed a $1.4 billion transaction. This sum, which far exceeded the imaginations of most contemporary citizens, became a symbol of the new giantism in the American economy.[14]

With the rise of big business, the term "private enterprise" acquired a different meaning. Whereas once it had signified liberty, freedom, and individualism, it now meant danger as well—the threat of giant corporations. Suddenly, big business seemed to

or water—best suited for par-
ion appears to have interfered
rst simply through ignorance,
ics of transportation was not
nost errors derived from the
r" to all parties. The commis-
n such a way as to ensure that
ıter—received not those por-
riate to its routes, but instead
ss at the expense of economic
ets, the ICC typified the prac-
ımerican history.[13]

problems dealt with by state
xpanded to include troubles
nough the profound economic
as the "rise of big business"
850s, it continued to move
ısing revolutions in manufac-
ges occurred between about
no single manufacturing en-
uring industry, had attained
ople. Before the 1880s, even
no more than a few hundred
ompanies were usually capi-
s. Yet, within a single genera-
1890, each of several large
0 workers. By 1900, John D.
had grown into a huge mul-
$122 million. By 1904, James
pany completed a series of
took it to a capitalization of
1890). And in 1901, the crea-
ation climaxed a $1.4 billion
ceeded the imaginations of
symbol of the new giantism

erm "private enterprise" ac-
once it had signified liberty,
meant danger as well—the
ly, big business seemed to

menace America. Large corporations represented that same centralized power against which the founding fathers had fought their revolution. Perhaps inevitably, American big business evoked a powerful regulatory response.

Overcapacity and the Rise of the Trusts

In the popular language of the late nineteenth and early twentieth century, such companies as Standard Oil, American Tobacco, and United States Steel became known as "trusts." Although most combinations did not take the precise legal form of a trust (in which trustees for several different companies hold the common stock of those companies and thereby exercise control over them),[1] all such corporations did share the characteristics of centralized management and very large size. The process by which companies in certain industries did or did not grow to great size is not fully understood, even today. But it is much better understood now than it was in the early twentieth century. Within the past few years, several streams of important research in economics, history, and business administration have converged, to make possible a much fuller understanding of the rise of big business and of attempts to regulate it. This story is not a simple one to tell but, because it is one so rich in importance, paradox, and surprise, it well repays the necessary effort.

The trust movement—that is, the powerful tendency of businessmen to combine with their competitors in associations or mergers—grew out of a particular problem of industrialization. This was the problem of periodic industrial overcapacity, tied to the boom-and-bust cycles of the late nineteenth century. Just as worldwide overcapacity lies behind the periodic "sickness" of such contemporary industries as steel, fibers, footwear, and automobiles, so in the late nineteenth century industrial overcapacity caused serious disruptions in the economies of all industrialized nations.

The underlying cause of overcapacity was the industrial revolution, which initially took the form of a revolution in production and productivity. In the first phase of the industrial revolution, progress in the technology of production far outran similar developments in distribution, marketing, and consumer purchasing power. Cyrus McCormick's reaper, invented in the 1830s, laid the

foundation for a national and international system of commercial grain agriculture; but the new system of distribution did not develop until a national railroad network emerged for marketing the grain (there were 23 miles of track in America in 1830 and 208,152 by 1890). Changes occurred at uneven rates, and these different paces of change created serious periodic imbalances between nations' capacities to produce and their abilities to consume. In some respects, of course, the differences were natural and inevitable. The rise of the department store (1870s) could hardly have preceded the invention of the sewing machine (1844). And the vast array of consumer goods that poured forth from the production of thousands of American factories could not flow to their users until a corresponding revolution had occurred in packaging and wholesaling.[15]

Machines that could reap grain, stitch cloth, and wrap consumer goods were but three examples of a profound worldwide revolution in productivity. This revolution substituted machine tools for human craftsmen, interchangeable parts for hand-tooled components, and the energy of coal for those of wood, water, and animals. The rising productivity, in turn, often brought overcapacity. The response, among businessmen in every industrial nation, was to combine with one another in schemes designed to limit the total output of their plants, maintain the price levels of their goods, and discourage the entry of new companies into their lines of business.

Such tendencies are discoverable in all industries, all countries, and all times. Adam Smith's famous description of the phenomenon was already trite when he made it in 1776: "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."[16] But the tendencies Smith observed seemed mild indeed when compared with the manic compulsions stimulated afterward by the revolution in productivity, which made the potential rewards of industrial success far greater than anything possible in Smith's era. It had the same magnifying effect on the potential cost of failure: the immense capital investment represented by a large modern factory or string of factories raised the penalty for failure beyond anything Smith could have contemplated. The collapse of, say, a large steel company would cost tens of millions of dollars in idled physical plant

national system of commercial
tem of distribution did not de-
ork emerged for marketing the
n America in 1830 and 208,152
even rates, and these different
riodic imbalances between na-
ir abilities to consume. In some
s were natural and inevitable.
1870s) could hardly have pre-
machine (1844). And the vast
d forth from the production of
uld not flow to their users until
urred in packaging and whole-

, stitch cloth, and wrap con-
ples of a profound worldwide
evolution substituted machine
angeable parts for hand-tooled
for those of wood, water, and
y, in turn, often brought
g businessmen in every indus-
ne another in schemes designed
ants, maintain the price levels
entry of new companies into

in all industries, all countries,
s description of the phenome-
de it in 1776: "People of the
even for merriment and diver-
conspiracy against the public,
es."[16] But the tendencies Smith
en compared with the manic
y the revolution in productiv-
vards of industrial success far
Smith's era. It had the same
cost of failure: the immense
large modern factory or string
ailure beyond anything Smith
apse of, say, a large steel com-
dollars in idled physical plant

and would throw thousands of employees out of work. Thus industrialists felt a powerful urge to maintain a market for their products—if necessary by temporarily selling below costs, if possible by cooperating with each other for the mutual protection of capital.

Among European industrialists, this inclination to combine in self-defense against overcapacity led to very different results from those in the United States. Although the response was not precisely the same in every European country, governments there generally accepted business combinations as legitimate. European price and production cartels usually enjoyed official sanction, so that the legal machinery of the state could be used to enforce contractual articles of cartelization against rebellious price cutters. In Europe, therefore, with its large public bureaucracies long established, industrial overcapacity became simply another problem for a mature political state to manage. The official sanctioning of cartels provided a convenient way for national governments to prevent direct damage to individual firms and thus to soften and stabilize the otherwise wrenching process of industrialization.[17]

In Europe, typical political battles during the late nineteenth century pitted a fairly united business community against a powerful labor movement. In America, by contrast, the most conspicuous political warfare of that time matched *one group of businessmen against another:* carriers versus shippers, commodity farmers versus mortgage bankers, small wholesalers and shopkeepers against large firms whose marketing divisions were eliminating the traditional roles of local jobbers and retailers. The small size of the United States government, as compared to its European counterparts, meant that no adequate public bureaucracy existed to manage such industrial conflicts. (In 1871, on the eve of the trust movement, only 51,020 civilians worked for the federal government, of whom 36,696 were postal employees. The remaining 14,424 constituted the national government for a country whose population exceeded 40 million.) Under American conditions, the emerging problem of industrial overcapacity, compounded by the boom-and-bust business cycle, moved immediately into the realm of public controversy. To a degree often underestimated by scholars, overcapacity influenced nearly every major economic issue of the period: not only the trust question, but also the perennial and divisive battles over the protective tariff, the railroad rate prob-

lem, and the imperial quest for foreign markets to absorb surplus production.[18]

Even so, the initial response to industrial overcapacity took much the same form in America as in Europe. American businessmen, like their counterparts abroad, energetically combined with each other in loose cartels designed to limit production, maintain prices, and divide markets. These cartels were intended to protect all member companies from the fearful consequences of collective overcapacity. By preventing bankruptcies and by making mergers unnecessary, the cartels could preserve the individual identity of all cooperating companies. During the 1870s and 1880s, the number of formal and informal associations erected for this purpose within the United States numbered in the thousands. Seldom, however, did these early cartels accomplish their purpose because almost every one of them encountered legal obstacles. The national culture was so opposed to "monopoly" and "restraint of trade" that American courts refused to enforce cartel arrangements against recalcitrant members.[19] Thus, unlike their European counterparts, American businessmen participating in cartels were left free to cut their prices in violation of the cartel agreements or to sell their products outside geographic areas determined by the cartel. In 1890, passage of the Sherman Antitrust Act served to formalize the common law's hostility to cartels, as well as to compel the Department of Justice to become an active opponent of cartel-like associations. The Sherman Act soon began to have an enormous, if paradoxical, impact.

The New Vocabulary

To understand that impact, it is necessary to use certain terms of modern economic analysis. Before proceeding further with the story of the trust movement, it will be appropriate to pause briefly and investigate the meaning and application of this new economic vocabulary. It is important to remember that most of these terms did not originate with the trust movement. Instead, they have evolved over the years since passage of the Sherman Act, as part of an enormous effort of research in the fields of economics and business administration.

In applying such terms in historical analysis, we are something like modern astronomers studying the pre-Copernican theories of

ROADS TO TRUSTS

oreign markets to absorb surplus

to industrial overcapacity took
a as in Europe. American busi-
abroad, energetically combined
s designed to limit production,
ets. These cartels were intended
from the fearful consequences of
enting bankruptcies and by mak-
els could preserve the individual
panies. During the 1870s and
informal associations erected for
ates numbered in the thousands.
cartels accomplish their purpose
em encountered legal obstacles.
ed to "monopoly" and "restraint
efused to enforce cartel arrange-
ers.[19] Thus, unlike their Euro-
inessmen participating in cartels
in violation of the cartel agree-
outside geographic areas deter-
age of the Sherman Antitrust Act
law's hostility to cartels, as well
ustice to become an active oppo-
he Sherman Act soon began to
impact.

ecessary to use certain terms of
re proceeding further with the
l be appropriate to pause briefly
pplication of this new economic
ember that most of these terms
movement. Instead, they have
ge of the Sherman Act, as part of
he fields of economics and busi-

rical analysis, we are something
the pre-Copernican theories of

Aristotle or Ptolemy. That is, we are testing the assumptions and insights of an earlier generation of observers through the use of methods developed in a subsequent scientific revolution to which they had no access. Although it may seem unfair to judge early twentieth-century observers by standards of economic reasoning they themselves did not fully comprehend, that price must be paid in return for our own understanding of the trust question. Even so, we should remember, in our effort to minimize the unfairness and achieve an appreciation of the problems faced by this earlier generation, that the movements and trends best described by the new vocabulary were not all equally prominent at the same time. At any given moment, some dominated the scene while others were latent. It is only through the double lenses of our own retrospect *and* the contemporaneous view that we are able to reach a mature historical evaluation of the elements involved in the trust question.

For us, the vocabulary essential to an understanding of that question has a minimum of seven relevant terms:

1. *Productive efficiency.* This term refers to the amount of work accomplished in relation to the effort expended. The underlying principle is common in mechanical engineering, as in the comparison that steam engines are more efficient than internal-combustion engines because they deliver more work per unit of fuel consumed. In economics, the meaning of productive efficiency is similar, but much broader. It includes anything that businesses can do to reduce the cost or improve the quality of their products, without any offsetting effect. Such changes may be organizational, technical, or of some other kind. One example of productive efficiency is machine mass production on assembly lines.[20]

2. *Allocative efficiency.* This term relates to the way in which an entire economy operates. In an allocatively efficient economy, no further rearrangement of prices, outputs, or distribution of goods and services can possibly make one consumer better off in terms of his own desires without making another worse off in terms of his. With the highest allocative efficiency, prices are set close to the marginal costs of production—that is, the cost of the last unit produced. Maximum allocative efficiency is equivalent to maximum "consumer welfare"; and in economic theory, any policy that promotes allocative efficiency and consumer welfare car-

ries a very strong favorable presumption. Similarly, any policy tending to diminish them should, if such a policy is to be justified, bring extremely beneficial offsetting effects of a noneconomic nature. Of course, allocative efficiency and consumer welfare apply chiefly to "economic man" and say little—at least directly—about aesthetic values, income distribution, ideological preferences, or any number of other legitimate human concerns. Even so, the concepts have important implications for practical politics as well as for theoretical economics: in modern times, high government regard for allocative efficiency and consumer welfare has been closely correlated with democracy; and low regard, with authoritarianism.[21]

3. *Scale economies.* There are two types of scale economies; in both, the larger the operation, the greater the productive efficiency. The first type is a *physical* scale economy, related to the amount or size of equipment used in a business operation. In the case of railroads, as Adams saw so clearly, the larger the number of cars in a given freight train, the lower the ton-per-mile cost of moving a given item of cargo. In some industries, such as oil refining (and others involving the flow of liquids or gases), economies of scale also derive from certain advantages in geometry. The volume of a pipe, for example, increases with the square of its radius. If the radius is doubled, the volume is quadrupled. Thus the bigger the pipe, the greater the productive efficiency, all other things being equal. The second type of scale economy is not physical but *organizational,* deriving from what Adam Smith called "the division of labor." It focuses on method rather than on equipment. The more particularized an operation is, the more efficiently it can be done by a specialized tool, machine, or skilled worker. The larger the undertaking, the more numerous and specialized its constituent labor force and machinery can be and, therefore, the greater the potential productive efficiency.

4. *Vertical integration.* This term refers to the gathering of many different business functions within a single firm. A company that conducts all of its business operations itself, ranging from the derivation of its raw materials to the selling of its finished manufactured products at retail, is said to have achieved complete vertical integration. A company's movement toward vertical integration almost always takes the form of incremental decisions to extend the functions performed by the existing business. An oil-refining

esumption. Similarly, any policy
, if such a policy is to be justified,
ting effects of a noneconomic na-
ncy and consumer welfare apply
ay little—at least directly—about
ution, ideological preferences, or
human concerns. Even so, the
tions for practical politics as well
modern times, high government
and consumer welfare has been
cy; and low regard, with authori-

two types of scale economies; in
the greater the productive effi-
al scale economy, related to the
d in a business operation. In the
clearly, the larger the number of
lower the ton-per-mile cost of
some industries, such as oil refin-
of liquids or gases), economies of
ntages in geometry. The volume
ith the square of its radius. If the
quadrupled. Thus the bigger the
efficiency, all other things being
conomy is not physical but *orga-*
am Smith called "the division of
r than on equipment. The more
more efficiently it can be done
r skilled worker. The larger the
and specialized its constituent
and, therefore, the greater the

refers to the gathering of many
a single firm. A company that
ons itself, ranging from the deri-
selling of its finished manufac-
ave achieved complete vertical
ent toward vertical integration
ncremental decisions to extend
xisting business. An oil-refining

company, for example, might decide to purchase tank trunks for transporting its gasoline to market; later it also decides to build service stations from which the gasoline will be sold at retail. Both decisions represent vertical integration.

A fully integrated oil company would own exploration and production equipment for the finding and pumping of crude oil; tanker fleets and pipelines to transport it; refineries to turn it into gasoline, kerosene, and heating oil; tank trucks from which to wholesale these products; and service stations for sales to retail customers. The company would pursue this strategy of vertical integration in order to match its organizational scale economies with its physical ones. If the company did not integrate vertically, then the scale-economy potential in refining might be interrupted randomly by a host of unforeseen events: an abrupt stoppage of crude-oil supply, a sharp rise in its market price, the unavailability of transportation at a critical time, a collapse of the distribution system, or some other sudden bottleneck in the movement of the product from one stage to another. With vertical integration, on the other hand, every part of the operation (exploration, production, transportation, refining, wholesaling, retailing) would be matched in size to every other part, and the company's own managers would supervise the flow of oil through the system. A continuous movement of the product at constant speed and at declining costs of production should be assured.

Even so, vertical integration is not necessarily good for all companies. It seems to work well as a business strategy only for firms in those industries that possess a significant scale economy at some stage in the production or marketing process. A business integrates vertically for the purpose of ensuring that its advantage from the scale economy is maximized and not interrupted. But vertical integration can also become dangerous to the survival of a firm, since it has the disadvantage of putting all a company's eggs in one very large basket, and then risking that entire basket to the contingency of a declining market for its products.

5. *Horizontal combination.* This term signifies agreements among producers of a certain item: either to limit production of the item or to maintain its price level, or to do both. (The word "horizontal" refers to all participants' being in the same line of business, producing the same item; as opposed to "vertical," which refers to different business functions—production, whole-

saling, retailing, and so forth.) In the late nineteenth century, when manufacturers began to make use of the new technology and new sources of energy in their factories, shops, and refineries, they discovered to their astonishment that production had suddenly become the least of their worries. They could produce huge quantities of oil, steel, sewing machines, cigarettes, and hundreds of other items. But could they also sell this enormous new production? And, if so, could they still keep their prices up, even as vast quantities of their goods kept flowing onto the market? Many producers found an apparently affirmative answer to these questions in mutual cooperation—that is, one or another type of horizontal combination. On the surface, these associations seemed attractive. Yet horizontal combinations had two enormous flaws. Whereas they might help participating companies, they would hurt nonparticipating ones; even worse, they would reduce consumer welfare by creating artificial scarcities. Total output of specific products would be restricted, prices kept up, and allocative efficiency diminished.[22]

In practice, a particular horizontal combination might take one of several different forms. At its simplest, it consisted of an informal agreement among, say, all the coal dealers in a single city, who made a mutual promise not to sell a ton of coal below a certain minimum price. A more complex horizontal combination appeared in America as a European-style cartel—ten or twelve steel companies sharing output quotas and price agreements. The ultimate horizontal combination proved to be huge firms such as American Tobacco Company, which represented the final stage in a series of mergers that brought all important cigarette companies into one industry giant. This so-called tight form, involving actual mergers of corporate entities, became known as "horizontal integration." In fact, all horizontal integration involved horizontal combination, but not vice versa. Most horizontal combinations were not tight mergers, but loose associations of corporations that still retained their individual legal identities. The difference between loose horizontal combinations and tight ones turned out to be extraordinarily important.[23]

6. *Center firms.* This term denotes those new types of giant companies—such as Standard Oil, American Tobacco, United States Steel, and a few hundred others—that grew to large size *and remained large.*[24] These center firms tended to have a set of

In the late nineteenth century,
ake use of the new technology and
actories, shops, and refineries, they
nt that production had suddenly
s. They could produce huge quan-
ines, cigarettes, and hundreds of
o sell this enormous new produc-
keep their prices up, even as vast
wing onto the market? Many pro-
rmative answer to these questions
one or another type of horizontal
ese associations seemed attractive.
d two enormous flaws. Whereas
ompanies, they would hurt non-
hey would reduce consumer wel-
cities. Total output of specific
rices kept up, and allocative effi-

ntal combination might take one
simplest, it consisted of an infor-
the coal dealers in a single city,
to sell a ton of coal below a cer-
nplex horizontal combination ap-
-style cartel—ten or twelve steel
and price agreements. The ulti-
oved to be huge firms such as
ich represented the final stage in
ll important cigarette companies
alled tight form, involving actual
came known as "horizontal inte-
integration involved horizontal
. Most horizontal combinations
associations of corporations that
al identities. The difference be-
ons and tight ones turned out to

notes those new types of giant
il, American Tobacco, United
others—that grew to large size
er firms tended to have a set of

common characteristics. All were capital-intensive, requiring very large outlays of investment. All were technologically advanced, in the sense that their production facilities were dominated either by continuous process production (Standard Oil, American Sugar Refining, United States Steel) or by some combination of large-batch production with machine mass production (Singer Sewing Machine, American Tobacco, Quaker Oats, Pillsbury Flour, International Harvester). Consequently, all center firms enjoyed some significant scale economy in the production or packaging process, and all were vertically integrated. Planning for the future, center firms also adopted an unusually long-range perspective. Because of the enormous investments required, survival was a far more important goal than was short-run profit. Accordingly, many center firms planned their business strategies on the basis of five-year horizons. In addition, many of them engaged in organized research and the development of new products. To manage their many internal functions, they began to organize complex managerial hierarchies.[25]

7. *Peripheral firms.* This term refers to companies that were small, labor-intensive, managerially thin, and bereft of scale economies. Concerned for survival, they looked more to this year's profits than to five-year plans. Peripheral firms, in other words, represented the opposite kind of company from center firms. The distinction between peripheral and center is not a perfect one, but the two terms may be considered as the polar ends of a spectrum: very pronounced differences are evident from one end of the spectrum to the other, but the differences become less dramatic toward the middle. The major point, however, is that center firms and peripheral firms represent two fundamentally different types of businesses.[26]

To be "peripheral" does not necessarily mean to be unimportant. Many peripheral firms are major sources of employment and thus vital to the national economy. Such industries as textiles, furniture, clothing, food service, building materials, hotels, and automobile repair are characterized by a large number of relatively small firms. These peripheral firms compete with each other very much in the fashion of Adam Smith's classical model, which remains the model of competition described in economics textbooks. Companies in these industries, along with numerous other kinds of peripherals—ranging from restaurants and family retail outlets to

specialized parts manufacturers and subcontractors for the center firms—form the backbone of small business in all capitalist economies. These industries, like the companies in them, are also called peripheral, to distinguish them from such center industries as oil, steel, automobile manufacturing, and electrical machinery.[27]

The New Research

As we have noted, the modern discipline of economics has helped us to understand the rise of big business, especially by supplying some of the necessary vocabulary. But economic research has also displayed some important limitations. Despite advances in the subdiscipline of industrial organization, the economic theory of oligopoly is not yet mature. Mainstream theory still tends to assign too little significance to the gains in productive efficiency arising from changes in the forms of business organizations, including vertical integration.[28]

One very important contribution to research on center firms has come not from economics but from history, especially from the world's leading historian of business, Alfred D. Chandler, Jr. In a series of articles and books written from the 1950s to the 1980s, Chandler and his students have systematically examined the evolution of the business system in the United States and in other major capitalist economies. From this study of more than one thousand center firms has emerged an enormous body of information about their evolution, structure, business strategies, and methods of internal governance. The implications of Chandler's work are far-reaching, but its principal message is simple and forthright. The economist Robert Heilbroner described Chandler's Pulitizer Prize book, *The Visible Hand,* as "a major contribution to economics, as well as to 'business history,' because it provides powerful insights into the ways in which the imperatives of capitalism shaped at least one aspect of the business world—*its tendency to grow into giant companies in some industries but not in others.*"[29]

For a better understanding of the trust issue, three of Chandler's propositions are especially helpful. First, the industry structure characteristic of the American center economy evolved largely during the forty-year period between 1880 and 1920. The shakeout of center firms and peripheral firms into a more or less stable

and subcontractors for the center
all business in all capitalist econo-
companies in them, are also called
from such center industries as oil,
, and electrical machinery.[27]

iscipline of economics has helped
business, especially by supplying
y. But economic research has also
tations. Despite advances in the
zation, the economic theory of ol-
stream theory still tends to assign
s in productive efficiency arising
usiness organizations, including

on to research on center firms has
rom history, especially from the
ess, Alfred D. Chandler, Jr. In a
en from the 1950s to the 1980s,
ystematically examined the evo-
the United States and in other
m this study of more than one
d an enormous body of informa-
ucture, business strategies, and
The implications of Chandler's
rincipal message is simple and
Heilbroner described Chandler's
Hand, as "a major contribution to
ss history,' because it provides
n which the imperatives of capi-
of the business world—*its ten-
ies in some industries but not in*

e trust issue, three of Chandler's
ul. First, the industry structure
enter economy evolved largely
veen 1880 and 1920. The shake-
firms into a more or less stable

configuration did not end abruptly in 1920, but the major phase of this evolution was complete by that time. Evidence of a remarkable industrial stability may be found by comparing a list of the 200 largest U.S. manufacturing companies (measured in assets) in 1917 with a similar list for 1973:

a. In 1917, 22 of the largest 200 companies were in the petroleum business. In 1973, 22 of the largest 200 were still in petroleum. For the most part, they were the same 22.
b. In 1917, 5 of the largest 200 companies were in the rubber business. In 1973, 5 of the largest 200 were still in rubber. Four of these five were the same (Goodyear, Goodrich, Firestone, Uniroyal). The fifth, Fisk, merged with Uniroyal—then called United States Rubber—in 1939.
c. In 1917, 20 of the largest 200 companies were in the machinery business. In 1973, 18 of the largest 200 were still in machinery. For the most part, they were the same companies.
d. In 1917, 30 of the largest 200 companies were in the food products business. In 1973, 22 of the largest 200 were still in food products, and several other food companies remained in the top 200 as parts of conglomerates.
e. In 1917, 26 of the largest 200 companies were in the transportation equipment business. In 1973, 20 were still in transportation equipment (down, partly as a result of mergers, from 26 in 1948).[30]

These numbers suggest that the companies which succeeded initially in center industries later enjoyed a considerable advantage over firms entering the same industries. Their advantages included scale economies and lower unit costs achieved through long production runs. If the companies also rationalized their production process by concentrating operations in their lowest-cost plants—and most did just this—then their advantages increased even more. Finally, when the companies proceeded to integrate vertically, thereby protecting the advantages they already held in production, they became inordinately difficult for competitors to dislodge. Considering the numerical evidence, which shows that the companies maintained their position over a very long period, the fact of their continuing dominance becomes a historical phenomenon of enormous help in understanding the trust question.

A second proposition is that companies in every manufacturing

industry attempted the same sort of growth strategy described above. That is, the movement toward horizontal integration was universally attractive. Firms in peripheral industries tried it as often as did those in center industries. Only the outcome proved different: unlike center firms, peripherals almost never succeeded. (*None* of the largest 200 companies in either 1917 or 1973, for example, was in furniture, printing, or similar peripheral industries.) Thus the difference between center and peripheral must be regarded as crucial to understanding the true nature of the trust question.[31]

A third proposition has to do with the evolution of business organizations in other countries. After having established that certain patterns characterized the structure of American industry in the twentieth century, Chandler went on to discover that the very same patterns obtain, with surprisingly few modifications, in other major market economies. These similarities appear despite different markets, political cultures, and sources of supply. The numbers cited above concerning the stable makeup of the largest 200 American manufacturing companies resemble comparable sets of numbers compiled for the same period in Germany, France, Britain, and Japan. Only the pace and timing of appearance vary, and even these do not differ radically. Furthermore, the pattern of industries themselves tends to be similar to that in the United States, as the following list shows:

	U.S. (192 companies)	*Abroad* (187 companies)
Center industries		
Transportation equipment	22	23
Electrical machinery	20	25
Stone, clay, and glass	7	8
Tobacco	3	4
Chemicals	24	28
Rubber	5	5
Petroleum	14	12
Peripheral industries		
Furniture	0	0
Printing	0	0

This list includes all manufacturing companies (worldwide) that as of 1973 had at least 20,000 employees. Approximately one half of the total of 379 were American firms, and half had their headquarters elsewhere.[32]

ort of growth strategy described
ard horizontal integration was uni-
ipheral industries tried it as often
. Only the outcome proved differ-
als almost never succeeded. (*None*
either 1917 or 1973, for example,
milar peripheral industries.) Thus
and peripheral must be regarded
rue nature of the trust question.[31]
with the evolution of business or-
After having established that cer-
structure of American industry in
went on to discover that the very
isingly few modifications, in other
similarities appear despite differ-
and sources of supply. The num-
stable makeup of the largest 200
nies resemble comparable sets of
period in Germany, France, Brit-
d timing of appearance vary, and
. Furthermore, the pattern of in-
milar to that in the United States,

U.S. (92 companies)	*Abroad* (187 companies)
22	23
20	25
7	8
3	4
24	28
5	5
14	12
0	0
0	0

g companies (worldwide) that as
yees. Approximately one half of
ms, and half had their headquar-

The striking similarities between industrial experience in the United States and in other market economies suggest strongly that the economic and technological characteristics of certain industries encourage them to assume either a center or peripheral configuration and to maintain that configuration over a long period of time. These characteristics now seem much more important than do differences in legal systems or national cultures; in fact they appear to determine the relative size and organizational structure of firms within the industries represented. This conclusion is of great importance to anyone interested in assessing the historical record of big business in the United States, a record that includes all attempts to regulate monopolistic "trusts."

The Old Assumptions

In the minds of many members of the generation that came to maturity during the 1880s and 1890s, the huge new companies we now call center firms seemed somehow unnatural. This generation, like all its predecessors, had grown up in an era occupied almost exclusively (the railroads being the sole exception) by what I have identified as peripheral firms. With their meager resources and lack of political power, such firms posed no threat whatever to the country's democratic heritage. The sudden appearance of center firms, on the other hand, brought novel and alarming practices: degradation of human labor in such industries as steel (where immigrant laborers put in 72-hour workweeks); unscrupulous manipulation of stocks and bonds in industries such as railroads and utilities (where "frenzied finance" seemed common); and abrupt losses of community control, not only over industries but over individual center firms, whose resources often dwarfed those of city and even state governments.

In the early years of industrialization, the trusts seemed to be mysterious mutations, the consequences of some evil tampering with the natural order of things. They were not merely economic freaks but also sinister new political forces—powers that had to be opposed in the name of American democracy. Such a conclusion seemed especially compelling once the ruthless, inhuman methods used in the formation of Standard Oil, American Tobacco, and other huge companies had reached public attention. Unlike loose cartels, which protected the interests of all participating com-

panies, these tightly organized, horizontally and vertically integrated giants swallowed up numerous competitors and pushed aside traditional wholesalers and retailers. If the low prices of center firms brought unprecedented benefits to masses of consumers, then those same prices also injured large numbers of competing small producers, wholesalers, and retailers.

Yet the central assumption shared by most contemporary critics was simply that trusts were *unnatural,* the bastard offsprings of unscrupulous promoters. Without the benefit of experience and of a vocabulary that could clarify important distinctions in the business revolution—center firms compared with peripheral, productive efficiency compared with allocative, vertical integration compared with horizontal, and so on—those observers had only their personal sensibilities and traditional political ideologies to guide them. Both their personal and political values concerning the nature of liberty, the meaning of opportunity, and the promise of America seemed to be seriously threatened by the new monster trusts.

By about the middle of the 1880s, public sentiment in America had crystallized: something had to be done about trusts. Courts began to be more attentive to violations of common-law prohibitions against the collusive fixing of prices and planned limitations of output. In 1887, Congress created the Interstate Commerce Commission as a new means of dealing with the perennial railroad problem. Three years later, Congress passed the milestone Sherman Antitrust Act, which seemed clear, concise, and to the point. The Sherman Act declared illegal "every contract, combination in the form of trust or otherwise, or conspiracy, in restraint of trade or commerce." It defined as a criminal "every person who shall monopolize, or attempt to monopolize, or combine or conspire with any other person or persons, to monopolize any part of the trade or commerce among the several States, or with foreign nations."[33]

Despite the apparent clarity of the Sherman Act, however, its implementation raised severe difficulties. Unlike the Act to Regulate Commerce of 1887, the Sherman law set up no new regulatory commission or other machinery, but instead relied on the existing structure of government. Enforcement was vested in the federal Department of Justice, but the attorneys general of the 1890s showed little inclination to prosecute aggressively. When

horizontally and vertically inte-
merous competitors and pushed
retailers. If the low prices of cen-
l benefits to masses of consumers,
ired large numbers of competing
d retailers.

ered by most contemporary critics
natural, the bastard offsprings of
t the benefit of experience and of
mportant distinctions in the busi-
compared with peripheral, pro-
h allocative, vertical integration
so on—those observers had only
raditional political ideologies to
and political values concerning
g of opportunity, and the promise
y threatened by the new monster

80s, public sentiment in America
to be done about trusts. Courts
olations of common-law prohibi-
of prices and planned limitations
eated the Interstate Commerce
aling with the perennial railroad
gress passed the milestone Sher-
l clear, concise, and to the point.
"every contract, combination in
conspiracy, in restraint of trade
riminal "every person who shall
polize, or combine or conspire
, to monopolize any part of the
veral States, or with foreign na-

f the Sherman Act, however, its
iculties. Unlike the Act to Regu-
man law set up no new regula-
hinery, but instead relied on
nt. Enforcement was vested in the
ut the attorneys general of the
o prosecute aggressively. When

they did prosecute, their favored targets were not the feared center firms but local combinations of peripheral firms, whose price-fixing practices were easy to demonstrate. During the Sherman Act's first twenty years, by far the majority of prosecutions were aimed at loose cartels of small companies. The six or seven "big cases" of this period, however, did pertain to railroad combinations and to large trusts in sugar, beef, oil, and meatpacking. But in almost none of these cases did the court decison emerge in the form that antitrust advocates had hoped to see. The apparent clarity of the statutory language became lost in verbal uncertainties. Courts struggled with the meaning of such terms as "restraint of trade" and "monopolization"; and judges' opinions created a semantic confusion that persisted for years afterward. Soon the obvious ineffectiveness of early antitrust policy began to stimulate a new movement to do something more. Meanwhile, the severe economic depression of the 1890s added another note of urgency, and finally both the populists of that decade and the progressives of the early twentieth century made trust reform a pivotal and decidedly durable political issue. Indeed, the trust question dominated political discourse in America from the 1880s until the start of the First World War.[34]

CHAPTER 7

Kahn and the Economist's Hour

In *Nixon Agonistes,* his bestseller of 1970, the journalist Garry Wills argued that "America is distinguished by a 'market' mode of thought in all its public (and even private) life." Wills went on to identify four market systems, each with its reigning prophet: the economic market system (Adam Smith), the political (Woodrow Wilson), the moral (Ralph Waldo Emerson), and the intellectual (John Stuart Mill). These four systems, Wills wrote, combined to yield up the strange and anomalous figure of Richard M. Nixon.[1]

Nixon Agonistes now seems both profound and half-baked. More important, however, the book's popular success signaled an unmistakable restoration of "the market" to intellectual respectability. For the principles that Wills made acceptable to a popular audience also found an enthusiastic if less boisterous welcome in the halls of academe. There a group of innovative scholars had spent the preceding years sending a similar message to students and colleagues. In economics the "Chicago school" of free-market theorists gradually gained reinforcements from a corps of young monetarists who became impatient with the policy activism favored by Keynesians. The Chicago school exerted only moderate influence before 1970, although occasionally a member might score a popular hit, as Milton Friedman did with *Capitalism and Freedom* (1962). By and large, however, the Chicago economists hardly entered the consciousness of most Americans, who took in what they knew about economics from the witty John Kenneth Galbraith. Even in 1970, few people realized that time and circumstance were now on the side of the Chicagoans. But over the next dozen years, the Chicago school came to exercise great influ-

PTER 7

Economist's Hour

eller of 1970, the journalist Garry
s distinguished by a 'market' mode
even private) life." Wills went on
each with its reigning prophet: the
n Smith), the political (Woodrow
do Emerson), and the intellectual
ystems, Wills wrote, combined to
alous figure of Richard M. Nixon.[1]
both profound and half-baked.
ook's popular success signaled an
market" to intellectual respecta-
Vills made acceptable to a popular
stic if less boisterous welcome in
group of innovative scholars had
ing a similar message to students
"Chicago school" of free-market
orcements from a corps of young
ient with the policy activism fa-
go school exerted only moderate
occasionally a member might
riedman did with *Capitalism and*
owever, the Chicago economists
of most Americans, who took in
cs from the witty John Kenneth
eople realized that time and cir-
of the Chicagoans. But over the
hool came to exercise great influ-

ence on public policy in America and on regulatory policy in particular. During this brief period, three Chicagoans (Milton Friedman, Theodore W. Schultz, and George J. Stigler) were awarded Nobel Prizes in Economic Science.

In 1970, when *Nixon Agonistes* was published, *The Economics of Regulation* by Alfred E. Kahn also reached print. Kahn taught at Cornell, and he had been a liberal Democrat since the New Deal. In no sense did he belong to the Chicago school. Yet much in his book appealed to Friedman, Schultz, Stigler, and other Chicagoans: in particular, the arguments that regulation and competition are often at odds; that government officials often misunderstand the economic consequences of their decisions; and that market incentives are usually preferable to command and control regulations.[2]

At the time his book appeared, Kahn was not well known beyond the Cornell campus, where he served as dean of arts and sciences and as professor of economics. A few years later, he was a national celebrity. His success rested on the power and timeliness of his ideas, combined with a magnetic personality, which won him a series of political appointments. The first appointment came in 1974, when Kahn was named chairman of the New York Public Service Commission—at the time the most important state regulatory post in the country (and the highest paid, state or federal). In 1977 and 1978, he presided over the Civil Aeronautics Board, which then carried through the first major deregulation of an entire American industry. From 1978 until 1980, he chaired the president's Council on Wage and Price Stability and became known to the public as "the nation's chief inflation fighter." He was a media favorite, a fixture on the evening news.

A good deal of Kahn's newsworthiness derived from the simple fact that he was great copy. He loved Gilbert and Sullivan, enjoyed singing in amateur productions of *Iolanthe* and *Yeomen of the Guard*, and could quote hilarious lyrics from obscure operettas. In his best form Kahn was the model of a modern media general. When as chief inflation fighter he broached the possibility of "a deep, deep depression" and suffered the inevitable White House reprimand, he proffered a substitute word: "banana." That witticism prompted a second deluge of coverage. *Newsweek* called him "Alfred 'Bananas' Kahn," while a second magazine ran "The Sayings of Chairman Kahn." Even this did not end the game;

apprised of protests from banana interests, Kahn took fresh refuge in a new term, "kumquat."[3]

Later, when a Polish cardinal became Pope John Paul II, thus igniting a round of Polish jokes, Kahn—with his mind on the aberrant forecasts of his colleagues—solemnly declared that the new Pope had conquered Rome by telling economist jokes. For Kahn, one-liners became a trademark, even though he resisted the media's attempts "to make me a character," a "Walking Mouth." Profiled in the magazine *People* (surely one of few regulators ever selected for the honor), he complained afterward, "All life is a concatenation of ephemeralities." And during his tenure at the Civil Aeronautics Board, Kahn astonished Frank Borman and other Eastern Airlines executives, who were trying to educate him about the merits of different aircraft, by admitting that "I really don't know one plane from the other. To me they are all marginal costs with wings." Words seemed never to fail him.[4]

Yet to Kahn the principle of marginal costs—a fundamental tool of regulatory economics—represented no ephemerality and certainly no laughing matter. Unlike many other commissioners who shared his humanist values, he attached great importance to this favorite economic idea. He argued that in its neglect of marginal costs, more than at any other point, regulation had gone wrong. Once in office, Kahn determined to oversee improvements in regulatory practice by personally supervising the application of marginal-cost pricing.

Sound regulatory policy, he never tired of explaining, requires that buyers pay the marginal cost of all the goods and services they receive. If five units of an item cost $40 to produce and six units cost $60, then the marginal cost of the item is not $8 or $10 but $20. If the sixth unit is priced at $10 (that is, at average cost), consumers will purchase too many units—often not just one too many, but several—and since consumers have only a certain amount of money to spend, they will be able to buy too few units of other items, relative to what they would do under allocative efficiency. When goods and services are not priced according to marginal costs, therefore, consumers will automatically bring about a misallocation of society's resources. In order to prevent this unhappy result, Kahn believed, the prices of all goods and services should be set "at the margin"—that is, they should be pegged to the cost of producing one more unit at a particular time.

interests, Kahn took fresh refuge

became Pope John Paul II, thus
Kahn—with his mind on the aber-
-solemnly declared that the new
lling economist jokes. For Kahn,
, even though he resisted the
character," a "Walking Mouth."
surely one of few regulators ever
plained afterward, "All life is a
" And during his tenure at the
astonished Frank Borman and
who were trying to educate him
raft, by admitting that "I really
her. To me they are all marginal
never to fail him.[4]
rginal costs—a fundamental tool
ented no ephemerality and cer-
many other commissioners who
tached great importance to this
d that in its neglect of marginal
int, regulation had gone wrong.
o oversee improvements in regu-
ervising the application of mar-

ver tired of explaining, requires
of all the goods and services they
ost $40 to produce and six units
of the item is not $8 or $10 but
10 (that is, at average cost), con-
ts—often not just one too many,
have only a certain amount of
e to buy too few units of other
d do under allocative efficiency.
t priced according to marginal
tomatically bring about a misal-
order to prevent this unhappy
of all goods and services should
ey should be pegged to the cost
articular time.

But Kahn went further than this. He argued that marginal-cost pricing in regulated industries should also include the social cost of minimizing such "externalities" as air and water pollution. For an item such as electricity, pricing below marginal social cost would have not only the usual result of encouraging excessive consumption; it would also lead to the additional construction of expensive, polluting power plants, all the while driving the electric company toward bankruptcy. Remarkable as it now seems, this process was already underway in the state of New York in 1974, when Kahn became chairman of its Public Service Commission.

Kahn worked to persuade both producers and consumers of electricity that only some drastic change in the pricing system could avert disaster. Eventually he succeeded in getting marginal-cost principles partially incorporated into standard rate designs, so that they were factored into price levels. The resulting prices took the form of time-of-day differentials, designed in particular to discourage the careless use of air conditioning, which requires large amounts of electricity. Marginal-cost pricing nearly always raised difficult problems of implementation; but even so, in the case of New York, some of the practical changes were dramatic, as the difference between off-peak and peak hours now reflected the social cost of each type of use. In one of Kahn's "proudest accomplishments," for example, the New York Public Service Commission made summer air conditioning extremely expensive to consumers. As Kahn explained it,

> In Long Island, if you are a large residential user of electricity, you used to pay roughly five cents a kilowatt-hour, a damn high price no matter when you consumed it. Now you pay 3½ cents on a normal summer day per kilowatt-hour . . . and 2½ cents at night; and on summer days, when the temperature gets above 82 degrees, you pay 30 cents per kilowatt-hour. Now, that represents a much closer reflection to the comparative economic costs of your being supplied. Notice, this is a ratio of 12 to 1, as between night and a hot summer day, a much closer approximation to the respective marginal costs of supply. Think what that does to people's conservation habits. Think what it does to encourage the development of a technology that will enable you to use your electricity at night.[5]

The new pricing schedule discouraged careless use, checked the growth of demand, and diminished the need for new construction to serve the peak load.

Kahn liked to illustrate the point by comparing electric utilities to retail stores. To him, charging a flat rate for electricity regardless of the time of consumption seemed "like charging a flat price per pound for all items in a grocery or department store. What would happen if everything that came out of the cow—steak, hamburger, suet, bones, and hide—were priced at average cost per pound?" His answer, of course, was that everyone would buy steak and no one would want suet or bones. A basic principle emerged: "The only economic function of price is to influence behavior—to elicit supply and to regulate demand." No Chicago economist could have said it more clearly.[6]

A Scholar's Odyssey

That Alfred Kahn should become the nation's best known evangelist for marginal-cost principles—a set of ideas seldom associated with Kahn's liberal noneconomist friends—may seem ironic. Yet an explanation can be found in the long intellectual journey that marked his life. Kahn was born in Paterson, New Jersey, in 1917, the son of a Russian Jewish immigrant who worked in a silk mill. Young Alfred (or Fred, as he was called) soon proved to be an intellectual prodigy. At the age of fifteen he completed high school in New York City. At eighteen he graduated *summa cum laude* and first in his class from New York University, and before he turned twenty took a master's degree at NYU. He pursued further graduate work at the University of Missouri, then moved on to Yale for his doctorate.

During the same years, he also worked at a series of research jobs in Washington, for the Brookings Institution, the Department of Commerce, and the Antitrust Division of the Justice Department. Early in his career, Kahn began to acquire firsthand experience with the kind of policy-oriented social science research that was linking universities, foundations, and public bureaucracies into an integrated network with real power. For the next thirty years, he would continue to shuttle among the bases within this system. He worked with the War Production Board, the Federal Trade Commission, the staff of the Council of Economic Advisers, and as a member of the three-person Economic Advisory Council of American Telephone and Telegraph. And, whenever he needed to, Kahn returned to a stable anchorage: Cornell University and the life of a teacher and publishing scholar. He and his wife lived

IST'S HOUR

nt by comparing electric utilities
a flat rate for electricity regard-
eemed "like charging a flat price
cery or department store. What
t came out of the cow—steak,
e—were priced at average cost
se, was that everyone would buy
uet or bones. A basic principle
ıction of price is to influence be-
regulate demand." No Chicago
e clearly.[6]

the nation's best known evange-
a set of ideas seldom associated
friends—may seem ironic. Yet
ıe long intellectual journey that
Paterson, New Jersey, in 1917,
grant who worked in a silk mill.
called) soon proved to be an in-
ifteen he completed high school
e graduated *summa cum laude*
York University, and before he
ree at NYU. He pursued further
of Missouri, then moved on to

worked at a series of research
ngs Institution, the Department
Division of the Justice Depart-
gan to acquire firsthand experi-
ıted social science research that
ions, and public bureaucracies
real power. For the next thirty
le among the bases within this
Production Board, the Federal
Council of Economic Advisers,
on Economic Advisory Council
raph. And, whenever he needed
horage: Cornell University and
g scholar. He and his wife lived

in Ithaca, near the campus, and found it a wholesome place in which to raise their three children.[7]

Between 1940 and 1974, when he took his seat as chairman of the New York Public Service Commission, Kahn wrote two books and coauthored two others. All four proved to be major works. In addition, he managed to publish some two dozen articles and comments, most of them in the *American Economic Review,* the leading journal of the discipline. Nearly all of this writing reflected, in one way or another, the influence of his two principal advisers, Myron Watkins and Joseph Schumpeter. It also evidenced deep intellectual debts to the pioneering institutional economists, Thorstein Veblen and John Maurice Clark.

Kahn had begun graduate study in the 1930s, the last decade during which many of the leading economists in the United States were still avowed institutionalists. The giants in this distinctively American tradition—Richard T. Ely and John R. Commons of Wisconsin, Wesley Clair Mitchell and John Maurice Clark of Columbia, Schumpeter of Harvard—characteristically looked less to a deductive theory of prices and markets than to the social influence of economic institutions. Taking a cue from Veblen, they studied the sociological and anthropological aspects of such institutions as the leisure class, the business cycle, the capitalist legal system, the labor union, the business corporation, and the regulatory agency. This last "institution" proved especially important, and many institutional economists interested themselves in regulation. John R. Commons, Balthasar H. Meyer, and Rexford G. Tugwell, for example, actually helped to write regulatory legislation; they and many others also served as members of regulatory bodies.

In the late 1930s, when Kahn was coming of age as an economist, the institutionalist approach was moving steadily out of fashion, in the face of a revolution in macroeconomics—that subdiscipline devoted to the study of national economies. Of course, the Great Depression helped to stimulate this change by encouraging the new line of such work, including the masterpiece of macroeconomics, John Maynard Keynes's *General Theory of Employment, Interest, and Money* (1936). Within a decade, institutionalism had lost its eminence. Within three decades, Keynesian macroeconomics held the fort, overshadowing all other subdisciplines. Kahn stood as witness to the change.

In the meantime, classical microeconomics had slowly begun to

recover from an earlier slump of its own. Common topics of that subdiscipline included the theories of prices and markets pioneered by Adam Smith, John Stuart Mill, and Alfred Marshall. These giants of an earlier era had long dominated the field of economics. Over the first half of the twentieth century, however, the classical approach suffered three severe blows. The first came from institutionalism. Veblen, among others, savagely attacked Adam Smith's proposition that an invisible hand moved the free market automatically to serve the public interest, with no need for government intervention. The institutionalists asserted that here classicism was simplistic and merely deductive; it sacrificed both common sense and empirical observation to an alleged theoretical rigor that on close inspection turned out to be chimerical. As a case in point, argued the institutionalists, Smith and Mill offered no help for the human costs of industrialization. More than anything else, classicism's implicit assumption that whatever is, is right angered the institutionalists, who pointed to the Great Depression as a confirmation of their views. Clearly, the invisible hand—if there was one—had guided the world of the 1930s to disaster.

As if the institutionalist attack were not bad enough, classicism, like institutionalism itself, now began to suffer from Keynesian macroeconomics. For thirty years after the depression, many of the best advanced students were attracted to the new Keynesian theories. The remarkable performance of national economies offered a series of practical experiments for the benefit of those scholars and planners who thought in terms of aggregates. National economies were growing very rapidly, and only the "new economics" of Keynes seemed to offer clear explanations. Keynesian economists in many countries became the honored prophets of society.

At the start of the 1970s, however, the period of miracle growth abruptly ended, as a variety of powerful forces struck almost simultaneously: rising inflation, the first "oil shock," and intensified competition from Japan and other advanced industrial nations. Suddenly, in America, the old topics of classical microeconomics began to reassert their original value, as unfamiliar concerns about scarcity, overregulation, and the functioning of different types of markets again emerged to become the most serious problems for policy. In a short time, classical and "neoclassical" approaches (as

the redeveloping philosophy came to be called) took center stage. And with the revival of the classical framework came a resurgence of interest in marginal-cost pricing, already a favorite idea of Alfred Kahn's.

In fact, almost the entire intellectual history of twentieth-century economics played a part in Kahn's professional career. He entered graduate school during the late 1930s, at the ebb tide of institutionalism but before the flood of Keynesianism. Kahn understood microeconomics, yet felt a greater fascination with the macroeconomic side of his discipline. Even so, he emerged from his training as a disciple of neither school but as a broadly informed empiricist, equipped with an eclectic mixture of research tools, special interests, and what Keynes liked to call "propensities."

One propensity, inherited from his institutionalist forebears, committed Kahn to dissatisfaction with the status quo. In his first published article, an attack on American patent law appearing in the *American Economic Review* in 1940, the young scholar showed his reformism by protesting that the law unduly rewarded holders of patent pools, such as General Electric and other giant companies. Their "great research laboratories are only incidentally technological centers. From the business standpoint they are patent factories: they manufacture the raw material of monopoly." With its echoes of Veblen, this article was unmistakably the credo of an institutionalist. At the same time, it also bore the marks of a youthful impetuosity and a habit, which its author never outgrew, of saying more than he should—of issuing torrents of words, in part on extraneous subjects.[8]

Kahn's first book, *Great Britain in the World Economy* (1946), reflected broad research, careful thought, and the author's persistent effort to superimpose a theoretical framework onto an empirically dominated philosophy of institutionalism. The book seemed at once technical, historical, and theoretical. Well-written and bursting with energy, it was innovative chiefly in its eclecticism.[9] Like several of Kahn's subsequent publications, it also expressed his deep interest in development economics and in macroeconomics generally. In fact, except for the critical essay on patent laws, all of his work during the 1940s was directed at macroeconomic or international themes. Kahn himself moved from assorted government jobs in wartime Washington to a two-year stint teaching at

Ripon College in Wisconsin and later, in 1947, to an assistant professorship at Cornell, where he would remain. Many times, even after he became established as a scholar, he considered going to law school. He thought that a law degree, added to his training in economics, would increase his understanding of (and opportunity to influence) the making of public policy. As it turned out, he gained ample understanding and influence even without formal legal training.[10]

In the early 1950s, Kahn's interests began to shift toward regulation and marginal-cost pricing. The evidence appears in a second book and several articles on antitrust. Within the profession of economics, antitrust forms part of a larger field called industrial organization, which tends to combine some of the old policy concerns of the institutionalists with a more rigorous and theoretical analysis of market structures. The topic is microeconomic and heavily freighted with the ideological overtones that surround every discussion of the role of big business in a democratic society. If the central value implicit in classical microtheory was allocative and productive efficiency, the central value of antitrust—at least as the practical basis for enforcement of antitrust laws—seemed broader and fuzzier. For Kahn this new concern posed a dilemma he could not easily resolve. As a liberal Democrat writing in a time of national prosperity, he now found himself preoccupied less with the improvement of efficiency than with the promotion of social justice. In particular, he believed that national policy should not be misled by what he regarded as misdirected paeans to the American business system, such as David E. Lilienthal's popular book, *Big Business—A New Era* (1953) and John Kenneth Galbraith's *American Capitalism: The Concept of Countervailing Power* (1952).[11] Typical of the time in which they were written, both books derided antitrust controls as ineffective in promoting efficiency and, even worse, as harmful to American businesses threatened by international competition. Other analysts from both the political right and the left also argued for overhaul of the antitrust laws.[12]

Uncharacteristically, Kahn responded by defending the status quo. He insisted that the antitrust laws worked well and should not be altered. In a book he wrote with his colleague Joel Dirlam and in articles for the *Harvard Law Review*, the *Journal of Political Economy*, *Fortune*, and other periodicals, Kahn outlined a

thorough justification of the antitrust system. His argument took as its text not the deductive theories of microeconomics but the actual practice of antitrust control represented by leading judicial decisions.[13]

His book entitled *Fair Competition: The Law and Economics of Antitrust Policy* (1954) provides the most complete analysis of the problem. Here the burden of the argument, in particular its insistence that economics offers a poor guide to practice, departed from the approach Kahn would take in his later scholarly work and in his practice as a regulator. This episode turned out to be the intellectual crossroads of his life. In the pages of *Fair Competition* may be seen a struggle between the sentimental feelings of the liberal activist and the reasoned arguments of the analytical economist. In this case, for one of the few times in Kahn's professional life, sentiment routed reason.

The argument began on sufficiently sound footing. "One cannot simply equate the 'public interest' in a democracy with the 'consumer interest,' defining the latter as the interest of all citizens 'in getting more and better goods for consumption at ever lower prices.' " But from this solid platform Kahn and Dirlam stepped off into a quagmire. They asserted that even though the Sherman law promoted active government intervention in business decisions, "it represented, paradoxically, a departure from laissez faire in the ultimate interests of laissez faire." Through the device of "preserving the competitive regulator it [antitrust] would in the end minimize coercion."[14] Although the authors had a point here, they failed to recognize the connection between ends and means. Employing coercion as a means of minimizing coercion did not make sense. Whatever else antitrust prosecutions may have signified, they certainly were not undertaken "in the ultimate interests of laissez faire." Kahn and Dirlam also appear to have misunderstood the actual effects of antitrust policy. They too easily assumed that center firms must be the usual targets of prosecutors, ignoring the much greater impact antitrust actions had on loose horizontal combinations of peripheral firms—an impact that indirectly but powerfully promoted the rise of that very bigness in business which they themselves opposed.

Kahn and Dirlam worked themselves into a logical dilemma very much like the one experienced by Louis Brandeis forty years before. Their argument has a number of Brandeisian overtones: a

marked preference for small size, an assumption that consumers could be helplessly manipulated, and an abhorrence of advertising. As the authors describe what they see:

> In our society the individual's consumption pattern and the businessman's costs are surely not determined by truly untrammeled choices or truly free bargaining. Instead it is the business culture which imposes more or less invariable standards of the "good life" on consumers and fixes interest rates, wages, and other costs of doing business. And constantly disrupting the consumer's tranquillity are the advertising agencies . . . A large portion of our "highest per capita income in the world" consists in expenditures that are necessary to offset the horrible consequences of our unplanned, ugly, dirty, crime-ridden cities, and to keep up with the ubiquitous ad-man.[15]

Like Brandeis, Kahn and Dirlam moved from a disaffection with some of the aesthetic horrors of modern life to a rejection of the consumer values that seemed to underlie those horrors.

In a remarkable abnegation of their own profession, the authors went on to argue that because economists could never achieve unanimity, economics itself "does not offer clear-cut objective criteria for antitrust superior to those which have long prevailed" (that is, the legal criteria developed in judicial decisions). Certainly Kahn and Dirlam were correct to point out that economists, like other students of public affairs, did disagree about criteria for effective antitrust policy. But their argument for the irrelevance of economics on the grounds of economists' disagreements seems comparable to insisting that medical research has nothing to do with the care of patients so long as researchers continue to dispute over forms of treatment.[16]

Reading Kahn's book now, a generation after its completion, one reaches the conclusion that its analysis is largely confused and lacking in rigor. As a reviewer noted, the book falls between two chairs because "it mixes, in unspecified proportions, vague moral values and unsystematic economics."[17] To Kahn and Dirlam this kind of attack was especially painful, since they knew that in many respects they had written a courageous book. They had taken on all comers regardless of discipline, reputation, or popularity: falling between chairs sometimes cannot be avoided when scholars work at the borders of several disciplines.[18] It is now clear, however, that the real problem was not the lack of a disciplinary fit, but that the book exposed a significant division in Kahn's

an assumption that consumers
nd an abhorrence of advertis-
hey see:

umption pattern and the busi-
ed by truly untrammeled choices
e business culture which imposes
e "good life" on consumers and
osts of doing business. And con-
quillity are the advertising agen-
per capita income in the world"
sary to offset the horrible conse-
crime-ridden cities, and to keep

oved from a disaffection with
dern life to a rejection of the
derlie those horrors.
eir own profession, the authors
onomists could never achieve
t offer clear-cut objective cri-
which have long prevailed"
d in judicial decisions). Cer-
t to point out that economists,
did disagree about criteria for
rgument for the irrelevance of
nomists' disagreements seems
al research has nothing to do
esearchers continue to dispute

neration after its completion,
nalysis is largely confused and
d, the book falls between two
fied proportions, vague moral
"[17] To Kahn and Dirlam this
ful, since they knew that in
courageous book. They had
scipline, reputation, or popu-
mes cannot be avoided when
l disciplines.[18] It is now clear,
not the lack of a disciplinary
ignificant division in Kahn's

thinking about the problem of antitrust policy. That division, we have seen, originated in the mixed legacy of his graduate training; it grew more pronounced in the clash between his personal values and his increasing analytical powers.

In the years after publication of this book, Kahn's confidence in the rigors of deductive and theoretical research in neoclassical economics began to overcome his strong personal objections to the choices made by American consumers. Twenty-five years after the appearance of *Fair Competition,* he wrote a letter to the editor of a college newspaper, chiding the young journalist in language that Kahn himself would have done well to heed a quarter century earlier:

> You tended in your article to mix together your own tastes and values (most of which I share) with objective analysis; it is important to try to keep these two separate. For example, it is popular to define some kinds of consumption as wasteful and undesirable and others as necessary. But one of the values that we hold important in our society is the substantial freedom of each consumer to choose what to consume within the constraints of income. We let the individual decide how many children to have, how many cars, how much food. We let the individual decide how hard to work and on what to work jointly with any willing employer. We let people save and invest, work for wages or take entrepreneurial risk. We leave definition of want and waste to the market.[19]

These lessons came hard for Kahn. But he learned them well and then proceeded to teach them to others.

Toward Marginalism and a Masterwork

Alfred Kahn's transformation took place gradually over the 1950s and 1960s. In part, this change reflects important shifts in the basic economics of those industries that captured his attention; in part, change was in the man, as he grew older and developed new interests in fields dependent on neoclassical methods. His public declaration of this transition came in 1970, with the appearance of his *Economics of Regulation,* one of the most important books ever written on the subject. Here Kahn managed brilliantly to fuse the economics of institutionalism and classicism. That achievement, in its turn, led to a new phase of his career: in 1974, he entered public office as a regulator.

By the middle 1950s, shortly after the publication of his book on

antitrust, Kahn had begun to interest himself in the related area of regulation, particularly of the oil and gas industries. Again the evidence may be found in a book, which he wrote with another co-author and published in 1959 as *Integration and Competition in the Petroleum Industry*. Spinoffs from Kahn's research included articles on the control of crude oil production, an examination of the oil depletion allowance, and—most important for his future interests—a critical analysis of the Federal Power Commission and its role in the regulation of natural gas. Here he faced one of the most controversial and convoluted subjects in the entire history of regulation in America. The stakes at issue in oil and gas regulation were very high; by the late 1950s, the issues had developed such complexity that institutionalist economic methods could no longer be used to study them adequately. One reason for this is that oil and gas are joint products. They occur together in nature and the drilling for one often results in production of the other, without regard to what the driller intended. If one or both are to be subjected to utility-like price regulation, then joint production costs must be allocated between the two; this problem alone requires a mixture of economic and accounting theory that moves beyond the reach of institutional analysis. Indeed, any systematic thinking about the price issue requires a sophisticated application of neoclassical price theory, including the concepts of cross-elasticities, substitution, and marginalism.

Yet as a scholar steeped in institutionalist methods, Kahn enjoyed certain advantages. In the real world, neither oil nor gas was priced solely through the interaction of supply and demand. Instead, the prices of both were influenced by regulation. Oil was "prorationed" to market demand by the policies of state and federal agencies. It also was subject to a series of laws designed to help certain parts of the industry: schemes ranging from depletion allowances and other tax advantages to oil import quotas. By itself, natural gas also was regulated; it came under the purview of the Federal Power Commission, which engaged in a constant struggle to rationalize its pricing. Just at the time Kahn became interested in the subject, the FPC seemed to be failing in this job. As James Landis, in his *Report to the President-Elect* of 1960, reminded his readers, the FPC "represents the outstanding example in the federal government of the breakdown of the administrative process." The point was not lost on one reader at least.[20]

MIST'S HOUR

terest himself in the related area of oil and gas industries. Again the k, which he wrote with another co- s *Integration and Competition in* 's from Kahn's research included oil production, an examination of d—most important for his future the Federal Power Commission natural gas. Here he faced one of oluted subjects in the entire his- The stakes at issue in oil and gas e late 1950s, the issues had devel- titutionalist economic methods them adequately. One reason for roducts. They occur together in ften results in production of the e driller intended. If one or both price regulation, then joint pro- between the two; this problem mic and accounting theory that tional analysis. Indeed, any sys- ssue requires a sophisticated ap- eory, including the concepts of marginalism.

itutionalist methods, Kahn en- l world, neither oil nor gas was ion of supply and demand. In- uenced by regulation. Oil was y the policies of state and fed- o a series of laws designed to chemes ranging from depletion s to oil import quotas. By itself, ame under the purview of the engaged in a constant struggle time Kahn became interested e failing in this job. As James *t-Elect* of 1960, reminded his tstanding example in the fed- f the administrative process." r at least.[20]

Kahn knew that if he were going to do serious work on the regulation of oil and gas, he must master every relevant tool of economic analysis. To be sure, some of these tools belonged to institutional economics. But even more essential, he must learn to use better the methods of neoclassical microeconomics. His research into the pricing mechanisms of oil and gas had to be based on the theory of price. At this point in his career, Kahn began the final stage of his intellectual odyssey.

He now became directly involved in the difficult problem of pricing natural gas. In testimony before the Federal Power Commission during the early 1960s, he introduced a set of ideas that later became institutionalized as the "two-tier pricing system." Under FPC regulation, "old" gas, drawn from fields already discovered and producing, would be priced more cheaply than "new" gas, from fields brought into production more recently. Although at the time this seemed to Kahn a worthwhile idea (since it would prevent big companies from reaping unearned benefits from rising gas prices), in fact the two-tier arrangement raised a host of problems. Not only was the idea itself grounded in mistaken economic theory, but once in operation it also proved impossible for regulators to administer. Kahn later expressed his regret that, in attempting to impose discipline on large gas companies, he had "bequeathed to the nation an impractical system." In overestimating the actual degree of concentration in the natural gas industry, he had set in motion a policy that inadvertently penalized small companies along with large ones. Intellectually, it amounted to the same type of mistake made fifty years before by Brandeis, and indeed the same kind made only six years earlier by Kahn himself in his book on antitrust policy.[21]

But Kahn was learning fast now. He continued to study the natural gas issue, as he immersed himself in the vast multidisciplinary literature of regulation and started intensive work on his new book. As he later recalled, "My sense of the economics of all this [pricing of natural gas] didn't really get straight until I wrote my *Economics of Regulation.*" By 1967, Kahn had begun to chart the course of his masterwork and to present his early conclusions in several exploratory papers, each one delivered first at one of the many conferences on regulatory reform being held at this time. In 1970, the first volume of his great book appeared in print (volume two followed in 1971). *The Economics of Regulation: Principles*

and Institutions soon took its place as the leading textbook in the field of regulation. Measured by its impact on practitioners—both business executives and commissioners—it remains the most influential work written on the subject. Reviewers praised Kahn's "enormous job of scholarship," his "mature, confident and witty humanism," his "wholesome discipline and restraint." The book examined every aspect of economic regulation, from utilities and airlines to railroads and trucks; from externalities and the theory of second-best to the economics and politics of valuation and pricing.[22]

The triumph lies in Kahn's ability to join two subdisciplines of economics in what had long been regarded as an impossible synthesis. His intention is clear: "Together, the two volumes are an attempt to join neoclassical theory with 'institutional economics.' The latter is aimless if it is not informed by theory. And a normative theory of public policy is not of much use if it cannot be related to the selection of the best set of social arrangements for achieving those norms. Therefore, each volume [one subtitled *Principles*, the other, *Institutions*] is written with continuous reference to the other."[23] Fusing neoclassicism with institutionalism meant fusing economics with law, with history, with political science. It also meant analyzing the experience of public ownership in America and bringing together a colossal mass of heterogeneous material, much of it laden with misinformation and tainted with controversy. The two volumes totaled almost 600 large pages, with more than 1300 footnotes. Dozens of these footnotes were little articles in themselves: discussions of the legal, political, and economic literature of regulation and extended analyses of judicial decisions. Kahn brought all this off in a virtuoso performance, and thus made his escape from life's "concatenation of ephemeralities."

Most of *Principles* is devoted to an exposition of marginal-cost pricing. This approach was still unusual, but it foreshadowed Kahn's priorities when he later served as a regulator. Now, he wrote, "The economist who asks a politician 'Won't you let me advise you on your legislative program?' deserves another question in response: 'What do you *know*?' And only the economist who can answer 'Well, I can tell you that if you pass a law that says such-and-such, these are the things that will probably happen,' or 'If you do nothing about such-and-such, this is what will probably happen,' deserves to have his offer taken seriously." The message

ace as the leading textbook in the
its impact on practitioners—both
ssioners—it remains the most in-
ubject. Reviewers praised Kahn's
his "mature, confident and witty
cipline and restraint." The book
mic regulation, from utilities and
from externalities and the theory
and politics of valuation and pric-

ility to join two subdisciplines of
n regarded as an impossible syn-
ogether, the two volumes are an
ry with 'institutional economics.'
rmed by theory. And a normative
uch use if it cannot be related to
cial arrangements for achieving
me [one subtitled *Principles,* the
ith continuous reference to the
th institutionalism meant fusing
, with political science. It also
of public ownership in America
mass of heterogeneous material,
ation and tainted with contro-
most 600 large pages, with more
ese footnotes were little articles
egal, political, and economic lit-
d analyses of judicial decisions.
oso performance, and thus made
ion of ephemeralities."
an exposition of marginal-cost
unusual, but it foreshadowed
erved as a regulator. Now, he
a politician 'Won't you let me
ram?' deserves another question
' And only the economist who
hat if you pass a law that says
that will probably happen,' or
such, this is what will probably
taken seriously." The message

seems clear: in the arena of practical regulation, Kahn knew with a rare certainty that the economist's hour lay at hand.[24]

He could hardly have timed his book better. The two volumes appeared in 1970 and 1971, the early years of a period destined to be plagued by a set of economic problems newly christened as "stagflation." The marvelous postwar era of rapid growth had suddenly ended. Now, for utilities as for other industries, the future called for reduced markets. This prospect clashed with the prevailing business strategies of many utilities, especially electric companies. These companies had anticipated—based on steady growth of demand and increasing economies of scale in their new plants—that they would continue to build large projects and would need ever-growing amounts of capital. In terms of net capital investment, electric power was America's largest industry and also its most capital-intensive, requiring the greatest amount of investment to produce a given dollar of revenue. With the total invested capital measured in hundreds of billions of dollars, utilities faced a continuous obligation to pay their heavy burden of debt. For a generation after World War II, this problem did not seem serious. So long as sales increased and fuel remained inexpensive, so long as inflation remained moderate and interest rates on borrowed capital stayed low, so long as construction costs held steady and technological innovation continued—under these conditions the task of utility management was simple: to promote a steady growth of demand and to build the generating capacity to meet it.

By the time Kahn's *Economics of Regulation* reached the public, every element underlying this task had begun to disappear. National economic growth, which had averaged more than 4 percent a year for thirty years, dropped to 2.6 percent in 1969 and then further, to a negative .3 percent for 1970. (The average for the 1970s stood at 2.87 percent, nearly a third lower than the rate for the preceding three decades.) And the national inflation rate, which had averaged 2.5 percent a year in the 1950s and 1960s, climbed to 5.3 percent in 1971 and averaged 6.7 percent for the decade of the 1970s. These national trends meant that the economics of the utility business was about to be turned upside down.[25]

Stagflation represented bad news for utilities, and three additional developments made things even worse. First, the environ-

mental movement: again the watershed year was 1970, when the celebration of Earth Day ushered in the creation of a new federal Environmental Protection Agency. The sudden popularity of environmental concerns hit utilities hard. Power plants pollute the environment in almost every conceivable way—indirectly through the strip mining of the coal they burn, directly through smokestack emissions, thermal pollution of cooling water, and leakage of radioactive materials. To prevent pollution, even to control its sources, soon becomes extremely costly. A utility must pay for nuclear-waste facilities, scrubbers for high-sulfur coal, and towering stacks (some more than 1000 feet high), and all are expensive. A national consensus, however, required that environmental quality be improved, and it formed just when economic growth began slowing and when inflation and interest rates began rising.

A second problem emerged from the sudden slowing of technological progress in the design and fabrication of generating equipment. For almost a century after the construction of Thomas Edison's first electric power plant, technological progress brought steady gains in productivity. Engineers and utility managers had come to take such gains for granted, thinking they would continue indefinitely. Beginning in the 1960s, however, a technological plateau was reached in the industry, as yearly gains came to an unexpected halt, particularly in those large-scale electric generating plants on which most power companies relied. Between 1967 and 1976, utility executives discovered that the most technologically advanced plants, which used steam at very high temperatures to turn gigantic turbines, went out of service almost three times as often as smaller-scale plants. Bigger came to mean not necessarily better, only less reliable.[26]

Then late in 1973 the price of imported crude oil shot up to four times the summer 1973 figure. Utilities were stunned, especially those located in the northeastern United States—where the companies had earlier converted to oil as a fuel less polluting than coal—and where Alfred Kahn was about to become a key regulator.[27] With the shock of the new oil pricing, the last comfortable assumption of regulatory commissioners and utility managers disappeared, along with every shred of confidence in low interest rates, low inflation, high economic growth, and constantly improving technology.

atershed year was 1970, when the
ed in the creation of a new federal
ıcy. The sudden popularity of en-
es hard. Power plants pollute the
ıceivable way—indirectly through
ey burn, directly through smoke-
n of cooling water, and leakage of
nt pollution, even to control its
y costly. A utility must pay for
ers for high-sulfur coal, and
an 1000 feet high), and all are
however, required that environ-
d it formed just when economic
ı inflation and interest rates be-

m the sudden slowing of techno-
fabrication of generating equip-
er the construction of Thomas
, technological progress brought
ineers and utility managers had
d, thinking they would continue
0s, however, a technological pla-
as yearly gains came to an unex-
large-scale electric generating
panies relied. Between 1967 and
d that the most technologically
m at very high temperatures to
f service almost three times as
r came to mean not necessarily

ported crude oil shot up to four
ilities were stunned, especially
United States—where the com-
il as a fuel less polluting than
about to become a key regula-
il pricing, the last comfortable
oners and utility managers dis-
of confidence in low interest
c growth, and constantly im-

The impact of these multiple revolutions transformed the entire economic picture. The overall effect on power companies can be gauged by the rapid rise in their combined requests for new rate increases.[28]

Quarter ending	*Cases filed nationally during quarter*	*Total increases requested*
3/31/70	12	$89 million
3/31/72	22	$171 million
3/31/74	45	$638 million

Because of accelerating inflation, utilities that based their rates on historical costs, as most did, had difficulty in meeting current expenses. Even when a regulatory agency did grant the requested rate increase in full, the time lag between petition and decision guaranteed that the companies would keep running faster and faster in pursuit of outdated targets. By 1974, the length of procedural delays averaged eleven months; and when inflation approached double digits, "regulatory lag" soon became a serious problem. All this set the stage for a warm reception of Kahn's work, which addressed itself directly to the revolutions in the underlying economics of utilities and of regulation in general.[29]

Ratemaking versus Stagflation

Kahn knew about the utilities' mushrooming revenue requirements, but he was even more concerned about the structure of their rates. Here, as he wrote in *Economics of Regulation*, Kahn could point to an area in which the professional economist might make a distinctive contribution, by starting from a foundation of marginal-cost pricing. The existing rate structures contained an implicit assumption that the costs of production would always be declining, as indeed they had done until the late 1960s. During those happier years, utility companies had experimented with declining-block rates, under which the price of a kilowatt-hour would be reduced with every increase in the total consumed, much like volume discounts for large purchases in retail stores.

A pioneer in this field, the publicly owned Tennessee Valley Authority, introduced as early as 1933 its radical declining-block rate structure:[30]

First 50 kwh per mo.	3.00 cents per kwh
Next 150 kwh	2.00 cents per kwh
Next 200 kwh	1.00 cents per kwh
Next 1000 kwh	0.40 cents per kwh

The central mission of the TVA was to encourage the widest possible usage of electricity, thereby improving the standard of living in the depressed Tennessee Valley region. This 3¢-2¢-1¢ rate schedule, based on a commercial jingle of the period that touted 3-2-1 Shoe Polish, directly promoted the goal. Fully electrified homes, each charged at an average rate of about 0.7¢ per kilowatt-hour, represented TVA's objective, at a time when the national average stood at 5.5¢. The agency took an enormous gamble: that, by slashing prices, it could multiply consumption and, further, that such increased consumption would permit savings based on scale economies in generation, which in turn would justify the price cuts already made. Only a government body could afford to take such a bold step, and the gamble did pay off. In one district located in Mississippi, as rates dropped from 7.40¢ to 1.58¢, usage more than tripled. By 1937, three years after the start of the experiment, the price of electricity in the entire TVA region stood at less than half the national average, and usage was about 60 percent greater.[31]

These results proved revolutionary. News of the TVA experiment stimulated thinking all over the world, especially the thinking of economists on the responses of utility customers to strong price signals. Kahn himself, while still a graduate student, became enamored of the experiment. One of his boyhood idols had been Senator George W. Norris, the "father of the TVA." And as Kahn remarked years later, "TVA did what regulatory commissions could not require private companies to do: set low rates and see what happened to demand and cost."[32]

Before fixing their rates, TVA's planners had made two important assumptions. First, all of the agency's generating capacity, including new production from hydroelectric dams under construction, would actually be sold. Only if this power was sold would the lower rates cover TVA's costs; that is, the "marginal cost curve" could tend downward only if usage tended upward. (Henry Ford had demonstrated the same principle years earlier, by making Model T's at less and less cost and then continuously cutting the selling price of each manufactured unit.) Second, TVA

3.00 cents per kwh
2.00 cents per kwh
1.00 cents per kwh
0.40 cents per kwh

was to encourage the widest pos-
improving the standard of living
lley region. This 3¢-2¢-1¢ rate
l jingle of the period that touted
noted the goal. Fully electrified
age rate of about 0.7¢ per kilo-
bjective, at a time when the na-
The agency took an enormous
, it could multiply consumption
consumption would permit sav-
generation, which in turn would
e. Only a government body could
nd the gamble did pay off. In one
s rates dropped from 7.40¢ to
y 1937, three years after the start
ctricity in the entire TVA region
al average, and usage was about

nary. News of the TVA experi-
the world, especially the think-
es of utility customers to strong
still a graduate student, became
e of his boyhood idols had been
ather of the TVA." And as Kahn
l what regulatory commissions
ies to do: set low rates and see
st."[32]

planners had made two impor-
e agency's generating capacity,
hydroelectric dams under con-
l. Only if this power was sold
A's costs; that is, the "marginal
l only if usage tended upward.
ne same principle years earlier,
less cost and then continuously
anufactured unit.) Second, TVA

assumed that the annual "load factor" for its system would soon reach 50 percent. Load factor, a vitally important statistic for every utility, expresses average usage by all customers during any given period as a percentage of their usage during the heaviest period of the year (or, simply, as a percentage of the utility's capacity). For electric companies today, the peak-load period often comes in midsummer, when residential consumers use their air conditioners. Thus, if the average load on a company's system is 75,000 kilowatts and the peak load 150,000, then load factor is 50 percent. Eventually TVA achieved such success that its load factor reached 70 percent, an extremely high figure among electric utilities.[33]

Such numbers are watched closely by all power companies because of the peculiar economics of the industry. Electricity, unlike most other goods and services, cannot be stored efficiently. It must be generated, transmitted, and consumed—in almost the same instant. This means that the companies must have enough generating capacity to meet not only the usual demands of their customers (average load) but also the highest demands (peak load). Since the companies are often required by law to serve all customers, whatever their demands, utility managers and regulatory commissioners have tended to assume that they themselves have little discretion in raising or lowering either the supply of electricity or the demand for it. Instead, company managers planned for the future by forecasting the total demand their customers would impose, then scheduling the construction of new generating plants to meet the pace of added demand. So the task of managing an electric utility largely amounted to planning and financing new construction. Managers gave little thought to measures that might affect consumer demand, except to promote ever greater consumption. This, in turn, had the effect of promoting more and more construction of new plants.

Utility managers acted in this way for several reasons. First, they felt the normal urge of business executives to encourage the consumption of whatever they sell. Second, and less obvious, they could factor the full cost of capital expenditures into their company's rate base and thus augment the potential revenues requested from the regulatory agency.[34] Finally, they could take advantage of the rapid rate of innovation in the technology of electric generation, particularly innovation within extremely large plants. Scale economies in this industry always seemed impressive;

in the post–World War II period, however, they looked to be unbelievably advantageous. Kahn pointed this out in one of his early papers on the subject. He said that the annual productivity increases in electric utilities averaged an astonishingly high 5.5 percent over the first half of the twentieth century. This figure dwarfed the 1.7 percent for the American economy as a whole during the same period. (At 5.5 percent per year, productivity will double every thirteen years. At 1.7 percent, it will double only every forty-one years.) The technology of electric power generation, therefore, was far outstripping innovation elsewhere in the economy—even *before* the construction of the gigantic and extremely efficient power plants that characterized the postwar period.[35]

The results of these innovations and scale economies could be read in a steady decline of the consumer's real costs. As Kahn later said, the real price of energy (not electricity alone) had dropped by 43 percent between 1951 and 1971, a period in which consumer prices as a whole rose by 56 percent.[36] Electricity seemed uniquely to be a bigger bargain every year. But then, at the close of the 1960s, this felicitous situation abruptly changed, and the real costs moved steadily higher during the 1970s and early 1980s. Inflation, stagnation, the environmental movement, rising interest rates, technological stasis, and the skyrocketing cost of fuel struck the industry simultaneously. Utilities' marginal-cost curves, which had shifted downward since the first Edison stations had opened nearly a century before, now moved ominously upward. Construction of new plants became almost prohibitively expensive. This meant that, for the first time in the industry's history, electricity coming from new plants would cost more than equivalent amounts from existing sources. Thus the declining-block rate structures in effect throughout the United States now gave exactly the wrong signals to consumers. Utility advertising campaigns, moreover, continued to urge greater consumption, which would inevitably lead to greater peak loads and increased demand for the construction of new power plants. The cycle seemed insane.

Into this setting, Kahn launched his *Economics of Regulation* in 1970 and then his own career as a regulator in 1974, when he became chairman of the New York Public Service Commission. For him the complex situation proved exhilarating rather than depressing; he knew that he had answers to the questions that baffled

d, however, they looked to be un-
pointed this out in one of his early
that the annual productivity in-
ged an astonishingly high 5.5 per-
twentieth century. This figure
American economy as a whole
ercent per year, productivity will
1.7 percent, it will double only
nology of electric power genera-
ing innovation elsewhere in the
truction of the gigantic and ex-
at characterized the postwar pe-

as and scale economies could be
sumer's real costs. As Kahn later
t electricity alone) had dropped
d 1971, a period in which con-
56 percent.[36] Electricity seemed
very year. But then, at the close
tion abruptly changed, and the
uring the 1970s and early 1980s.
nental movement, rising interest
skyrocketing cost of fuel struck
ies' marginal-cost curves, which
irst Edison stations had opened
oved ominously upward. Con-
almost prohibitively expensive.
in the industry's history, elec-
ould cost more than equivalent
Thus the declining-block rate
he United States now gave ex-
mers. Utility advertising cam-
ge greater consumption, which
eak loads and increased demand
ver plants. The cycle seemed

his *Economics of Regulation* in
regulator in 1974, when he be-
ublic Service Commission. For
d exhilarating rather than de-
ers to the questions that baffled

utility managers and other commissioners. In the last academic paper he published before taking his seat in Albany, Kahn returned to the theme of the economist's role in making policy. Clearly, Kahn felt very sure of himself, as he pointed to the place where answers would be found: "Applied microeconomics is the exciting new frontier of public policy."[37]

The New York Public Service Commission

Together with California and Wisconsin, New York for many years was acknowledged as a leader of regulatory practice. This had been the case since the early part of the twentieth century, when the movement to regulate electric and gas utilities began to spread throughout the country. Later on, a bewildering variety of other industries fell under state commissions, and in many states the same agency exercised jurisdiction over these industries as well as over electric and gas utilities. The New York commission in 1974 regulated more than forty, including telephones, buses, water supply systems, docks and wharves, and warehouses. The degree of authority actually exercised, of course, varied widely according to the industry, both in New York and elsewhere. It also varied according to the vigor and resources of the state commissions.[38]

In many other states, regulatory commissions served as dumping grounds for political hacks and cronies of the governor. So long as regulation remained unimportant—which it did, relatively speaking, during the years of steady economic growth and declining costs for utility services—membership on the commissions did not receive much public attention. Even during the boom years of increasing prosperity, however, a commissionership in one of the three leading states, and especially a chairmanship, offered a prestigious and desirable post for a talented candidate.[39]

Most state commissions had three members, but the numbers in 1974 ranged from one in Oregon to seven in Massachusetts and South Carolina. The typical term of office was six years, with staggered appointments designed to minimize the disruption from turnover and to provide some insulation from partisan politics. Governors appointed the commissioners in three-fourths of the states, voters elected them in one-fourth. Some states specified that consumer interests must be explicitly represented among the commissioners, and about half of these state agencies, like most

federal commissions, by law had to have minority party representation.[40]

The state and federal commissions were filled with lawyers, who together accounted for more than 80 percent of the total membership of some federal commissions and at no agency less than 50 percent. Their representation on state commissions was very large as well, and their domination of important regulatory jobs is a phenomenon full of significance for the study of the values reflected in the commissions' policies. In particular, lawyers and legal patterns of thought must be seen to underlie a tendency of the commissions to emphasize procedural due process rather than economic efficiency. Alfred Kahn is only one of many economists who have made this point in criticizing regulatory performance.[41]

As bureaucratic institutions, the state commissions varied greatly in size and complexity. Some were only slightly larger in 1974 than Charles Francis Adams' Massachusetts agency had been in 1874. New Mexico's commission had only 17 staff members. The Massachusetts commission, on the other hand, had grown to a total of about 120 full-time employees.[42] The New York Public Service Commission was second in size only to California's. As chairman, Kahn headed an agency of some 650 persons, with an annual budget in 1974 exceeding $12 million. Kahn received the highest salary of any regulator in the United States—about $51,000—and in addition enjoyed a chauffeur-driven limousine and many other benefits. A listing of his professional staff, together with their individual specialties, provides us with a useful index to the nature of the agency at the time Kahn took charge. Notice the dominance of engineers, accountants, and lawyers, and the minor role played by professional economists.[43]

Professional staff		*Division of labor*	
Inspectors and investigators	103	Chairman's office	17
Engineers	57	Energy division	231
Auditors	54	Acctg. and utility finance	97
Accountants	28	Communications division	92
Attorneys	21	Ofc of administration	79
Administrative staff	14	Ofc of general counsel	46
Rate analysts	13	Ofc of environmental planning	41
Hearing examiners (admin. law judges, mostly lawyers)	13	Water division	34
Economists	8	Ofc of hearing examiners	20
		Ofc of economic research	12

IST'S HOUR

to have minority party represen-

ons were filled with lawyers, who
80 percent of the total member-
s and at no agency less than 50
state commissions was very large
nportant regulatory jobs is a phe-
e study of the values reflected in
cular, lawyers and legal patterns
e a tendency of the commissions
ocess rather than economic effi-
of many economists who have
ılatory performance.[41]
the state commissions varied
me were only slightly larger in
Massachusetts agency had been
had only 17 staff members. The
other hand, had grown to a total
[2] The New York Public Service
ly to California's. As chairman,
50 persons, with an annual bud-
Kahn received the highest sal-
States—about $51,000—and in
en limousine and many other
l staff, together with their indi-
a useful index to the nature of
harge. Notice the dominance of
, and the minor role played by

Division of labor

Chairman's office	17
Energy division	231
Acctg. and utility finance	97
Communications division	92
Dfc of administration	79
Dfc of general counsel	46
Dfc of environmental planning	41
Water division	34
Dfc of hearing examiners	20
Dfc of economic research	12

Even though its economic expertise far surpassed that of regulatory agencies in most states, the New York commission did not emphasize economic analysis in its work. For Kahn this situation posed a dilemma. He could not inaugurate major new hirings because the commission's budget was practically frozen. This left two alternatives. He could attempt to overhaul the present staff, letting go some of the inspectors, investigators, engineers, accountants, and lawyers, and hiring trained economists in their places. Or he could try to convince his existing staff members that economics must be relevant to their own jobs, even though many of them had been doing their work for years without any formal training in the discipline. Neither choice seemed to provide an easy road to popularity. Even so, the prospect of retooling his commission staff appeared to Kahn entirely feasible as he looked outward at the condition of the utility industry in 1974.

The Commission's Strategy

The combined impact on New York's electric utilities of inflation, higher fuel costs, technological stasis, and other forces may be gauged in the history of rates charged by the Consolidated Edison Company, the state's largest utility. Con Edison served Greater New York City, and there residential customers saw the average monthly electric bill double quickly during the early 1970s.[44]

1965	$ 9.68
1970	10.65
1974	22.08

In 1974, even with these higher rates, Con Edison still found itself in the worst financial shape in its history. So it naturally expected to request from the Public Service Commission still another large increase in rates during 1975.

Kahn anticipated this situation and came into office with a clear idea of what he should do. As a fundamental strategy, he planned to reform rate structures so as to make them reflect true costs. Of course he wanted to do this not for Con Edison only, or even for electric utilities alone, but also for gas and telephone rates. As an economist, he knew that such free services as telephone directory assistance were in fact not free at all. They were subsidized by

other charges that subscribers who never dialed Information paid as part of their bills. Kahn believed that each consumer should pay for services actually received and for no others. Yet his studies had convinced Kahn that such subsidies were being paid on a large scale. He found the rate structures of nearly all the state's utilities to be shot through with unfair distortions, resulting in general abuse: some consumers were subsidizing others, and nearly all customers were receiving signals to buy when they should be warned to conserve. Kahn accepted the chairmanship with a conviction that only decisive action by the commission could correct these signals and encourage conservation. If the commission acted quickly, moreover, it might help slow or even reverse New York electric utilities' ominous drift toward bankruptcy. Kahn's strategy was to make electric rates coincide with the companies' accelerating current costs, and not with obsolete, mythical patterns predicated on a continuing decline in costs.

Now the question became not what to do but rather how to do it. How could Kahn convince his fellow commissioners, his 650-member staff, the utilities themselves, and eight or ten different classes of consumers that the old practices of the companies and the commission were inappropriate to current reality? Generally in the American regulatory system, commissions act as collegial bodies, led by chairmen who have little more power than do each of their colleagues. Chairmen possess almost no executive authority; they cannot compel anyone (except a member of their personal staffs) to do anything. Such use of force did not suit Kahn's style in any situation. His strength was persuasion, not dictation. For twenty-seven years he had been one of Cornell's best teachers, as his classes filled with students eager to hear his famous lectures. He excelled in small seminars as well, helped by an engaging manner, wit, and conviction. These same qualities would work for him in the commission setting, as Kahn the master teacher determined to win over a new group of students: the commission staff.

To begin, he did almost no hiring or firing but instead worked with a few carefully chosen pupils, then a few more, spreading his message in widening circles. Even those who disagreed with a marginal-cost philosophy had to respect Kahn's learning, and nearly all staff members appreciated his informal manner. A visitor from the *New York Times* found him padding around the office in his socks. Kahn spent little time behind his desk, wrote this

ıo never dialed Information paid
d that each consumer should pay
for no others. Yet his studies had
lies were being paid on a large
s of nearly all the state's utilities
distortions, resulting in general
dizing others, and nearly all cus-
ɔuy when they should be warned
chairmanship with a conviction
commission could correct these
tion. If the commission acted
slow or even reverse New York
ward bankruptcy. Kahn's strat-
coincide with the companies'
ɔt with obsolete, mythical pat-
decline in costs.
vhat to do but rather how to do
fellow commissioners, his 650-
lves, and eight or ten different
practices of the companies and
te to current reality? Generally
n, commissions act as collegial
little more power than do each
ɔssess almost no executive au-
ɔne (except a member of their
Such use of force did not suit
rength was persuasion, not dic-
had been one of Cornell's best
udents eager to hear his famous
inars as well, helped by an en-
n. These same qualities would
setting, as Kahn the master
ew group of students: the com-

g or firing but instead worked
then a few more, spreading his
those who disagreed with a
respect Kahn's learning, and
d his informal manner. A visi-
him padding around the office
e behind his desk, wrote this

Times reporter, preferring instead "the informality of a bentwood rocking chair and a pair of couches that fill a corner of the office." The chairman "is an educator at heart who believes that he can convince any rational person of the rightness of his position."[45] Small wonder that the press felt an interest in this new kind of bureaucrat. Yet Kahn was not merely putting on a show for reporters. Such informal behavior was his pattern throughout three years on the Public Service Commission. He talked endlessly with his staff about the principles of marginal-cost pricing. He delighted in teaching them at every opportunity: over lunch, in meetings, even at the daily swims he initiated. Immersed in the water, visible only from the neck up, Kahn lectured on the technicalities of cross-substitution, elasticities of demand, and marginalism. For a professor in his late fifties, he seemed amazingly youthful and vigorous and, above all, sincere.[46]

Kahn's friendly manner proved especially important in his effort to convert the commission's rate engineers. Because of the technological history of electric utilities—improved technology as a basis for the steady growth of scale economies and declining rates—engineers became much enamored of the usual system of volume discounts for businesses and declining-block rates for home consumers. In the engineers' view, large users of electricity deserved to pay less per unit consumed. On other hand, the engineers also felt close to panic over the current situation, which they clearly saw as an unprecedented crisis.[47] They defined this crisis as one of "revenue erosion" for the companies: costs were accelerating faster than revenues, and this produced a conflict between the volume-discount rate structures and the upward movement of marginal-cost curves. Here Kahn's marginalist approach provided his engineers with a convenient solution to the problem of revenue erosion. For a time, the engineers remained skeptical about marginalism because so much was left to guesswork. Kahn told them that marginalism was indeed difficult to implement, that it was full of subjective judgments. His candor seemed appealing yet, given their training, the engineers naturally preferred precise solutions over rough approximations. To meet their objections, Kahn again and again asked them, "Do you want to be precisely wrong or approximately right?" Ultimately he won them over.[48]

When approached by people from outside the commission, Kahn took a conciliatory position. He insisted that all parties for-

get their anger of the moment and take a longer view. Much as Charles Francis Adams had done a century earlier, Kahn emphasized the unity of interests between companies and customers. Also like Adams he used the *Annual Reports* of the commission to set forth his strategy in explicit terms. The commission's aim, he wrote in his first report, was to set policy not in a series of dramatic and episodic rate cases, but in a continuum—to relate the past to the present and the future. In the 1950s and '60s, Kahn wrote, rate cases were rare, problems not serious. Both producers and consumers benefited from the declining-block rate structure. The world of the 1970s was very different. Inflation, tight money, and high interest rates had brought from the utilities an endless stream of requests for rate increases, and the response by consumers had been equally reflexive. "All too often," Kahn wrote, "the proponents [of] 'consumerism' interpret their responsibility as one loudly opposing any and all rate increases." This too was understandable. But because of strong pressure from consumers, "it should not be surprising that regulators in 1974 generally squeezed utility profits harder than may be in the consumer's long-term interest." Kahn expressed his determination to regulate with an even hand, favoring neither producers nor consumers.[49]

Generic Rate Hearings

The regulatory process in the United States has been desultory in character. State and federal commissions, somewhat like courts, have seldom initiated action on their own, but instead have responded to petitions or complaints. As a consequence of this reactive role, they find it difficult to change policies quickly, especially when such changes are opposed by powerful interest groups. These groups are often ready to prevent loss of advantage by falling back on litigation calculated to delay any change.

Kahn knew that, in New York, he could not allow for these normal delays, despite the high level of controversy that surrounded his ideas. The commission must move quickly, rather than wait for utility companies to initiate proceedings, especially since nothing could be expected of the companies except additional requests for rate increases. The major problem, Kahn believed, lay not with rate levels (the total amount of revenue needed by the utilities) but rather with rate structures (the mix of charges to different

nd take a longer view. Much as
a century earlier, Kahn empha-
veen companies and customers.
ual Reports of the commission to
erms. The commission's aim, he
et policy not in a series of dra-
t in a continuum—to relate the
re. In the 1950s and '60s, Kahn
ems not serious. Both producers
declining-block rate structure.
different. Inflation, tight money,
ht from the utilities an endless
es, and the response by consum-
ll too often," Kahn wrote, "the
rpret their responsibility as one
ncreases." This too was under-
pressure from consumers, "it
tors in 1974 generally squeezed
n the consumer's long-term in-
nation to regulate with an even
or consumers.[49]

d States has been desultory in
issions, somewhat like courts,
eir own, but instead have re-
As a consequence of this reac-
ge policies quickly, especially
by powerful interest groups.
vent loss of advantage by fall-
delay any change.
could not allow for these nor-
controversy that surrounded
quickly, rather than wait for
ings, especially since nothing
except additional requests for
Kahn believed, lay not with
enue needed by the utilities)
mix of charges to different

classes of customers spread over different blocks of electric power). These structures could not be altered by the usual process of increase after increase.[50]

Confronted with the dilemma, Kahn found his answer in the device of holding not individual rate cases but a "generic hearing" instead. His hearing would amount to an intensive study of the rate question, conducted outside the heated atmosphere typical of the usual rate case. Not one company but all companies would be involved, and testimony would include evidence from environmental experts and other specialists interested in the proceedings. Best of all from Kahn's viewpoint, he himself might be able to shape the procedures and influence the outcome far more than in the usual commission hearing. In short, he viewed the generic hearing as a novel opportunity to educate all parties on the benefits to them of marginal-cost pricing. Confidently relying on his own ability to convince doubters, Kahn foresaw enormous gains and very little risk.[51]

Yet he sought to reduce the risks still further by cultivating a set of useful allies, among them the Environmental Defense Fund. For several years this organization had been intervening in regulatory proceedings throughout the United States, at first from an almost purely adversarial position. Like many other environmental groups, it tended to oppose almost all projects, especially the construction of new power plants. As time wore on, some staff members of the Fund, in particular a young lawyer named Edward Berlin, began to search for ways in which the organization could make a more positive contribution.

This quest led Berlin to the device of marginal-cost pricing for electricity. But whereas Kahn's motivation in advocating marginalist principles was to hold down the need for rate increases, and to achieve allocative efficiency in the economic sense, that of Berlin and his colleagues emphasized the use of a new price system as the means of slowing new construction of generating plants. The traditional declining-block rate structure had encouraged greater use of electricity and more new construction. But now, if utility customers could be persuaded through price signals to reduce their consumption (or to shift consumption to off-peak hours), then existing plants could better serve their needs. New construction could be delayed, perhaps defeated altogether, and the environment would be protected. This general line of thinking led the

Environmental Defense Fund to adopt a spirited advocacy of marginal-cost pricing. Kahn already knew Berlin and others in the Fund through a consulting firm of economists with which Kahn himself had been affiliated. Berlin and the consulting firm had participated in a pioneering marginal-cost case in Wisconsin.[52] Based on that experience, the Fund quickly accepted Kahn's invitation and petitioned the New York Public Service Commission to hold generic hearings. In this way, the initiative for the new hearings seemed to derive from a broader base of interest than if it came from Kahn alone.

Still, assistance from the Fund could prove a mixed blessing. If the new chairman seemed to align himself with an interested party in a rate hearing, he would be courting serious trouble—and doubly so because of Kahn's prior connection with the consulting firm, which was also expected to take part. Kahn met this difficulty by broadening his support even more. He asked the chief executive officers of New York state's seven major utilities to come to his office in Albany for a conference. Gathering them all in one room, he explained his devotion to marginalist principles, listened to their views, and went on to persuade them to petition as a group for the generic rate hearing. They agreed to his choice, and Kahn had scored a great coup. Now the most important utility companies were joining with a leading environmental group to ask the commission to do what the chairman had wished in the first place.[53]

The various companies had different reasons for joining in this petition. Several had already become interested in marginalist principles themselves as possible alternatives to declining-block structures. Others regarded a generic hearing as a harmless academic exercise, and they wanted to appear cooperative. And a few others of the seven saw the hearing as a welcome opportunity to delay more serious actions by the commission. In all, the year 1974 was a very bad one for utilities, not only financially but also because the first oil crises of the 1970s had tarnished the already poor image of oil companies and electric utilities. Everywhere consumers were angry, and companies hoped that a generic hearing might last long enough for passions to cool. They knew that the normal rate case in New York lasted no more than eleven months from the calling of the first witness to the final order of the commission. But it seemed possible that a generic hearing could go on indefinitely.

adopt a spirited advocacy of
dy knew Berlin and others in the
of economists with which Kahn
in and the consulting firm had
rginal-cost case in Wisconsin.[52]
d quickly accepted Kahn's invi-
k Public Service Commission to
the initiative for the new hear-
ader base of interest than if it

could prove a mixed blessing. If
gn himself with an interested
e courting serious trouble—and
connection with the consulting
take part. Kahn met this diffi-
en more. He asked the chief ex-
seven major utilities to come to
ce. Gathering them all in one
marginalist principles, listened
ersuade them to petition as a
They agreed to his choice, and
ow the most important utility
ing environmental group to ask
airman had wished in the first

rent reasons for joining in this
ome interested in marginalist
lternatives to declining-block
ric hearing as a harmless aca-
appear cooperative. And a few
as a welcome opportunity to
mmission. In all, the year 1974
t only financially but also be-
had tarnished the already poor
utilities. Everywhere consum-
d that a generic hearing might
l. They knew that the normal
e than eleven months from the
l order of the commission. But
aring could go on indefinitely.

Kahn assumed office in the summer of 1974; his generic rate hearing began in January 1975 and ran in intermittent session for more than eighteen months. The entire written record comprised some 5000 pages of testimony and cross-examination. In all, twenty-two parties submitted briefs, many of them extremely detailed. Among numerous witnesses called to the stand were a score of experts on the theory of marginal-cost pricing, including some who thought it impractical. Kahn opened the proceedings by warning all interested parties that this hearing was not to be an exercise conducted solely for lawyers. In language again reminiscent of Adams' approach a century earlier, the chairman told the participants, "I may call for innovations in our procedures . . . I hope we can proceed as cooperative seekers of the truth, rather than as adversaries seeking an advantage over one another."[54]

So the hearing went. The chairman himself presided, and most of the sessions resembled academic seminars more than formal rate hearings. Kahn insisted on organizing groups of witnesses into panels, then lining up each panel face to face with a similar lineup of attorneys, who sat side by side at another table. The lawyers took turns asking questions. Kahn encouraged interruptions, especially his own, after the witnesses had given their presentations. As might have been expected, some of the lawyers objected to these unusual procedures. Counsel for one group of industries filed a formal motion asking that Kahn be disqualified from the hearings because of his "preconceived notions regarding the appropriate elements of rate design" and for having "departed from his quasi-judicial role as a regulator." The chairman, his challengers argued, had "indulged in extensive cross-examination" and made "extensive extrajudicial comments on the record." In other words, Kahn had failed to behave like a judge at a courtroom trial.[55]

In response to this motion, Kahn presented a statement addressed to the whole commission:

> I have adjured all parties to look upon at least the theoretical phase of the proceedings as an intellectual exercise, to treat it in the nature of a running seminar, in which our purpose was essentially to explore certain academic ideas. That is the spirit in which I have behaved whenever I have presided. I have never pretended to play the role of a passive receiver of evidence . . . I intervened freely—attempting, however, not to interrupt the train of a lawyer's cross-examination—to make certain that I understood what a witness was saying and what he was not saying, to

test his ideas, to offer hypotheses of my own in order to get his reaction . . . I propounded problems—questions of whose answers I was myself uncertain; and when the witness was uncertain about the answer I asked him, and all the others, to think about it. No witness, of whatever point of view, was ever denied an opportunity to express his opinion, as fully as he wished.[56]

As expected, the commission ruled against the motion to disqualify Kahn. Certainly, he did not feel at all defensive about his behavior. He could not have been happier about the way the hearings went, especially since publicity remained favorable throughout. In a remarkable parallel development, Edward Berlin of the Environmental Defense Fund was appointed to the commission a few months after the hearings began. Berlin excused himself from further participation and took no official role in the hearings. Kahn, on the other hand, not only continued to preside but also evangelized about the proceedings in many speeches to outside groups. During one address to a meeting of investment analysts, he confessed, "I am having the time of my life running that seminar."[57]

The generic hearing did not convince all parties about the benefits of marginal-cost rate structures. But it convinced many. The final lineup for and against can be summarized.[58]

Strongly affirmative
Long Island Lighting Company
Environmental Defense Fund
Federal Energy Administration
City of New York
Chemung County Neighborhood Legal Services

Affirmative but less strongly so
Consolidated Edison Company
Central Hudson Company (utility)
New York State Electric and Gas Co.
Orange and Rockland Co. (utility)
Public Service Commission Staff
State Consumer Protection Board

Strongly opposed
Airco (large industrial user of electricity)
General Services Administration (federal agency)
Industrial Power Consumers Conference
Multiple Intervenors (industrial customers' group)
Niagara-Mohawk (large utility)
Rockland County Industrial Energy Users Association

Although Kahn had obviously not won all the parties to his side, he had achieved the next best thing: he had divided the potential opponents. What might have been a united front of utilities standing against his proposals had been split into three distinct groups. At

this point, the only adamant opposition to marginal principles came from large industrial and commercial users of electricity, who feared paying higher rates under time-of-day pricing schemes. Otherwise, Kahn had begun to teach his lessons to an admiring and sometimes enthusiastic group of listeners. Even so, his work had only started.

Installing New Rate Structures

The hearing on marginal-cost principles opened the way to actions on several fronts at once. Choosing their arenas carefully, Kahn and his fellow commissioners moved to put their ideas into effect as quickly as they could, for at least one large electric utility. This utility would be a model for their demonstration. All other power companies, the commission ordered, must prepare and submit plans for implementing similar rates. The model chosen for electric rate reform, and the first major utility to file under the new plan, was the Long Island Lighting Company (LILCO). As indicated in the list above, LILCO strongly supported Kahn's marginalist position. The company knew it needed help; it had a poor load factor, a heavy burden of debt, and a rapidly growing demand for its electricity during the summer months. With both company and commission anxious to institute marginal-cost-based rates, no obstacle appeared to stand in the way of a quick reform of the declining-block structure.[59]

While he recognized these advantages, Kahn remained cautious. A single mistake in implementing the plan would give support to his opponents among the large industrial users. As he mulled over this problem in one of many brainstorming sessions with his staff, Kahn came up with the idea of giving a financial guarantee to all customer classes, large industrial users included: a solid reassurance that, *as a class*, no group would pay more under the new system than they had under the old. In this way, businesses would not have to subsidize householders or vice versa. All necessary adjustments would be confined to shifts within (and not between) customer classes.[60]

While the lawyers and rate engineers on the commission staff regarded the chairman's proposal as a stroke of political genius, the economists protested that such a guarantee would require a

significant departure from pure marginalist principles. But Kahn knew that the commission needed good will to nurture its mandate; it must take care not to create additional enemies. A guarantee of no increase for any class of customer meant that no solid phalanx of opposition to the commission would develop, and "rate reform" would not be used as a political vehicle to soak business. Of course, some of the industrial users would pay more under the new system, but at the same time others would pay less; the winners and losers could not be expected to combine in a common cause against marginalism. With these anticipations, Kahn decided to move ahead, and his staff economists reluctantly agreed.[61]

The chairman presented the new commission plan to about forty industrialists at another meeting in his Albany office. This time many members of the group arrived in an angry mood. Several were already on record as having demanded Kahn's impeachment. As assistant to the chairman recalls the scene:

> Fred's handling of that meeting was simply a delight to watch . . . Somebody asked whether or not there could be other values than efficiency taken into account in setting prices, like fairness. Fred said something to the effect that we've got to set prices some way, and that setting them in accordance with cost seems the most logical way to do it. If you want to argue for something else you're going to lose. He said that—he, simply with fist in hand, said [to the business customers] that—"you're going to lose." Then he posited the question, "Why?" and answered, "because if prices are set in some fashion other than on the basis of cost, then it's the political process which will set prices and the political process asks who will gain and who will lose. And let's see who's going to gain? Schools will gain, hospitals will gain, the small business user will gain, and the residential customer will gain. And who will lose? You are going to lose . . . it's not the little guy but the industrialists that they are going to sock it to . . . you ought to stand up and cheer this proposal, because it is the only thing that will save you."[62]

Kahn's mixture of logic with unsettling prediction drove his points home, one after another. Most of the industrialists caved in.

The new LILCO rates represented a compromise between pure marginalist principles and political reality. As a model, this structure proved practical and enlightened. Marginalist thinking was reflected powerfully in time-of-day pricing differentials that penalized summer peak users and rewarded off-peak consumers. Here the chronic marginalist problem of precise measurement

marginalist principles. But Kahn
ded good will to nurture its man-
eate additional enemies. A guaran-
of customer meant that no solid
mission would develop, and "rate
political vehicle to soak business.
users would pay more under the
e others would pay less; the win-
pected to combine in a common
h these anticipations, Kahn de-
f economists reluctantly agreed.[61]
new commission plan to about
eeting in his Albany office. This
arrived in an angry mood. Sev-
ving demanded Kahn's impeach-
n recalls the scene:

simply a delight to watch . . . Some-
uld be other values than efficiency
ike fairness. Fred said something to
some way, and that setting them in
logical way to do it. If you want to
g to lose. He said that—he, simply
customers] that—"you're going to
'Why?" and answered, "because if
an on the basis of cost, then it's the
and the political process asks who
see who's going to gain? Schools
ll business user will gain, and the
ho will lose? You are going to lose
rialists that they are going to sock
eer this proposal, because it is the

ing prediction drove his points
ne industrialists caved in.
d a compromise between pure
reality. As a model, this struc-
ned. Marginalist thinking was
ay pricing differentials that
rewarded off-peak consumers.
lem of precise measurement

(which came down to a problem of installing meters that would record usage by time of day) seemed to be solved by applying the new rates at first only to LILCO's 750 largest consumers, all of them business firms. These customers could afford the cost of new meters, which ranged from $400 to $4000. Within a few years, time-of-day rates had been applied to many other customers, and the price spread between peak and off-peak had reached a ratio of twelve to one (30¢ per kilowatt-hour on a hot summer day versus 2½¢ on a winter night). Kahn's marginalist approach had been vindicated by evidence drawn from real experience.[63]

The LILCO case proved to be a landmark in the history of regulation in the United States. For the first time, after an exhaustive generic hearing followed by a brief rate case involving a single company, marginal-cost principles had been put into operation on a large scale. Chairman Kahn deserved the credit. Working assiduously over a two-year period, he artfully shaped a political consensus among unusual bedfellows: environmentalists, electric utilities, consumer groups, business firms, and residential customers. At the same time, he neutralized all significant opposition to his own strategy. Much like James Landis at the SEC, Kahn deliberately modified his thinking in order to get the policy exactly right. But then he moved on with equal care to the issue of implementation. Thinking perceptively about the nature of the opposition both inside and outside the commission, Kahn isolated the interest of every opponent in turn. One by one, he enlisted each of them in his cause, winning now by force of intellect, now by charm and cajolery, now by political compromise. It was nothing less than a tour de force of regulatory strategy.

Still, a strategy for electric utilities represented only one part of Kahn's program. Even in the midst of the generic rate hearing, the commission found time to deal with the perennial problems of Consolidated Edison, as they ordered the elimination of declining-block rate structures for that company's millions of residential customers in New York City. Simultaneously, Kahn and his colleagues demanded that all electric utilities in the state prove that their declining-block rates were cost-justified. The commission next forbade the inclusion of gas and electric charges in rental payments for new apartments. Kahn's reasoning here was that, when tenants paid lump-sum rents that included utility bills, they would have no incentive to conserve. Given the psychology of landlord-tenant

relations, some renters actually delighted in wasting energy at their landlords' expense. Under the new rules, tenants would pay utility bills separately from rents. Finally, at Kahn's behest, the commission instructed several utilities to submit their practices to systematic review, done under contract by management-consulting firms. Here the purpose was to offset the effects of the cost-plus philosophy that still permeated companies under traditional rate regulation. If rates were to be based on costs, it was important for both the utilities and the commission to understand as fully as possible the nature and extent of costs. Since the companies seemed unable to explain their costs adequately and the commission lacked the necessary staff, both needed help. Again, Kahn's strategy resembled Landis' with the SEC: where Landis used accountants to gather essential data, Kahn used consulting companies. The approach opened the way for a broader application of Kahn's regulatory philosophy.[64]

New York Telephone

For many New Yorkers, the most vivid illustration of Kahn's approach to regulation came from his two highly publicized battles over telephone rates. In one, he fought against the New York Telephone Company, trying to alter its rate structure. The other battle pitted him against consumer groups, on the issue of charges for directory assistance.

Throughout most of his career, Kahn displayed as much interest in telecommunications as in any other area of regulation. He served on the economic advisory board of American Telephone and Telegraph, where he learned that the company made a policy of bundling its charges. That is, the Bell System, partly in keeping with its slogan "The System Is the Solution," did not like to publish cost data on individual parts of its business: long distance, local telephone service, Bell Labs, Western Electric, Yellow Pages advertising, and equipment rentals. Instead, Bell preferred to price its services as a "bundle," without direct reference to their individual costs. This practice offered the company wide discretion in deciding which segments of the business would subsidize other segments. Bundling provided a valuable management tool because it permitted Bell executives to raise or lower long-distance and local-service rates as market opportunities or political

ally delighted in wasting energy at
er the new rules, tenants would pay
rents. Finally, at Kahn's behest, the
utilities to submit their practices to
r contract by management-consult-
s to offset the effects of the cost-plus
d companies under traditional rate
based on costs, it was important for
nmission to understand as fully as
nt of costs. Since the companies
costs adequately and the commis-
both needed help. Again, Kahn's
h the SEC: where Landis used ac-
data, Kahn used consulting com-
ne way for a broader application of

st vivid illustration of Kahn's ap-
his two highly publicized battles
ought against the New York Tele-
its rate structure. The other battle
ups, on the issue of charges for di-

Kahn displayed as much interest
y other area of regulation. He
board of American Telephone
that the company made a policy
ne Bell System, partly in keeping
e Solution," did not like to pub-
s of its business: long distance,
Western Electric, Yellow Pages
als. Instead, Bell preferred to
ithout direct reference to their
ered the company wide discre-
of the business would subsidize
d a valuable management tool
ves to raise or lower long-dis-
arket opportunities or political

pressures required. For example, Bell often held down basic monthly charges to residential users—a politically popular thing to do—while subsidizing the resulting revenue shortfalls with excessive long-distance fees. Bundling did not, however, always work to the advantage of masses of consumers. Kahn had reason to believe that New York Telephone was using the practice to subsidize its business customers at the expense of residential users, specifically by undercharging for business switchboard services.

To remedy the many abuses associated with bundling, Kahn decided to take direct action. Accordingly, when he and his commission colleagues approved a telephone rate increase in 1975, they also ordered the company to prepare detailed, unbundled cost and distribution studies before its next application. When that application, which included a $393 million rate hike, did come in 1976, no such studies were included. The commission rejected the application outright, without even taking up the merits of New York Telephone's case. The company had been forewarned, said Kahn, that it must produce the cost data. "But they didn't give that to us. Why? Because the tradition that this was management's prerogative is very strong and . . . they have never done good cost studies. They want to be free to set their rates the way they want. It just goes to show how sloppy rate-making has been in the past—totally permissive to the companies. We just aren't going to permit that any more."[65]

Though critical of management's abuses, Kahn did not speak as a special-interest consumer advocate determined to punish companies and reduce prices without any reference to costs. He insisted that consumers pay their share, especially for directory assistance, rather than receiving it free from New York Telephone. As an economist, he knew that nothing was free. Directory assisitance, in fact, could not readily be automated, and the obligation to give information tied up hundreds of operators. The service was costing New York Telephone about $100 million a year, and the company simply passed along those charges to the consumer, indiscriminately across its entire list of subscribers. Acting on the fundamental principle that every individual must pay for what he receives, Kahn insisted that "free" directory assistance had to stop.[66]

Consumers responded about as he had expected: they were loud, hostile, and abusive. Kahn did not waver. New York Tele-

phone and the commission together worked out a compromise formula under which each caller who dialed Information would be charged 10¢. All customers, however, would begin each month with three free calls. If these calls were not used, the company agreed to credit each nonuser with a 30¢ rebate. While the amounts may seem tiny, the results proved that the issue was not trivial. Within the first two months under the new system, the volume of directory-assistance calls in New York declined by more than 40 percent. Eighty-five percent of residential customers and 70 percent of the business consumers received the full 30¢ per month rebate. Total saving to the company (and the consumer) amounted to an impressive $25 million per year.[67]

Even so, Kahn was not satisfied. He wished to push the company and the commission further, toward a complete marginal-cost pricing system. Under Kahn's plan, which took advantage of the sophisticated computer systems already installed at New York Telephone, callers would pay more during peak phoning periods, less during off-peak. They would also pay for distance—more for calls stretching across five miles than across only one. The new minimum charge for intrastate toll calls would entitle a caller to talk for two minutes, not three. For all telephones—residence and business—conversations of long duration would be more expensive than brief chats. Kahn's system was full of innovations. Like the directory-assistance proposal, this new plan did not win universal acclaim. One official of the American Telephone Consumers Council denounced it as "the beginning of the end of the consumers' right to rap in New York State . . . Telephone users will be running up charges like a taxi meter." Not even Kahn could have put it better. Should a twenty-mile taxi trip cost the same as one of two blocks? Again he showed no inclination to retreat: "It is our intention (and the Company's) eventually to go to a full-blown 'Message-Minute-Mile' pricing system, according to which charges per call would vary according to the marginal costs of merely setting up a call, of minutes of additional conversation time, and of distance." The chairman's voice sounded the same, on or off the telephone.[68]

After almost three years of favorable media attention in Albany, Kahn enjoyed a national reputation as an innovative regulator. During the same three-year period, old regulatory issues had sur-

gether worked out a compromise
er who dialed Information would be
owever, would begin each month
calls were not used, the company
r with a 30¢ rebate. While the
sults proved that the issue was not
ths under the new system, the vol-
ls in New York declined by more
rcent of residential customers and
sumers received the full 30¢ per
the company (and the consumer)
million per year.[67]
d. He wished to push the company
ward a complete marginal-cost
lan, which took advantage of the
lready installed at New York Tel-
during peak phoning periods, less
pay for distance—more for calls
across only one. The new mini-
lls would entitle a caller to talk
all telephones—residence and
duration would be more expen-
em was full of innovations. Like
this new plan did not win uni-
American Telephone Consumers
nning of the end of the consum-
te . . . Telephone users will be
er." Not even Kahn could have
taxi trip cost the same as one of
nclination to retreat: "It is our
entually to go to a full-blown
system, according to which
rding to the marginal costs of
tes of additional conversation
n's voice sounded the same, on

le media attention in Albany,
n as an innovative regulator.
old regulatory issues had sur-

faced in a new way—to become a national preoccupation. This happened primarily because of the oil crisis of 1973–74, the rise of inflation, and the deep economic recession of 1974–75. These events focused attention on the cost of almost all goods and services, and consequently on the price the nation was paying for its regulatory goals. Regulation suddenly assumed a political importance it had not attained since the Progressive Era. During the presidential campaign of 1976, candidates Gerald Ford and Jimmy Carter both fixed on regulatory reform as a central concern for the next administration. When Carter won the election, it was a forgone conclusion that many new faces would appear in federal regulatory agencies.

For himself, Kahn would have preferred to serve on the Federal Communications Commission, where he might have continued his crusade for marginalist principles in telecommunication pricing. Carter had other ideas: he wanted Kahn for the Civil Aeronautics Board, where some momentum for deregulation seemed to be underway but in danger of failing. Kahn first refused Carter's offer of appointment, pleading ignorance of the airline industry. But after a second invitation, which included specific assurances of the president's full backing, he accepted the job. His years on the New York Public Service Commission had provided ample recommendation for Kahn's new federal office. But now he had to learn his way in yet another complicated industry.[69]

The Airline Industry

The American air age began in 1903, with the Wright brothers' flight at Kitty Hawk. But it was Charles A. Lindbergh's solo trip across the Atlantic in 1927 that first made Americans aware of the possibilities in commercial aviation.[70] Airline companies, which had transported only 6,000 passengers in the entire year before Lindbergh's flight, carried 48,000 in 1928, 162,000 in 1929, and 383,000 in 1930. Not even the Great Depression halted the growth of the fledgling industry. By 1938, the year in which Congress created the Civil Aeronautics Authority (later Board), passenger traffic surpassed one million. Then it reached almost four million in 1941, when World War II temporarily set back further growth by giving priority to military demands.[71] Equally, though, the war reinforced the glamour of flying and helped to make the federal

government more aware of both its potentials and its risks. As in the earlier history of the railroads, airline accidents were often spectacular, and deaths became commonplace. Safety regulation thus presented an obvious arena for government action.[72]

Commercial aviation, however, was a very different kind of industry from railroading. The business was relatively cheap and easy to enter; all anyone needed was an airplane, a few pilots, and a ticket office. So airlines could not be considered natural monopolies in the fashion of early railroads or of gas and electric utilities. In theory, dozens or even hundreds of companies could compete for passengers and freight. This ease of entry played a large role in the history of airline regulation. It forced the Civil Aeronautics Board into the dilemma of either maintaining an orderly industry with a small number of participants or allowing market forces to shake out a multitude of competitors, with only the fittest eventually surviving. By the 1930s, some competitors of the fly-by-night variety had already given that term its literal meaning. In an industry posing grave threats to the passenger's life and limb, the job of regulation seemed difficult and messy.[73]

Nor was the market for airlines so simple a thing as, say, the market for lightbulbs. Airline companies sold something like a trip—really the promise of a guaranteed seat of a certain class, on a scheduled flight between "city-pairs." Each pair of cities represented a separate market, with customers in each, and the airline market became segmented in several different ways. Some companies specialized within single states (Texas International, Air Florida); some in regions (Ozark, Piedmont, Pacific Southwest); some in hub-and-spoke networks serving national markets (American, United, Delta). Others catered to charter traffic, leasing individual airplanes to travel groups by the trip or month. Of course, none of these details made the airline industry easy to understand, much less regulate. By the time Alfred Kahn came to the Civil Aeronautics Board, most of the significant markets were served by ten major scheduled airlines. Together, these ten flew to about 430 cities with more than 90,000 possible city-pairs, of which about 58,000 actually received sheduled service.[74]

In another way, the airlines were characterized by unusual volatility. Because air travel is expensive and because a large percentage of potential passengers are vacationers—and thus subject to changes of mind about the trip when money is short—the airline

th its potentials and its risks. As in
oads, airline accidents were often
e commonplace. Safety regulation
a for government action.[72]
er, was a very different kind of in-
usiness was relatively cheap and
l was an airplane, a few pilots, and
not be considered natural monop-
oads or of gas and electric utilities.
reds of companies could compete
ease of entry played a large role in
. It forced the Civil Aeronautics
r maintaining an orderly industry
ants or allowing market forces to
titors, with only the fittest even-
some competitors of the fly-by-
nat term its literal meaning. In an
the passenger's life and limb, the
and messy.[73]
es so simple a thing as, say, the
mpanies sold something like a
anteed seat of a certain class, on
pairs." Each pair of cities repre-
stomers in each, and the airline
eral different ways. Some com-
states (Texas International, Air
Piedmont, Pacific Southwest);
erving national markets (Ameri-
l to charter traffic, leasing indi-
y the trip or month. Of course,
ne industry easy to understand,
ed Kahn came to the Civil Aer-
ant markets were served by ten
, these ten flew to about 430
ble city-pairs, of which about
ervice.[74]
characterized by unusual vola-
e and because a large percent-
ationers—and thus subject to
n money is short—the airline

business has always had violent ups and downs. It remains among the most cyclical of industries. The historical fluctuations in the prices of airline securities on the New York Stock Exchange illustrate this characteristic: in the early 1960s, the average price of a share of airline common stock stood at about $5. The subsequent national economic boom, however, quickly drove prices up by a factor of nine—to an average of $47 a share in 1966. By the recession year of 1970, the price had dropped back down to $13.[75]

Nonetheless, the salient trend of the industry from the 1930s to 1977 was its enormous growth. Every indicator displayed the same message.[76]

Year	*Persons employed*	*Number of aircraft*	*Seats per aircraft*	*Revenues ($thousand)*	*Revenue (passenger-miles)*
1938	13,274	311	14.2	$57,997	0.533 mil.
1958	152,510	1,899	57.0	2,243,964	31,499 mil.
1977	310,674	2,420	148.1	19,212,228	187,681 mil.

The Civil Aeronautics Board

By 1977, when Kahn became chairman of the CAB, the airline industry was still growing. But it had also developed a number of chronic problems, chief among them a depressed load factor. For all flights by all major airlines in 1977, the composite load factor stood at only 55.5 percent, which meant that on average each plane was flying a little more than half full. Critics of airline regulation contended that this unsatisfactory performance could be traced directly to the policies of the CAB. To explain how this had happened, they cited the unusual circumstances in which the board had been created.[77]

The first important attempts to bring order to the industry began during the 1920s, under the stimulus of Secretary of Commerce Herbert Hoover. Safety rules were legislated in 1926, and Congress began to appropriate federal subsidies to airlines for the carrying of mail. By 1933, the postmaster general's administration of these subsidies had led to a concentration of the industry in a handful of important firms. For reasons of its own convenience, the Post Office placed nearly all of its twenty airmail contracts with three holding companies, and these three became the ancestors of today's Big Four airlines: United, American, Eastern, and

TWA. Because of this concentration, and more particularly because of the apparent evasion of competitive bidding requirements, the Senate conducted an extensive investigation which turned up such a mass of damning evidence that President Roosevelt canceled all airmail contracts. Roosevelt said that the Army Air Corps, not commercial airlines, would now carry the mail. This too proved unworkable, and in 1934 a new round of bidding began. Again the three holding companies walked off with most of the lucrative routes. Meanwhile, competition for shorter routes became very keen, and a consensus in Congress began to develop that the industry needed a comprehensive system of regulation for the management of all passenger-fare levels, airmail subsidies, and route structures. In 1934, the first New Deal bills for economic regulation of airlines were introduced, and in 1938 Congress passed the Civil Aeronautics Act.[78]

The principles expressed in the act of 1938 closely resembled those embodied in the Act to Regulate Commerce of 1887 and the Motor Carrier Act of 1935, legislation that had begun the economic regulation of railroads and trucking. In fact, had the Interstate Commerce Commission itself not been overburdened with these other two industries, Congress might well have vested the ICC with airline regulation as well. One congressman complained that the government was taking a big chance in starting a new agency. "Once you establish a commission . . . you have the devil's own time passing any act abolishing it."[79]

A brief report on the new legislation, issued by the Senate Committee on Commerce, explained that the bill provided for "the usual system of economic regulation," according to "the recognized and accepted principles of the regulation of public utilities, as applied to other forms of transportation." This language suggested that the committee regarded airlines, like railroads or electric utilities, as a natural-monopoly industry. Thus a mistaken economic assumption seems to have formed the foundation of the bill. The Senate report went on to say that "competition among air carriers is being carried to an extreme."[80] In this judgment the House report agreed. Instead of being awarded on the basis of ability to serve, said the House, airline routes went to the lowest bidder for air-mail contracts. As a result, "the air carriers, in their desire to secure the right to carry the mail over a new route, have made absurdly low bids, indeed, have virtually evinced a willing-

tration, and more particularly be-
n of competitive bidding require-
an extensive investigation which
ing evidence that President Roose-
acts. Roosevelt said that the Army
lines, would now carry the mail.
nd in 1934 a new round of bidding
companies walked off with most of
e, competition for shorter routes
asus in Congress began to develop
rehensive system of regulation for
-fare levels, airmail subsidies, and
irst New Deal bills for economic
roduced, and in 1938 Congress
t.[78]

ne act of 1938 closely resembled
gulate Commerce of 1887 and the
slation that had begun the eco-
l trucking. In fact, had the Inter-
elf not been overburdened with
ress might well have vested the
ll. One congressman complained
a big chance in starting a new
nmission . . . you have the devil's
ng it."[79]

gislation, issued by the Senate
ined that the bill provided for
ulation," according to "the rec-
f the regulation of public utili-
transportation." This language
arded airlines, like railroads or
poly industry. Thus a mistaken
e formed the foundation of the
say that "competition among
xtreme."[80] In this judgment the
ng awarded on the basis of abil-
routes went to the lowest bid-
sult, "the air carriers, in their
ne mail over a new route, have
ve virtually evinced a willing-

ness to pay for the privilege of carrying the mail over a particular route." In addition, the House committee was told by the president of the Air Transport Association (the industry's trade group) that $120 million of private investment funds had been poured into the existing system, and that half of that sum had simply been lost. "He further testified," said the committee report, "that unless legislation is enacted which would give the carriers reasonable assurance of the permanency of their operation and would protect them from cutthroat competition, a number of the air lines would soon be in serious financial trouble." Prospects for the industry seemed bleak.[81]

Clearly, in passing the Civil Aeronautics Act, Congress intended to bring stability to airlines. What is not clear is whether the legislature intended to cartelize the industry. Yet this did happen. During the forty years between passage of the act of 1938 and the appointment of Kahn to the CAB chairmanship, the overall effect of board policies tended to freeze the industry more or less in its configuration of 1938. One policy, for example, forbade price competition. Instead the CAB ordinarily required that all carriers flying a certain route charge the same rates for the same class of customer. This rule conformed to a basic principle of public-utility regulation, and, however odd the policy appeared later on, it flowed logically from the concerns expressed by Congress in 1938. A second policy had to do with the CAB's stance toward the entry of new companies into the business. Charged by Congress with the duty of ascertaining whether or not "the public interest, convenience, and necessity" mandated that new carriers should receive a certificate to operate, the board often ruled simply that no applicant met these tests. In fact, over the entire history of the CAB, no new trunkline carrier had been permitted to join the sixteen that existed in 1938. And those sixteen, later reduced to ten by a series of mergers, still dominated the industry in the 1970s. All these companies participated in the immense growth shown in the table above. All developed into large companies under the protective wing of the CAB. None wanted deregulation.[82]

What the airlines did want was new routes. And since the CAB declined to permit either rate competition or new entry, its regulatory activities came to focus on the question of which airlines would receive new routes. At the same time, the board exercised responsibility for the development of sound route structures for all

companies. This meant that profitable long hauls must be mixed with less lucrative short hauls, and that only a few companies—sometimes just one—should be allowed to fly any particular route: that is, a monopoly in some city-pair markets would be necessary to compensate the airline for other, unprofitable routes it was required to fly elsewhere.

To maintain this complex balancing act, the board was ultimately drawn into the adjudication of trivial decisions at the lowest levels of airline management. The CAB came to function more and more along the lines of a court, and its members began to think of themselves as judges, not as proponents of economic efficiency.[83] James Landis treated this problem in his 1960 report to Kennedy, citing inordinate delays as a characteristic of the CAB's proceedings. Landis went on to denounce the agency's failure to do "forward planning of the type necessary to promote our air commerce to its desired level of efficiency" and added that the Board's regulatory efforts were frustrated by a courtroom-style routine that prevented sensible resolution of issues.[84] The board's usual pattern of lengthy cross-examination, Landis concluded, was "a wasteful manner of establishing many of the basic facts." In making awards, the board "selects a route pattern for reasons that it frequently cannot even articulate." In contrast to the SEC of Landis' own design, the CAB simply had no strategy.[85]

The particular questions the board took upon itself to decide tended to preclude the development of an overall strategy. What difference did it make to public policy whether the third or fourth carrier selected to serve a particular city-pair was Eastern, Delta, or American? On what basis should such a decision be made? There was none, other than giving some desirable route to a carrier as compensation for an undesirable one. In such circumstances of arbitrariness the CAB, in writing its decisions, could rationalize its own cartel-like policies only through arbitrary lines of reasoning. Given the emphasis on procedure so characteristic of American administrative law, together with the wide substantive discretion usually enjoyed by regulatory agencies, almost any line of reasoning was likely to be upheld by appellate courts, provided only that the CAB's procedures had been according to form. This meant that the CAB could do almost anything it wanted so long as it went through the proper motions. By contrast, should it actually articulate its practice of cartelization, or openly acknowledge its

rofitable long hauls must be mixed
, and that only a few companies—
allowed to fly any particular route:
y-pair markets would be necessary
ther, unprofitable routes it was re-

alancing act, the board was ulti-
cation of trivial decisions at the
ment. The CAB came to function
of a court, and its members began
not as proponents of economic ef-
this problem in his 1960 report to
ys as a characteristic of the CAB's
denounce the agency's failure to
pe necessary to promote our air
f efficiency" and added that the
frustrated by a courtroom-style
esolution of issues.[84] The board's
mination, Landis concluded, was
ing many of the basic facts." In
s a route pattern for reasons that
late." In contrast to the SEC of
ply had no strategy.[85]
oard took upon itself to decide
ent of an overall strategy. What
olicy whether the third or fourth
lar city-pair was Eastern, Delta,
uld such a decision be made?
g some desirable route to a car-
able one. In such circumstances
g its decisions, could rationalize
rough arbitrary lines of reason-
lure so characteristic of Ameri-
r with the wide substantive
atory agencies, almost any line
d by appellate courts, provided
d been according to form. This
t anything it wanted so long as
By contrast, should it actually
on, or openly acknowledge its

implicit compensatory policies, then substantive challenges might arise and threaten the board's very foundation.[86]

Still another difficulty grew from mismatch between the CAB's courtlike atmosphere and the issue of "*ex parte* considerations." Whereas conventional judges must not consort with plaintiffs and defendants, regulatory commissioners sometimes had to make nonlegal types of decisions, especially on economic matters. If CAB members could not converse with airline executives, how would they learn about the industry? One of Kahn's colleagues on the CAB, himself a former judge, put the dilemma this way: "In some respects, we are held to the same rigid standards that a court is held, with regard to making decisions, and yet, we are determining policy in which all sorts of inputs should be received before you can make very much of a quality decision, and where I can obtain that it is a bit of a problem without some contact with the industry."[87] Over the years, several board members had resigned, citing frustration caused by the mixed roles they were expected to play. Yet, in this awkward situation, legal experience proved invaluable. If the primary role of the CAB was to be not the framing of a regulatory strategy but rather the assurance of a fair hearing and complete record, then legal procedures seemed natural as the stock in trade of the board. And lawyers did dominate its membership: of the thirty-four members who served between 1938 and the mid-1970s, twenty-four were trained in law. The second-place profession, engineering, accounted for only four.[88]

The Ferment of Reform at the CAB

When Alfred Kahn came to the CAB in 1977, change was in the air. Academic studies had long since established beyond much doubt that an airline cartel existed in America, as a result of CAB policies. These studies argued convincingly that, by restricting economic competition, the board encouraged greater competition in frills: fancier in-flight food and drink, more movies, additional attendants—all minor forms of service to the customer. Much more important, the board's policies also induced the airlines to schedule too many flights. This happened because, once a carrier had been "certificated" by the board to fly between a pair of cities, it was left free to fly as many trips along that route as it

wished. Since airline managers held as an article of faith that travelers gravitated to the company with the most departures at the most convenient times, competing airlines scheduled more and more flights. With the number of customers constant, they were forced to fly emptier airplanes. As a result, the industry drifted into its woeful load factor of 55.5 percent. For several years the CAB had responded to this situation by imposing an implicit route moratorium. That is, in addition to forbidding rate competition and denying all applications for entry into the industry, the board also refused to grant any new routes to the existing carriers. The route moratorium had loosened up a bit just before Kahn's arrival in Washington, but the practice of restricting new routes lingered as evidence of the board's single-minded opposition to competition.[89]

By the mid-1970s, several major airlines were in serious financial trouble. Much as in the electric utility industry, a series of converging problems—the skyrocketing price of fuel, the economic recession, and an expensive inventory of wide-body aircraft (Boeing 747s, Lockheed L-1011s, Douglas DC-10s)—had caught the airlines in a trap of rising costs and declining revenues. The CAB's inflexibility prevented rapid adjustment by the companies themselves. It was not that the board had caused all these problems, but rather that its opposition to any real competition made it impossible for the industry to adjust through normal market mechanisms.[90]

Public indictments of the CAB began to grow. In 1975, a subcommittee chaired by Senator Edward M. Kennedy subjected the industry and the board to searching scrutiny. The stage manager of the Kennedy hearings was Stephen Breyer, then a professor at Harvard Law School and later a federal circuit judge. A specialist in regulation and a man of vivid imagination and agile mind, Breyer wanted to demonstrate in a public arena that the CAB form of regulation did violence not only to the principles of competition but even to the administrative process it seemed designed to serve. Senator Kennedy, to whom Breyer fed dozens of "zinger" questions for a lineup of witnesses from the board and the industry, responded with a virtuoso performance. It was Kennedy's finest hour as a legislator, and his participation guaranteed the hearings wide publicity.[91] The clincher in the case against the CAB proved to be the example of airlines that operated intrastate

held as an article of faith that trav-
y with the most departures at the
eting airlines scheduled more and
of customers constant, they were
. As a result, the industry drifted
5.5 percent. For several years the
tion by imposing an implicit route
on to forbidding rate competition
entry into the industry, the board
outes to the existing carriers. The
up a bit just before Kahn's arrival
of restricting new routes lingered
e-minded opposition to competi-

jor airlines were in serious finan-
ctric utility industry, a series of
ocketing price of fuel, the eco-
e inventory of wide-body aircraft
, Douglas DC-10s)—had caught
sts and declining revenues. The
id adjustment by the companies
oard had caused all these prob-
n to any real competition made it
st through normal market mech-

began to grow. In 1975, a sub-
ward M. Kennedy subjected the
ng scrutiny. The stage manager
hen Breyer, then a professor at
ederal circuit judge. A specialist
d imagination and agile mind,
a public arena that the CAB
t only to the principles of com-
tive process it seemed designed
n Breyer fed dozens of "zinger"
from the board and the indus-
ormance. It was Kennedy's fin-
participation guaranteed the
ncher in the case against the
irlines that operated intrastate

routes beyond board jurisdiction. Two companies in particular, Southwest Airlines in Texas and Pacific Southwest in California, had developed successful business strategies that combined extremely low fares with high load factors. Breyer and Kennedy demonstrated startling differences between the fares customers paid in 1974 in city-pair markets regulated by the CAB and those in Texas and California, where price competition was permitted.[92]

City–pair	*Miles*	*Fare*	*No. of passengers*
Los Angeles–San Francisco	338	$18.75	7,483,419
Chicago–Minneapolis (CAB)	339	$38.89	1,424,621
New York–Pittsburgh (CAB)	335	$37.96	975,344
Houston–San Antonio	191	$13.89	490,000
Boston–New York (CAB)	191	$24.07	2,493,882
Reno–San Francisco (CAB)	192	$25.93	312,811

The Texas and California airlines achieved these remarkable results by offering numerous flights at cut rates. They started with the assumption that the market for air travel was much bigger than the trunklines or the CAB imagined—that the barrier to increased traffic was simply the high price of an airline ticket. They eliminated first-class accommodations and crowded as many additional seats into the cabins as possible without seriously interfering with passengers' comfort. As the Kennedy subcommittee pointed out, Pacific Southwest Airlines "puts 158 seats in a Boeing 727-200 jet aircraft and fills approximately 60 percent of these seats on average. American Airlines puts 121 seats in the same plane and flies it on average 55 percent full. Thus, when flying PSA, 95 passengers must share the cost of flying the airplane, while, on an American plane, 66 passengers must share the same cost." Obviously, the price of a ticket could be cut dramatically, and the price cut in turn would bring in new categories of customers. Explaining the phenomenon to the Kennedy subcommittee, an official from the Texas Aeronautics Commission testified: "These [passengers] are people who are coming to the airline instead of going to the bus . . . you kind of wait for them to tie the chicken coop on top of the airplane."[93]

The remarkable results of the cut-rate fares in Texas and California had been available to any interested observer long before Kennedy's 1975 hearings. Kahn, in fact, included an analysis of the Pacific Southwest story in the second volume of his *Economics of*

Regulation, published in 1971. He viewed the case histories as "the closest thing to a 'controlled experiment' in public policy," and especially appreciated the marginal-cost aspects of the experiment. As Kahn knew, the cost of flying a plane between a pair of cities was more or less fixed, no matter how many passengers made the trip. Obviously the airline could not give away tickets, but it could cut rates on a trial-and-error basis—charging less, for example, for off-peak travel. In a situation of depressed load factor and industry distress, the point of marginalism was to furnish a rationale for any sort of radical price cut. Once fares went down, new travelers would be induced to fly in such numbers that the airlines might soon reach an advantageous load factor.[94]

In all, the Texas and California experiences provided powerful ammunition to the Kennedy subcommittee and to other proponents of deregulation as well. A General Accounting Office study estimated that actual fares in national airline markets exceeded hypothetical fares derived by methods similar to those in Texas and California by 22 to 52 percent. Passengers traveling under the hypothetical rates would have saved between $1.4 billion and $1.8 billion annually during the years 1969–1974. The GAO study added that the lowered fares would have persuaded so many new travelers to fly that the total saving would have been more than $2 billion a year.[95]

By the time of Kahn's arrival in Washington, a broad and strikingly mixed coalition had already begun to lobby for deregulation. One group included a number of strange bedfellows: the National Association of Manufacturers, Common Cause, Sears Roebuck, the American Association of Retired Persons, the Aviation Consumer Action Project, the National Taxpayers Union, the Cooperative League of the U.S.A., the American Conservative Union, and the Public Interest Economics Group. "There is some amount of culture shock," conceded one member of this unusual alliance, "but we do all agree on this issue." A writer in *Fortune* called the deregulation advocates "an odd coalition of academic economists, Naderite consumerists, liberal Democrats, and conservative Republicans." Diverse as this group seems, they were already gathering momentum and political power.[96]

Shortly after Kahn's appointment, the Carter Administration held a "reform gala" at the White House to underscore its commitment to deregulation. Kahn took a prominent part. Holding

He viewed the case histories as
led experiment' in public policy,"
marginal-cost aspects of the ex-
st of flying a plane between a pair
no matter how many passengers
rline could not give away tickets,
and-error basis—charging less, for
situation of depressed load factor
f marginalism was to furnish a ra-
rice cut. Once fares went down,
to fly in such numbers that the
antageous load factor.[94]

a experiences provided powerful
committee and to other propo-
General Accounting Office study
ational airline markets exceeded
ethods similar to those in Texas
t. Passengers traveling under the
ed between $1.4 billion and $1.8
rs 1969–1974. The GAO study
ld have persuaded so many new
g would have been more than $2

Washington, a broad and strik-
begun to lobby for deregulation.
strange bedfellows: the National
mmon Cause, Sears Roebuck, the
Persons, the Aviation Consumer
payers Union, the Cooperative
an Conservative Union, and the
"There is some amount of cul-
r of this unusual alliance, "but
writer in *Fortune* called the
alition of academic economists,
mocrats, and conservative Re-
seems, they were already gath-
ver.[96]

nt, the Carter Administration
House to underscore its com-
k a prominent part. Holding

forth later on to a *New York Times* reporter, he explained why the movement united so many disparate interests:

> First of all, I guess it has been this experience of the last 10 years of disappointment with the notion that you solve your problems by passing a law and then going away. The second I suspect has been the sheer proliferation of regulations. I mean, we get new regulatory agencies every year, and it does grow kind of astronomically. A third has been a kind of pendulum effect . . . So you have these real conservatives who have always said, "We don't want the Government messing in here," joining with the liberals of the Ralph Nader kind who say, "Well, where competition could work, you should leave it to competition."[97]

Of course, not everyone wanted airline deregulation. The opponents still held the upper hand, even though two years had passed since the Kennedy hearings. Kahn realized that a crucial part of his job would be to convert skeptical majorities in several constituencies, including Congress and even the CAB itself. The final measure of his success would come from what he could do within his agency and from his attempts to break the logjam on Capitol Hill, where deregulation bills had been stalled for months. During the Kennedy hearings, the airline industry had presented a solid phalanx against deregulation. Without exception, top executives had testified in favor of continued regulation, many of them in vehement tones. Now, two years later, some tiny fissures in the proregulation front had begun to appear, in the representatives of Hughes Airwest and Frontier Airlines; and mighty United Airlines was showing signs of wavering. But a proregulation majority controlled the industry.

Ironies abounded. Delta, the most profitable of all airlines and by common agreement one of the nation's best-managed corporations, liked to profess from its Atlanta headquarters a free-enterprise, antiunion philosophy. Yet, in Washington, Delta remained the most vocal proponent of regulation. Its outspoken chairman denounced the deregulation movement as a wolf in sheep's clothing. Under the banner of "free enterprise" and "less government regulation," he said, the movement actually promised more regulation and perhaps eventual nationalization."[98] Industry trade papers offered equally conspiratorial theories. One journal outlined its answer to the "most mystifying question" of "why air transport regulation, eminently successful by any yardstick, has been singled out for sustained attack." A two-stage plan, said the

paper, had been hatched in the executive branch. First, deregulation would get the independent CAB out of the picture. The way would then be clear for manipulation by operatives from the White House and Department of Transportation. Yet in point of fact, the Department of Transportation stood on the side of the airlines. Secretary Brock Adams was an energetic advocate of regulation, and later on he became a problem for Kahn when the two men battled for influence over national transportation policy.[99]

During Kahn's early months as CAB chairman, his most difficult tasks lay within the board itself and with Congress. In the House, the chief obstacle to Kahn's program was Democratic Representative Elliott Levitas of Atlanta, who held a key committee post. (Journalists, mindful of Levitas' services to his Delta Airlines constituents, referred to him as a "Deltacrat.") Other congressmen also responded to the intense proregulation stance of the airline industry.[100] As things happened, the most solid support of continued tight regulation came not from airlines but from labor unions; and the unions put even heavier pressure on their representatives to shield them from deregulation. Airline employees, assisted by the cost-plus thinking and protectionist policies of the CAB, had come to receive exceptionally high pay. In 1976, trunkline workers—including all categories from baggage handlers to pilots—earned an average wage of $19,000, which put them at the very top of all employee groups in the United States. Not surprisingly, then, the industry's labor force, led by the Air Line Pilots' Association and including the unions of flight engineers, transport workers, machinists, and aerospace workers, conducted an active campaign against deregulation. Their well-publicized power in Washington presented Kahn with a problem unlike any he had encountered at the New York Public Service Commission.[101]

The Early Strategy

At first Kahn felt overwhelmed by the complex task of moving the airlines toward deregulation while at the same time lobbying Congress for new legislation.[102] But, as in New York, he quickly learned his way. One of the first things he did was to lay down the law to the CAB about the language it used in orders and correspondence. Over the years, the agency had involved itself in so many petty details, so many arbitrary decisions, that it seemed

executive branch. First, deregula-
CAB out of the picture. The way
ulation by operatives from the
f Transportation. Yet in point of
rtation stood on the side of the
was an energetic advocate of reg-
problem for Kahn when the two
national transportation policy.[99]
CAB chairman, his most difficult
nd with Congress. In the House,
am was Democratic Representa-
who held a key committee post.
ervices to his Delta Airlines con-
Deltacrat.") Other congressmen
regulation stance of the airline
he most solid support of contin-
n airlines but from labor unions;
ressure on their representatives
. Airline employees, assisted by
ionist policies of the CAB, had
h pay. In 1976, trunkline work-
n baggage handlers to pilots—
0, which put them at the very
United States. Not surprisingly,
by the Air Line Pilots' Associa-
ight engineers, transport work-
vorkers, conducted an active
heir well-publicized power in
a problem unlike any he had
ic Service Commission.[101]

he complex task of moving the
t the same time lobbying Con-
as in New York, he quickly
ngs he did was to lay down the
e it used in orders and corre-
ency had involved itself in so
ary decisions, that it seemed

unable to write convincing explanations and opinions. As Kahn read through the files, he became incredulous, then furious, with what seemed to be intentional obfuscation. Aware that he might be risking the success of his chairmanship, he nevertheless decided to launch a campaign for clarity. The opening gun was a memorandum to the CAB staff:

> One of my peculiarities, which I must beg you to indulge if I am to retain my sanity (possibly at the expense of yours!) is an abhorrence of the artificial and hyper-legal language that is sometimes known as bureaucratese or gobbledygook . . .
>
> May I ask you, please, to try very hard to write Board orders and, even more so, drafts of letters for my signature, in straightforward, quasi-conversational, humane prose—as though you are talking to or communicating with real people. I once asked a young lawyer who wanted us to say "we deem it inappropriate" to try that kind of language out on his children—and if they did not drive him out of the room with their derisive laughter, to disown them.[103]

Bureaucrats who doubted the new chairman's seriousness found their drafts repeatedly bounced back for revision. Letters were rewritten four and five times, as drafts with Kahn's corrections came back to writers with instructions to revise. Board employees said that the experience of writing for Kahn was like preparing a college term paper.[104]

His critics questioned whether the expenditure of all this energy was the best way for an agency head to be spending his time. Kahn, however, remained adamant: "If you can't explain what you are doing to people in simple English, you are probably doing something wrong." Part of the problem came from lawyers' domination of the regulatory process. "One cannot hope over night to wipe out the effects of three years of law school, or decades of non-lawyers trying badly to imitate lawyers (some of whom, by the way, write extremely well)." But, he emphasized, the point was not one of style alone. "I really have certain very profound not only esthetic but philosophical objections to people in Government hiding behind a cloud of pompous verbiage which creates a gulf between them and the people." Reporters enjoyed hearing such messages from the head of a bureaucracy, and the crusade for clarity received wide coverage.[105]

Kahn's second target in the early days of his chairmanship was the CAB's crushing workload. Almost eighteen years after the

Landis report, with its criticism of the agency's backlog, Kahn determined to learn why the workload remained such a persistent problem. He found that the CAB, by insisting on courtroom routine, frittered away time on "cumbersome procedures required to assure due process to all affected parties." Substantively, moreover, the board was adjudicating not only the important decisions that compelled due process, but a mass of trivia as well.[106] Here what Kahn said about the "picayune decisions" that the CAB had to make concerning airlines applied with equal accuracy to the ICC's regulation of trucking and to the Federal Power Commission's oversight of the natural gas industry:

> May a freight forwarder receive commissions from a cargo carrier? To how many travel agents may a tour operator give free passage to inspect an all-inclusive tour? And must those agents then visit and inspect every one of the accommodations in the package? May an air taxi acquire a 50-seat plane? May a supplemental carrier carry horses from Florida to somewhere in the Northeast? Should we let a scheduled carrier pick up stranded charter customers and carry them on seats that would otherwise be empty at charter rates? May an air taxi operate a single round-trip passenger charter flight between two points in substitution for a supplemental? Should we take review of an order by our Bureau of Enforcement dismissing this or that complaint, even though no party has appealed the dismissal? May a carrier introduce a special fare for skiers that refunds the cost of their ticket if there is no snow? May the employees of two financially affiliated airlines wear similar-looking uniforms? . . . Is it any wonder that I ask myself every day: is this action necessary? Is this what my mother raised me to do?[107]

Despite the disarming final question, Kahn was teaching a much larger lesson: "the inexorable tendency for regulation in the competitive market to spread." What he said applied to economic regulation in general:

> Control price, and the result will be artificial stimulus to entry. Control entry as well, and the result will be an artificial stimulus to compete by offering larger commissions to travel agents, advertising, scheduling, free meals, and bigger seats. The response of the complete regulator, then, is to limit advertising, control scheduling and travel agents' commissions, specify the size of the sandwiches and seats and the charge for inflight movies. Each time the dyke [sic] springs a leak, plug it with one of your fingers; just as a dynamic industry will perpetually find ways of opening new holes in the dyke, so an ingenious regulator will never run out of regulatory fingers.[108]

ı of the agency's backlog, Kahn de-
rkload remained such a persistent
AB, by insisting on courtroom rou-
ımbersome procedures required to
ted parties." Substantively, more-
g not only the important decisions
t a mass of trivia as well.[106] Here
yune decisions" that the CAB had
plied with equal accuracy to the
d to the Federal Power Commis-
ıs industry:

ɔmmissions from a cargo carrier? To
operator give free passage to inspect
e agents then visit and inspect every
package? May an air taxi acquire a
carrier carry horses from Florida to
l we let a scheduled carrier pick up
r them on seats that would otherwise
air taxi operate a single round-trip
ɔ points in substitution for a supple-
ın order by our Bureau of Enforce-
ɔlaint, even though no party has
r introduce a special fare for skiers
there is no snow? May the employ-
irlines wear similar-looking uni-
sk myself every day: is this action
ised me to do?[107]

ion, Kahn was teaching a much
lency for regulation in the com-
ıe said applied to economic reg-

rtificial stimulus to entry. Control
ı artificial stimulus to compete by
gents, advertising, scheduling, free
of the complete regulator, then, is
g and travel agents' commissions,
l seats and the charge for inflight
gs a leak, plug it with one of your
perpetually find ways of opening
s regulator will never run out of

The master teacher knew how to hold an audience. It remained for him to make the CAB practice what he preached.

The New CAB

The Civil Aeronautics Board had been in ferment for some two years before Kahn arrived, ever since the Kennedy hearings of 1975. Kahn was the agency's fourth chairman in four years. His immediate predecessor, John Robson, had spoken out in favor of competitive new rate structures and more liberal entry policies. But beyond Robson's rhetorical "one-man reform crusade" (as a trade paper called it), he had not been an effective chairman. He was unpopular with much of the CAB staff; a number of legislators disliked him; and many airline executives despised him. Two highly regarded industry chiefs, Frank Borman of Eastern and Robert Six of Continental, were quoted as saying that Robson was perhaps the worst chairman in CAB history.[109]

This situation benefited Kahn. Not only did he step into the shoes of an unpopular Republican chairman, but as a Democrat he entered office with a board majority on his team. President Carter strengthened Kahn's hand by naming as a Republican member another deregulation-minded economist, Elizabeth E. Bailey, a young veteran of research at Bell Labs. As board colleagues, Bailey and Kahn sometimes disagreed on noneconomic matters, such as smoking by airline passengers. (Bailey took a tough antismoking approach, whereas Kahn favored local option on the grounds that the airlines themselves should respond to customers' preferences.) But on the important questions of regulatory policy, such as freer entry and competitive fares, the two thought alike.[110]

With the backing of his two Democratic colleagues plus Bailey, Kahn took quick steps to reorganize the board. Whereas in New York he had not had this option, in Washington he saw no alternative. He was convinced that the excessive influence of lawyers and the absence of first-rate economists made it impossible for the agency to respond to the economic crisis besetting the industry.[111] In size the CAB was only a little larger than the New York Public Service Commission. Its staff of some 800 was divided between 438 professional, technical, and research personnel, of whom about 120 were lawyers or persons "trained in law," and a remainder of nonprofessionals. The organization included nine staff offices and six line bureaus, each with the power to act:

Enforcement, Economics, Accounts and Statistics, International Affairs, Operating Rights, and Administrative Law Judges. The board's headquarters, located on Connecticut Avenue in Washington, conveniently housed the entire organization.[112] Under Kahn's reorganization plan, the board would retain three of the existing six bureaus (Accounts and Statistics, International Affairs, and Administrative Law Judges) and abolish the others. In their stead, two new bureaus would be formed and reinforced with specialists in economics: Pricing and Domestic Aviation, and Consumer Protection.[113]

More important than any of these was another Kahn brainchild, an Office of Economic Analysis. In asking for additional personnel to staff this unit, Kahn wrote to the chairman of the Civil Service Commission, "I find it simply incredible that the Board has had no genuine Bureau of Economics: the Bureau that until recently had this title was actually concerned only with fares and tariffs and was and is run entirely by lawyers." Kahn elaborated in a letter to a former student who was also a prominent economist. "The reason I set up an Office of Economic Analysis is that I wanted to have objective economists watching carefully what we are doing and how it works out, prepared to identify problems, rather than simply to shut our eyes and rely on faith." Kahn added that "liberals like me" had been too prone in the past to rely on government to solve all problems. Now, he wrote, "I do not want . . . to fall into the opposite error of simply substituting the cliche 'leave it to the market' to take the place of using my own eyes."[114]

Kahn's next step was to bring in a cadre of young bureau chiefs devoted to deregulating the airlines. To head the Office of Economic Analysis, he appointed Darius Gaskins, a well-trained economist with prior experience at the Federal Trade Commission. As general counsel, he hired Philip Bakes, who, as Stephen Breyer's deputy, had played an important role in the Kennedy hearings. And as chief of the Bureau of Pricing and Domestic Aviation, Kahn brought in Michael E. Levine, who twelve years earlier had written for the *Yale Law Journal* the pioneering article on the experience of Pacific Southwest Airlines. Entitled "Is Regulation Necessary?" Levine's article had anticipated with uncanny accuracy the national developments of the succeeding decade.[115]

These appointments of deregulation partisans left no question of Kahn's resolve to make things happen at the CAB. As one of the

unts and Statistics, International
Administrative Law Judges. The
Connecticut Avenue in Washing-
tire organization.[112] Under Kahn's
would retain three of the existing
ics, International Affairs, and Ad-
bolish the others. In their stead,
ed and reinforced with specialists
stic Aviation, and Consumer Pro-

ese was another Kahn brainchild,
n asking for additional personnel
he chairman of the Civil Service
edible that the Board has had no
e Bureau that until recently had
only with fares and tariffs and
." Kahn elaborated in a letter to
prominent economist. "The rea-
ic Analysis is that I wanted to
ng carefully what we are doing
identify problems, rather than
faith." Kahn added that "liber-
the past to rely on government
ote, "I do not want . . . to fall
bstituting the cliche 'leave it to
ng my own eyes.'"[114]
a cadre of young bureau chiefs
s. To head the Office of Eco-
s Gaskins, a well-trained econ-
Federal Trade Commission. As
kes, who, as Stephen Breyer's
ole in the Kennedy hearings.
cing and Domestic Aviation,
who twelve years earlier had
pioneering article on the ex-
nes. Entitled "Is Regulation
icipated with uncanny accu-
e succeeding decade.[115]
n partisans left no question of
n at the CAB. As one of the

new appointees himself put it, "All the bomb throwers are here and in place. You have to judge this agency starting from this point forward; now there are no excuses." A new Civil Aeronautics Board had been born.[116]

At least at one level, the mission for Kahn and his new CAB seemed clear: to create a competitive airline industry in place of the forty-year-old noncompetitive model, so largely defined by public-utility regulation. To do this, the CAB must first melt a frozen rate structure and then encourage competition between airline companies on the basis of price. Routes also had to be made genuinely competitive, so that carriers could choose to serve or not serve particular city-pairs. Finally, new entry into the industry must be invited from all qualified applicants. Each of these tasks for the CAB flowed from a single source: the study of economic reasoning to which Kahn had devoted his professional life.

What remained less clear was the exact method by which he should proceed. Difficult questions of implementation constantly nagged at Kahn, much as they had plagued James Landis and his SEC colleagues during the New Deal. But whereas Landis had devised a successful strategy to impose a regulatory regime on an industry accustomed to laissez faire, Kahn had to do just the opposite. He was about to impose competition as a replacement for overregulation. But where should he begin, and at what pace could he move? Should deregulation come gradually, so as to minimize trauma, or in one "big bang," so as to reach his target quickly? Should Kahn focus his attention on Congress and try to get the existing statutes changed? Or, since the present laws seemed to permit a good deal of tinkering, might it not be better to work within the CAB itself? These questions seemed quite enough to confound any new chairman. Kahn had little time for reflection, however, because the immediate question pressed: should his first move come on rates, routes, or entry? All three must be liberalized—economic theory was clear on that—but in what order?

His pragmatic answer grew out of what happened. The issue of rates drilled the first hole in the aging dike, and reform, if not revolution, had begun. Shortly before Kahn came to Washington in 1977, Texas International Airlines had proposed an off-peak "Peanuts" fare for some of its city-pairs within Texas. The company

also flew routes outside the state, and for that reason it was subject to CAB jurisdiction in all of its operations. The Peanuts fare was designed to compete with the low-cost rates already being offered by Southwest Airlines, the intrastate carrier that had played a starring role in the Kennedy hearings. A few weeks after Texas International's Peanuts application, American Airlines proposed a "Super Saver" coach fare between New York and the west coast, with discounts ranging from 35 to 40 percent, depending on the day of travel. In the last weeks of the Robson chairmanship, the CAB launched an investigation of both these proposals. Then, in a significant reversal of policy, the board agreed to let the proposed rates become effective, pending the outcome of the full investigation.[117]

Now, with Kahn on the scene, the dike burst. As the CAB's *Annual Report* declared, in the kind of understatement typical of such documents, "Thereafter, discounts became more widespread as competing carriers either matched these [Super Saver] proposals or initiated promotional fares of their own. Before long, other reduced fares—'Simple Saver,' 'Super Coach,' 'Super-APEX,' and 'Budget'—proliferated across the Nation and the North Atlantic." Once it became apparent that load factors were indeed climbing rapidly—just as theory had predicted—discount fever swept through the industry. By the summer of 1978, one year after Kahn's arrival, about half of all coach-class revenue-passenger-miles in the United States were being flown at discount rates, with effects especially evident during summer vacation months.[118]

Month	*Percentage of total fares subject to discount*	*Month*	*Percentage of total fares subject to discount*
October 1977	25.9%	April 1978	35.7%
November	24.6	May	36.8
December	31.8	June	41.0
January 1978	31.5	July	56.9
February	32.0	August	56.0
March	35.3	September	48.0

Load factor for the trunklines exhibited its largest jump in history, climbing from 55.5 percent for the year ending in September 1977 to 60.7 percent for the year ending in September 1978. This gain was extremely significant. It meant that on average nearly 10 per-

nd for that reason it was subject
erations. The Peanuts fare was
-cost rates already being offered
tate carrier that had played a
gs. A few weeks after Texas In-
American Airlines proposed a
New York and the west coast,
40 percent, depending on the
the Robson chairmanship, the
oth these proposals. Then, in a
oard agreed to let the proposed
e outcome of the full investiga-

e dike burst. As the CAB's *An-*
of understatement typical of
unts became more widespread
ed these [Super Saver] propos-
their own. Before long, other
er Coach,' 'Super-APEX,' and
ation and the North Atlantic."
factors were indeed climbing
icted—discount fever swept
mer of 1978, one year after
oach-class revenue-passenger-
g flown at discount rates, with
nmer vacation months.[118]

nth	*Percentage of total fares subject to discount*
ril 1978	35.7%
y	36.8
e	41.0
y	56.9
gust	56.0
tember	48.0

ed its largest jump in history,
ar ending in September 1977
n September 1978. This gain
at on average nearly 10 per-

cent more passengers were flying on each plane, but at practically no additional expense to the companies. Even with the discount fares, therefore, the positive impact on airline earnings proved to be dramatic. Operating profits nearly doubled for the trunklines as a whole, climbing from $490 million for the year ending in September 1977 to $905 million for that ending in September 1978.[119] From Kahn's viewpoint, of course, this was the best possible outcome. Even so, what had happened was not quite as simple as the numbers alone seemed to suggest, and some troubling questions persisted. How long would the airlines' prosperity continue? Would load factors continue to climb? And what was the effect on the quality of service in ever more crowded airplanes?[120]

For the fifteen months between Kahn's arrival in Washington and the reporting of these September 1978 figures, several things stand out. One is his determination to make the experiment go forward. "Historically," Kahn said in one speech, "the Board has insisted on second-guessing decisions by individual carriers to offer price reductions." That role, he made clear, had ended. Now carriers could not only reduce prices on their own, but in applying for rate increases they could use projected rather than historical costs as a basis for their application. This represented a major departure from traditional regulatory practice, and for Kahn it repeated one of the changes he had earlier pushed through at the New York Public Service Commission. In a time of inflation, the use of projected rather than historical costs was not merely a more realistic policy; it was also a device calculated to reduce the number of new applications for rate increases, thus cutting the CAB's workload. The board's endorsement of this point, by a vote of 5–0, had major implications for all American industries subject to rate-of-return regulation.[121]

A second concern is evident in Kahn's speeches and in letters to his friends. He became fascinated with the new system of rates proposed by the airlines, in part because they did not evolve in the way that economic theory had predicted. Instead of across-the-board cuts for all passengers, the companies rushed to offer a bewildering variety of special fares on particular days of the week, available to passengers on a first-come-first-served basis. "I realize that all this makes for a messy fare structure," Kahn wrote. "Before coming to the Board, I confess, I regarded all these variations as an abomination." Yet to him they now represented a decided

improvement over the traditional "rigid, non-competitive, uniformly high fare level." What still worried him, at least theoretically speaking, was that the crowded condition of the airplanes might create a new type of discriminatory situation. As he put it to an economist friend, "As load factors rise, I don't quite see how the market process will work to satisfy the demands of people who are willing to pay fares equivalent to the cost of providing service at, say, an average 55% load factor, yet are getting service at 75% or 80%, with the difference this implies in the availability of reservations up to flight time." Here Kahn's institutionalist background helped a bit. "I come from a tradition in which it was considered important to recognize that competition is rarely perfect, and typically works extremely imperfectly."[122]

There were also some signs that basic fares might come down, along with the special rates. Competition might not be working perfectly, but certainly it was working well. If the board had prohibited discriminatory price cutting in the form of special fares (and discrimination indeed violated the rules of classical public-utility regulation), there would have been no breakthrough at all. As Kahn wrote to a law professor who had objected to this discrimination, "Observe the fact that the Super Savers are now available between all major cities of the country; observe the Chickenfeed fare, which is available to anyone who calls in time, regardless of race, creed, previous condition of servitude, or length of hair; and observe, finally, and most satisfactory, the beginnings of competition in the basic fares themselves."[123]

Everything considered, Kahn was having a wonderful time, just as he did in New York. When Barry Goldwater wrote him to complain about the reduced quality of life aboard crowded airplanes, Kahn replied that some discomfort was inevitable. "When a cartel-like regime begins to break up, it breaks up grudgingly, selectively, sloppily ... When you have further doubts about the efficiency of a free market system, please do not hesitate to convey them to me. I also warmly recommend some earlier speeches and writings of one Senator Barry Goldwater." In response to a similar letter from one of his own best friends, who had complained about the experience of flying to Denver seated next to what the friend called a "hippy," Kahn wrote that the experience need not be unpleasant for everyone: "Since I have not received any complaints from the hippy, I assume the distaste was not reciprocated."[124]

ional "rigid, non-competitive, uni-
still worried him, at least theoreti-
crowded condition of the airplanes
criminatory situation. As he put it to
factors rise, I don't quite see how
satisfy the demands of people who
ent to the cost of providing service
ctor, yet are getting service at 75%
implies in the availability of reser-
Kahn's institutionalist background
adition in which it was considered
petition is rarely perfect, and typ-
tly."[122]

hat basic fares might come down,
mpetition might not be working
orking well. If the board had pro-
ting in the form of special fares
ted the rules of classical public-
ave been no breakthrough at all.
who had objected to this discrimi-
Super Savers are now available
untry; observe the Chickenfeed
who calls in time, regardless of
servitude, or length of hair; and
ory, the beginnings of competi-
."[123]

as having a wonderful time, just
y Goldwater wrote him to com-
life aboard crowded airplanes,
was inevitable. "When a car-
it breaks up grudgingly, selec-
ve further doubts about the
lease do not hesitate to convey
end some earlier speeches and
vater." In response to a similar
ds, who had complained about
eated next to what the friend
ne experience need not be un-
not received any complaints
ste was not reciprocated."[124]

Even in the midst of the revolution in rate competition, Kahn and his fellow board members still had to confront the question of how to open up entry into the airline industry by new competitors. Within the CAB, debates broke out over the best way to coordinate these two basic parts of the overall strategy for making the industry competitive. As it turned out, the rapid pace of rate competition forced an acceleration of new entry. As Kahn put it, coordinating rates and entry was "like equalizing the two sides of a mustache: one can do it much more rapidly by cutting down on the longer side than by extending the shorter one!" His answer sounds easier, though, than events made the change seem at the time.[125]

Liberalization of entry, which was only one side of the mustache, included within itself two different possibilities. On the one hand, existing carriers could apply for new routes, and some 600 such applications were pending before the board, many of them dating back several years. As noted before, this backlog reflected the agency's inability during the pre-Kahn era to deal adequately with its assigned responsibilities. Some of the 600 were "totally insincere," said Kahn; others were "actively sought for and prosecuted." But all had to be dealt with, and the prospect of taking them up piecemeal, through the laborious hearings typical of CAB procedure, did not seem pleasant.[126]

On the other hand, a second form of entry could also occur, through a new company's application for admission to the club. Although the CAB had admitted no new trunk carriers in its forty-year history, it had certificated scores of new regional airlines. Once the board began to encourage rate competition, both would-be new entrants and enterprising executives of existing regional airlines started making big plans. Working feverishly, they calculated how they might cut prevailing rates, steal some of the business from trunk carriers, and create new classes of passengers, on the Texas and California models. This frenetic activity within the industry presented the board with some vexing problems, but also with an unusual opportunity to showcase its new policy of competition. "I am reasonably persuaded," said Kahn, "that if we are to make genuine progress toward effective competition, we have to institute some system of automatic, discretionary entry into markets." Only in this way would airline executives be forced to think like other business managers—to plan which services

their companies should provide and which should be left for competitors; and to determine how fare mixtures and route structures might be rationalized, so as to forestall every competitive challenge. Here Kahn's goal led him to inaugurate a policy of granting permissive and nonexclusive entry to airlines applying for new city-pair markets. "If two carriers are applying for a particular route, the traffic on which, it appears, is large enough to justify only one, I suggest that we should carefully be considering *permitting both* but *requiring neither.*" Kahn's idea represented a double reversal of traditional CAB practice. But it would discourage airlines from making preemptive bids for particular markets in order to prevent others from gaining a toehold.[127]

An opportunity to implement these principles came in the form of two dramatic cases involving underused airports in Chicago and Oakland. In the Chicago case, two different applicants asked the CAB for permission to fly routes between Midway Airport and several cities in the Middle West. Midway, situated in the heart of Chicago, was much more convenient for local travelers than was the giant O'Hare Airport located on the city's outskirts. But the major carriers did not like to fly out of Midway because they wanted to build additional traffic through connecting flights with their own scheduled services to other cities, services for which O'Hare provided the center of their hub-and-spoke systems. Since the new applicants had no such networks, they saw an opportunity to create a market niche for themselves in flights between Midway and several medium-sized midwestern cities. Both airlines quoted rates to the CAB 50 percent below those of existing carriers. As soon as these new entrants filed, the carriers already licensed to serve Chicago announced that they would reduce fares in line with the prospective new competition. This step raised the question of whether either of the two applicants could survive in a war with incumbent airlines.[128]

Because of the precedent that any CAB decision would set, the issue of survival held significance beyond the case at hand. As Kahn described it later, "Several civic parties urged us to protect one or both of the innovators by giving them for a year or two the exclusive right to service Midway Airport; some of them originally proposed that we also prohibit incumbent carriers even from matching the low fares at O'Hare. The innovators, they argued, needed and deserved a period of exclusive right to exploit their

and which should be left for com-
fare mixtures and route structures
forestall every competitive chal-
to inaugurate a policy of granting
try to airlines applying for new
ers are applying for a particular
ppears, is large enough to justify
l carefully be considering *permit-*
Kahn's idea represented a double
ice. But it would discourage air-
ls for particular markets in order
toehold.[127]

hese principles came in the form
nderused airports in Chicago and
o different applicants asked the
between Midway Airport and
Midway, situated in the heart of
ent for local travelers than was
on the city's outskirts. But the
out of Midway because they
through connecting flights with
other cities, services for which
ir hub-and-spoke systems. Since
works, they saw an opportunity
elves in flights between Midway
tern cities. Both airlines quoted
w those of existing carriers. As
he carriers already licensed to
ey would reduce fares in line
tion. This step raised the ques-
pplicants could survive in a war

y CAB decision would set, the
beyond the case at hand. As
vic parties urged us to protect
ing them for a year or two the
irport; some of them originally
ncumbent carriers even from
The innovators, they argued,
xclusive right to exploit their

new idea." If the innovators did not get this protection, the incumbents might drive them out of business, then "drift back to O'Hare" in order to maximize connecting reservations. For Kahn, the Midway affair also illustrated his objections to gradual deregulation on a case-by-case basis. "Despite our use of extraordinarily expedited procedures, these applications had been pending for almost two years," during which the incumbents, who already operated over the routes at issue, had the advantage.[129]

Confronted with this situation, said Kahn, "We grasped the opportunity to make our first major grant of universal authority to all applicants, in the belief that this would ensure the fullest and most rapid possible exploitation of the market." In sorting out winners and losers, "the competitive market would do a better job than we." But in order to give the new entrants some breathing room, the board went beyond the case at hand, which involved six city-pairs, and extended multiple permissive authority to include seventeen additional markets. This was designed to prevent the incumbents from blanketing all routes and thereby preempting the potential business of the new companies. "The idea was to open up so many [opportunities that] the incumbents would simply run out of blankets." The new companies embraced this idea, even though during the case they had insisted that they needed temporary exclusive authority. One executive commented, "It's going to be harder for [the incumbents] to grab us in a bigger fishbowl." Added Kahn: "Precisely as we intended."[130]

The Oakland case was conducted along the same lines as Midway, with expedited procedures, multiple permissive entry, and no special protection for innovators. If some companies failed, then so be it. As the CAB's 65-page proposal for Oakland concluded, "We cannot agree to define healthy competition as that state where the fortunes of the competitors fluctuate but no competitor ever goes to the wall." This manifesto expressed a fundamental difference between the policies of the Kahn board and those of the old CAB. Whereas traditionally the CAB had looked on the prospect of an airline bankruptcy with horror, Kahn regarded bankruptcy as a normal result of competitive business pressures: "In healthy competition, producers who are inefficient or make bad decisons may fail, but efficient and well-managed producers can operate profitably." The CAB's achievement in Oakland, together with similar ones at Midway and elsewhere,

made 1978 the year in which, as the agency's *Annual Report* put it, "the CAB route program entered its most active phase in the history of the Board." In the process, Kahn had given up all faith in the merits of gradual deregulation. He had shifted instead to deregulation at a breakneck pace, and the board's policy of proceeding on several fronts simultaneously seemed to be working beyond anyone's wildest expectation. The airlines were enjoying their most prosperous business in many years. And the CAB, said the *Annual Report*, in an explosion of polysyllables but with full accuracy, had "introduced unprecedented substantive and procedural innovations."[131]

Due Process and Competition

These gains did not come easily. The CAB had cut its procedural decision-making time in half, Kahn wrote a close friend, but there had been "blood on the floor." Overall, the chairman reported, the board's efforts to cut red tape "have come under attack from much of the practicing bar." In no case, however, had the CAB abandoned its traditional hearings process, and the repeated suggestions to the contrary always angered Kahn.[132]

Looking back now, it seems clear that the experience of the CAB in 1977 and 1978 provides an exceptionally vivid example of change along one of the main currents in American regulatory history: tradeoffs between economic efficiency and procedural due process. From the very beginning, as the Massachusetts experience of Charles Francis Adams demonstrated, all attempts to regulate economic behavior toward the goal of greater efficiency encountered obstacles based on traditional values of American life, liberty, and property—protection of which was guaranteed under the Bill of Rights and the Fourteenth Amendment. In the 1870s, Adams had sidestepped this difficult question by refusing to assert regulatory powers that would have interfered with the constitutional rights of persons and thus have drawn his commission into court. The Massachusetts Board of Railroad Commissioners emphasized publicity and shunned coercion. In that way it remained out of court, its powers ironically maximized through its own policy of restraint. But the tradeoffs between economic efficiency and due process continued long after Adams' time. In some ways, the entire history of responsible regulation is encapsulated in the ongoing effort to strike a proper balance between the two.

s the agency's *Annual Report* put
tered its most active phase in the
ocess, Kahn had given up all faith
lation. He had shifted instead to
ce, and the board's policy of pro-
neously seemed to be working be-
ion. The airlines were enjoying
n many years. And the CAB, said
ion of polysyllables but with full
ecedented substantive and proce-

The CAB had cut its procedural
n wrote a close friend, but there
erall, the chairman reported, the
ve come under attack from much
e, however, had the CAB aban-
ocess, and the repeated sugges-
ed Kahn.[132]

lear that the experience of the
exceptionally vivid example of
rrents in American regulatory
ic efficiency and procedural due
as the Massachusetts experience
trated, all attempts to regulate
l of greater efficiency encoun-
al values of American life, lib-
which was guaranteed under
nth Amendment. In the 1870s,
t question by refusing to assert
e interfered with the constitu-
ve drawn his commission into
Railroad Commissioners em-
rcion. In that way it remained
aximized through its own pol-
ween economic efficiency and
lams' time. In some ways, the
ion is encapsulated in the on-
ce between the two.

For Louis D. Brandeis, in contrast to Adams, the keys to regulatory definition making and standard setting did not emerge as a result of informal commission rulings, but rather from court decisions and procedures. For him, the essence of regulatory routine could be found in established legal rules of evidence, testimony, and cross-examination. The adversarial theory—two strong advocates arguing for their versions of truth before impartial judges and juries—lay at the heart of the Anglo-American conception of liberty, an ideal that Brandeis would not willingly undercut. Brandeis admonished the first members of the Federal Trade Commission not to give advance advice to business managers, but to rely instead on established judicial procedures as the best method for setting the limits of competition. This devotion to cherished principles symbolized not only Brandeis' stand but also the collective view of the American bar toward regulation itself.

In the early years of administrative regulation, American courts took a stern stance toward the upstart commissions. Partly from jealous guardianship of power, partly from unfamiliarity with administrative needs, judges often limited regulatory discretion so severely that the commissions had to return to the legislatures again and again for grants of clarifying powers. By the time of the New Deal, however, the discretion of administrative agencies had begun to spread beyond previous boundaries, as the powers of commissions and boards multiplied. James Landis celebrated this phenomenon in *The Administrative Process*, which caught the tenor of the times. Unlike Brandeis, Landis embraced advance advice and made it the normal routine of the SEC. By the end of World War II, then, the advocates of wide regulatory discretion and minimal judicial review were in command. For a brief moment the "conservative" opponents of this discretion—usually found in bench and bar—seemed to have lost the fight.

At just this point in American legal history, the old establishment staged a counterattack, the traditional viewpoint having been strengthened by two recent developments. One was a wave of national conservatism, a common aftermath of wars. The second brought to the foreground a reinforcing battalion of "liberal" lawyers who were growing uncomfortable with government threats to civil liberties. Conscious of the rise of totalitarianism in the 1930s and of its persistence in eastern Europe during the 1940s, civil libertarians now joined with the old opponents of

commission power to oppose broad governmental discretion. The result was an effective political coalition that pushed the Administrative Procedure Act of 1946 through Congress.[133]

That act proved to be a milestone in the history of American administrative government. Over the next generation, not only regulatory decision making but nearly all administrative action came under the new rules of notice, reasonableness, and opportunity for all affected parties to give evidence and testimony. Later, in the mid-1960s, still another judicial revolution moved regulatory practice even farther down the track of procedural safeguards. This was the revolution in "standing," which brought new participants and issues into regulatory proceedings. Now matters of environmental impact, affirmative action, and rights of other affected parties added new layers of procedural requirements to the already sizable burden of the regulatory process.[134] In all of this, of course, the influence of lawyers on regulation—always extremely powerful—reached its historical zenith. Lawyers seemed indispensable, and the corporation or agency that tried to do its business without expert legal advice was simply being foolish.

Alfred Kahn viewed this situation with displeasure when he undertook to reform the regulatory scene. In New York, his chief response took the practical shape of the generic rate hearing. At the CAB, Kahn launched a frontal assault on what he regarded as the flagrant and broad-gauged abuse of due process, by public and private sectors alike. "It seems to this layman," he said, "that 'due process' is defined in an asymmetrical fashion: it seems inherently to give protection to the parties who benefit from delay, and to be injurious to the parties—typically the public at large—who are *adversely* affected by delay . . . In short, the requirements of *legal* due process, interposing the heavy hand of government between an idea and its application in the market, are directly antithetical to competition."[135]

Kahn favored decision making in the open; thus he championed the sunshine laws enacted by Congress in the late 1970s. But he knew that the basic problem transcended the issue of openness. Rather it had to do, as he wrote the chairman of the House Subcommittee on Aviation, with "being placed in the position of having to tell a businessman to wait two years until we bureaucrats have shuffled enough papers to permit us to reach a decision

ad governmental discretion. The
coalition that pushed the Admin-
hrough Congress.[133]

tone in the history of American
r the next generation, not only
nearly all administrative action
ce, reasonableness, and opportu-
e evidence and testimony. Later,
dicial revolution moved regula-
the track of procedural safe-
"standing," which brought new
tory proceedings. Now matters
tive action, and rights of other
rs of procedural requirements
he regulatory process.[134] In all
lawyers on regulation—always
s historical zenith. Lawyers
rporation or agency that tried
legal advice was simply being

n with displeasure when he un-
ene. In New York, his chief re-
he generic rate hearing. At the
ılt on what he regarded as the
f due process, by public and
is layman," he said, "that 'due
al fashion: it seems inherently
benefit from delay, and to be
the public at large—who are
ort, the requirements of *legal*
hand of government between
rket, are directly antithetical

he open; thus he championed
ess in the late 1970s. But he
ended the issue of openness.
chairman of the House Sub-
placed in the position of hav-
o years until we bureaucrats
mit us to reach a decision

which probably shouldn't be any of our concern in the first place."[136] Kahn's best opportunity for a full expression of his thoughts on due process came when the American Bar Association invited him to speak to its members in New York. As a title for the session, planners had selected "Regulatory Reform at the Civil Aeronautics Board—With or Without Legislation—With or Without Due Process." Their provocative gesture gave Kahn the opening he had been waiting for. His speech, despite its manifest passion, represented his mature thoughts on the issue, framed after a career of careful writing about regulation and after more than four years' experience as chairman of two important regulatory agencies.

Ever the showman, Kahn began his address by quoting, in Scots dialect, four stanzas from Robert Burns's satiric poem, "To a Mouse." Such a "wee, sleekit, cowrin, tim'rous beastie" had devised "the inhospitable, smart-alecky title." Having captured the attention of the assembled lawyers, Kahn moved on to an analysis of their motives. Referring to route and entry procedures, "where we encounter by far the most complaints from some members of the aviation bar," he said that he and his colleagues had tried to substitute innovative methods for the "Perry Masonisms that characterize the typical CAB proceeding." The bar, deprived of long and "lucrative" procedures, had protested. And the lawyers' real objections, Kahn suggested, were not procedural but substantive; they did not wish to see so much new entry into the industry. "Indeed, I cannot help feeling that some among you, determined to resist our present policies with all the weapons at your command, have seized upon procedural arguments because you really can't refute our new policies substantively—just as in the past some of you used laws designed to protect the environment or to encourage energy conservation as devices to shield your clients from competition."[137]

Kahn continued his address, blistering the captive audience with the heat of his powerful attack:

> Are we, in removing the restraints that for 40 years have kept people out of the airline industry, in proposing to open up markets to all fit applicants, and relieving incumbents of burdensome restrictions on their operating rights, are we running afoul of the Fifth Amendment's injunction against depriving persons of "life, liberty or property without due process of law"? . . . Are we in so doing depriving the incumbent carriers

of property, with or without due process? I find no justification in either the Constitution or the Federal Aviation Act for the assumption that the certificates of public convenience and necessity that we issue under the Act for providing air service over particular routes need be exclusive, or that the Board is under the slightest constitutional obligation to safeguard their value, as protected property under the Fifth Amendment. I have been told by people who have been at the CAB a long time . . . that in the past the Board would often choose among competing applicants for the right to operate a particular route in secret sessions, held in a closed room from which all staff were rigidly excluded; that somehow out of that process emerged a name attached to the route in question; that the Chairman—or perhaps his assistant—would then pick up the telephone and call the General Counsel and tell him who the lucky winner was, and nothing more; that then a lawyer on the General Counsel's staff, amply supplied with blank legal tablets and a generous selection of cliches—some, like "beyond-area benefits," "route strengthening" or "subsidy need reduction," tried and true, others the desperate product of a feverish imagination—would construct a work of fiction that would then be published as the Board's opinion. Need I add that any resemblance between it and the Board's actual reasons for its decision would be purely coincidental? And then the courts solemnly reviewed these opinions, accepting the fiction that they truly explained the Board's decision, to determine whether the proffered reasons were supported by substantial evidence of record.

Well, we have come to grips with the institutional decision. The Sunshine law has helped force us to do so, but I would like to take some small credit too . . . I am particularly proud, for example, of our decision in the Midway case—not mainly because it is well written, although I think it is, but because it lays out, step by painful step, the actual agonizing process we went through in deciding to certificate all applicants, and to protect nobody from anybody . . . I believe that one substantive regulatory principle on which we can all agree is the principle of minimizing coercion: that when the government presumes to interfere with peoples' freedom of action, it should bear a heavy burden of proof that the restriction is genuinely necessary . . . The dispensation of favors to a selected few is a political act, not a judicial one . . . None of this exempts a regulatory commission from the requirement to be scrupulously fair. But due process in the fashioning of economic *policy* is not the same as due process in a criminal trial, and the attempt to hedge it about with the same kinds of procedural restrictions perpetrates all sorts of unfairnesses, in the name of due process.[138]

The discomfort among the audience can only be imagined, though many listeners must have admired Kahn's candor.

The CAB's Constituencies

Like Adams and Landis before him, Kahn considered it vitally important that his policies be made clear to his various constituencies. He spent an enormous amount of time making speeches, writing letters, and cultivating the press. In fact, one of his greatest strengths—perhaps his greatest—was as communicator and educator. Seldom did he mince words or try to sweet-talk hostile interest groups. Convinced of the correctness of his cause, Kahn attempted to explain his position, point out the inconsistencies of opposing arguments, and persuade through reasoned argument. As a teacher, he never retired.

More than any other group, the airline labor unions learned Kahn's lessons the hard way. Quite accurately sensing that wages would decline if the industry suddenly became competitive, the unions did not want to surrender the cartel-like protection afforded them by the old CAB. Kahn, despite the prolabor sympathies characteristic of his background as a liberal Democrat, met their objections head on. He conceded that regulation did raise the compensation received by labor, and not just airline labor. The reason for this, Kahn believed, could be found in the cost-plus character of traditional regulation, which induced companies to pass along input costs to their customers and which discouraged them from effective searches for cost reduction. Regulators, for their part, permitted such a policy because they wanted to keep peace within the industry and because they themselves were accustomed to cost-plus ways of thinking.[139]

After voicing his lack of sympathy with this view, Kahn was deluged with hostile mail from labor. He proved more than ready to answer his critics. In his typically open and didactic manner, he responded to one pilot who had protested against Kahn's statements about excessive pay:

> You argue in rebuttal that being a pilot requires a high degree of skill, long training, and great responsibility and entails arduous working conditions. That is absolutely correct. From these facts, you conclude that the present high wages and short hours are justified. I disagree. The sort of value-of-service pricing you suggest is not the way that wages are set in the competitive sectors of our economy. Wages, like the prices of other goods and services, are determined on the basis of supply and demand. The long training and skill required to be a successful pilot will

always limit the supply, and thus justify wages markedly higher than average in our economy—but not necessarily as much higher as today.[140]

Again, Kahn told a protesting executive of the Air Line Pilots Association that he was "badly misinformed" about what the CAB was trying to do. After denying that he himself was antilabor, Kahn insisted: "If I'm anti-anything, I'm anti-excessive government interference . . . And I am *particularly* against government being used to protect powerful business interests by giving them special grants of monopoly privilege." His dander up, Kahn added that he resented "the cheap argument you repeat about my having tenure at Cornell . . . I don't *need* tenure at Cornell." But the central lesson was, "Lower prices induced by more competition mean *more* jobs, not fewer: don't you forget that when you say it is you who speak for labor, and not I."[141]

Nor did Kahn limit his candor to distant correspondents or lone dissenters. Informed that Pan American World Airways had flouted the CAB's advice on discount tickets, he said, "I'm inclined to tell Pan Am to go to hell." When Jimmy Carter reversed the CAB's recommendation about an international route award, Kahn openly criticized the president. "There have been three major international decisions, and he has overturned me on all three." Expressing himself as thoroughly disappointed, Kahn told reporters that Carter "is the President and has the right to make a judgment, and I have a right to disagree with it, which I do." The chairman said later, on a television program, that Carter's decision "casts a shadow on the integrity of the [CAB] process."[142] Even so, throughout his tenure at the CAB, Kahn's relationship with the president remained warm and close. Kahn usually communicated with the White House on only a few kinds of issues, such as international aviation, where the president possessed much broader regulatory authority than over domestic policies. A related issue concerned the attempts of Secretary of Transportation Brock Adams to move into the domain of both the CAB and the State Department (which led American negotiations on international airline matters). Kahn vigorously repulsed Adams' foray.[143]

A third area of communication between Kahn and the White House concerned the appointment of new board members. In 1978, the terms of his two Democratic colleagues ended, and Kahn

stify wages markedly higher than av-
cessarily as much higher as today.[140]

ecutive of the Air Line Pilots As-
sinformed" about what the CAB
that he himself was antilabor,
hing, I'm anti-excessive govern-
particularly against government
usiness interests by giving them
ge." His dander up, Kahn added
nent you repeat about my having
tenure at Cornell." But the cen-
uced by more competition mean
rget that when you say it is you
1

o distant correspondents or lone
American World Airways had
nt tickets, he said, "I'm inclined
hen Jimmy Carter reversed the
nternational route award, Kahn
There have been three major in-
verturned me on all three." Ex-
appointed, Kahn told reporters
s the right to make a judgment,
it, which I do." The chairman
, that Carter's decision "casts
the [CAB] process."[142] Even
CAB, Kahn's relationship with
close. Kahn usually communi-
nly a few kinds of issues, such
he president possessed much
over domestic policies. A re-
ts of Secretary of Transpor-
the domain of both the CAB
h led American negotiations
Kahn vigorously repulsed

etween Kahn and the White
of new board members. In
ic colleagues ended, and Kahn

was concerned about suitable replacements. Not only his party majority but also his policies were at stake. "Above everything else," he wrote to Hamilton Jordan, "what I need on the Board is professionalism. Every step on the path to regulatory reform is politically charged. The *only* way we can persevere is if each Board appointee has a firm, independent professional background and qualifications, and is in a postion to make strong, intellectually-grounded professional judgments." So Kahn took an active hand, proposing specific names for President Carter's consideration. "I know you already recognize that my job in months and years ahead will be impossible if you do not actively involve me in this process."[144] A little later in 1978, Kahn sketched for the White House the general situation:

> Remember we have a nervous Congress; our decisions are subject to Court review; and the airlines and the financial community are constantly demanding assurances that we are not a bunch of wild-eyed ideologues, and that we know what we are doing . . . [We have adopted] genuinely novel procedural and substantive policies. That process can work well only if each Member participates actively . . . What kind of people do I need, specifically? I need people with special skills in making highly technical public economic policy—with training and/or experience in the law, economics, regulation. I don't need politicians. Maybe I can stand one—two, absolutely not.[145]

In the end, Carter's appointments met with Kahn's approval. The president was on his side, in part because of the chairman's careful attention to White House relations.

Lobbying the Congress

Senators and representatives provided yet another set of constituents about whom Kahn must worry. In fact, because legislation to deregulate the industry was pending on Capitol Hill throughout his chairmanship, he spent more time dealing with Congress than with any other group except for the board itself.[146] The road to legislative change wound tortuously around the special interests opposing deregulation. The coalition of industry and labor spokesmen alone seemed an almost unbeatable opposition, even without their allies in the Department of Transportation and among remaining dissidents in the CAB. A related problem seemed the sheer weight of history. Over time, Congress had enacted an enor-

mous amount of regulatory legislation, beginning with the Act to Regulate Commerce of 1887. And in the ninety years since that beginning, the trend had all been one way: toward more of the same. Congress had never taken anything like the giant step backward that Kahn and his allies were now urging. Even the word "deregulation" had entered the national vocabulary only recently, and so far there was little besides bombast to stand behind it. Numerous deregulation bills had been introduced; but, as a trade paper reported late in 1977, "after three years of elephantine rhetoric," the lack of action made deregulation look like a mouse. Powerful presidential support would have helped, but Jimmy Carter was preoccupied with his energy legislation.[147]

Already Kahn had achieved spectacular success within the CAB, with his policies of rate competition, open entry, and new route authority. The industry responded with lower fares, higher earnings, and service initiatives. Yet this very success seemed to be giving ammunition to Kahn's opponents, who argued that the CAB's turnabout made new legislation unnecessary.[148] For Kahn himself, no argument could have sounded more ironic. The real fight, he knew, had just begun. For one thing, the board's innovative procedures must still meet the test of judicial review. The challengers who had appealed the CAB's Midway and Oakland decisions might well win because they had "40 years of CAB precedent to back them up." Not only that, but the appointment of different board members could quickly reverse the progress already made. Therefore Congress must act. A well-designed deregulation bill, Kahn concluded, will "secure these advances against a change of agency policy, and it will surely avoid a lot of difficult litigation challenging the legality of the reforms we have already introduced."[149]

Many members of Congress balked at deregulation because of their uncertainty about future airline service to small cities once deregulation was on the statute books. These legislators were acutely sensitive to their constituents' wishes for big-time airline service, a concern augmented by their own dependence on frequent jet flights for quick trips to and from home. Hundreds of legislators, particularly those from rural districts, were loath to tamper with the existing system. As noted earlier, the CAB's forty-year pattern of dealing with local service had been to maintain a balancing act among routes with heavy and light traffic.

ation, beginning with the Act to
d in the ninety years since that
n one way: toward more of the
nything like the giant step back-
ere now urging. Even the word
itional vocabulary only recently,
bombast to stand behind it. Nu-
een introduced; but, as a trade
three years of elephantine rhet-
leregulation look like a mouse.
ould have helped, but Jimmy
energy legislation.[147]

spectacular success within the
mpetition, open entry, and new
onded with lower fares, higher
et this very success seemed to be
ponents, who argued that the
ation unnecessary.[148] For Kahn
sounded more ironic. The real
r one thing, the board's innova-
he test of judicial review. The
e CAB's Midway and Oakland
hey had "40 years of CAB prec-
hat, but the appointment of dif-
ly reverse the progress already
t. A well-designed deregulation
these advances against a change
avoid a lot of difficult litigation
eforms we have already intro-

ked at deregulation because of
ine service to small cities once
books. These legislators were
nts' wishes for big-time airline
their own dependence on fre-
d from home. Hundreds of leg-
rural districts, were loath to
As noted earlier, the CAB's
ocal service had been to main-
s with heavy and light traffic.

Confronted with this issue, and with the fears on Capitol Hill, Kahn took the same line of attack that had served him so well in the past. He acknowledged the reality of the problem, laid out a theory for its solution, and then sought by reasoned analysis to change the minds of skeptics. In addition to his many appearances before subcommittees of both houses, he talked with scores of legislators and their staff assistants, called them repeatedly on the telephone, and wrote one letter after another in response to specific queries. Even under the existing system, Kahn told one senator, the airlines were flying more than they had to under the terms of their certificates. "There is absolutely no reason to doubt that the same would be true under the automatic entry provisions." Fares would drop, traffic would grow, and the frequency of service might actually increase. "Is this prediction merely hypothetical, merely based on deductive reasoning? I submit it is not. It is based on the lessons of economic history; this is how a competitive market operates, and you can see proof of it everywhere about you."[150]

The chairman's own congressman, representing the Cornell district in New York, wrote him to protest the CAB's decision permitting Allegheny Airlines (later U.S. Air) to reduce its service to Ithaca. Kahn replied that eventually Ithaca and similarly sized cities would receive better service from small commuter airlines than the CAB could ever force Allegheny to offer. True, the market adjustment would take some time, and gaps would occur during which there might be no service at all. But Kahn saw no alternative, given the overall goal of making the industry competitive: "Telling carriers they must compete and cannot look to the government for protection is totally inconsistent with telling them also that they must provide unremunerative services at their own expense."[151] The chairman lobbied endlessly on this point, and he tried to enlist the help of local airlines themselves. "In almost every conversation I have had with Congressmen," he said in a speech to the Commuter Airlines Association, "I have encountered a tendency to deprecate the value of service by small aircraft. They and their constituents seem to think it is somehow demeaning to lose service by large jets once or twice a day, and receive instead service by the obviously more suitable and efficient smaller commuter aircraft, even when that means service four and five times a day—this in the face of all the objective evi-

dence that travelers respond positively to such substitutions. I think you have a major public relations job here."[152]

Kahn himself, of course, excelled at public relations. He turned every contact with members of Congress into an opportunity for further lobbying. In one letter he sent to every member of the Senate, he noted that the CAB unanimously supported the deregulation bill, enclosed two pages summarizing key issues, attached fifteen pages of relevant testimony, and closed with an offer of further help: "I would like to extend to you an offer of my services and those of my staff to answer whatever questions you may have about the need for the legislation or any of its provisions. In particular, I know you are probably concerned about the provision of air service to the smaller communities within your state . . . If you would like information about that or any other aspect of the bill, please don't hesitate to call on me."[153]

Kahn was an effective lobbyist in part because he received such a good press. Congressmen liked to talk with him; he was interesting and considerate of their problems, and the media were making him into a national celebrity. He granted interviews freely and spent a good deal of time talking to reporters. Kahn understood their needs and prejudices, and played directly to both. He made a point of self-deprecation, but seldom failed to give each interviewer something new, some small exclusive. As a result, he received extremely wide coverage, nearly all of it favorable. Detailed feature articles about him appeared in the *New York Times*, the *Wall Street Journal*, the *Washington Post*, *Fortune*, *Forbes*, *Barron's*, *National Journal*, and many others. He was written up regularly in such trade publications as *Airline Reports*, *Aviation Daily*, and *Travel Weekly*. He appealed to even broader audiences, in writeups for *People* magazine and on television programs such as the "MacNeil-Lehrer Report" and William F. Buckley's "Firing Line." He was always a hit. After being told that the MacNeil-Lehrer producers had received a record number of requests for transcripts of his appearance, Kahn wrote a friend, "Maybe I should throw up my present job and do what I really would love to do—play musical comedy on whatever stage will accept me."[154] But in fact he stood on the right stage already. When a student who had taken his courses at Cornell thirty years before wrote him about his new celebrity, Kahn replied that he "was awfully pleased to hear your recollections of me as a teacher.

sitively to such substitutions. I
ations job here."[152]
d at public relations. He turned
Congress into an opportunity for
e sent to every member of the
animously supported the dereg-
ummarizing key issues, attached
, and closed with an offer of fur-
l to you an offer of my services
natever questions you may have
r any of its provisions. In partic-
cerned about the provision of air
es within your state . . . If you
or any other aspect of the bill,
."[153]
n part because he received such
talk with him; he was interest-
ms, and the media were making
granted interviews freely and
to reporters. Kahn understood
yed directly to both. He made a
dom failed to give each inter-
ll exclusive. As a result, he re-
, nearly all of it favorable.
m appeared in the *New York*
he *Washington Post*, *Fortune*,
and many others. He was writ-
ublications as *Airline Reports*,
. He appealed to even broader
nagazine and on television pro-
rer Report" and William F.
ways a hit. After being told that
l received a record number of
earance, Kahn wrote a friend,
esent job and do what I really
omedy on whatever stage will
od on the right stage already.
courses at Cornell thirty years
elebrity, Kahn replied that he
ecollections of me as a teacher.

In a way, I feel that I am still in that profession—but giving my lectures from a slightly different platform."[155]

The Act of 1978

In October 1978, Congress finally passed the Airline Deregulation Act. It formalized all the principles Kahn had been enforcing administratively. Now, at last, the future of these principles seemed secure against adverse judicial challenge. Kahn's only regret, he said later, was that the passage of the act denied him the opportunity of defending the CAB's policies at the Supreme Court, where he would have "presented the ultimate Brandeis brief." The new legislation made the setting of fares competitive, gave the carriers almost unlimited route authority, and opened up the airline business to free entry by new companies. In a gesture calculated to emphasize that the act meant business, Congress added a provision calling for the disappearance of the Civil Aeronautics Board after a suitable transition period, defined by the act to expire in 1985.[156]

For all these changes, it is unrealistic to credit Kahn alone. Many hands helped, and some had been toiling long before Kahn came to Washington. The Kennedy hearings of 1975, planned and executed by Stephen Breyer, certainly began the groundswell of support for deregulation. The subsequent joint efforts of Naderites, conservative Republicans, and liberal Democrats amounted to a rare spectacle of cooperation by diverse allies. And the persistent lesson of experience—the examples of success without tight regulation in California and Texas—provided the necessary demonstration without which deregulation perhaps never would have arrived.

Kahn's primary contributions lay in two areas. First, by setting an example through powerful administrative action, he made it impossible for the opposition to paint a horrible picture of a deregulated future. During his sixteen months at the CAB, the airlines enjoyed the most prosperous economic times in their recent history. Their load factors climbed rapidly, their profits doubled, and they could no longer keep claiming that deregulation meant disaster. One by one, the airlines began to soften their stance. Led by United, the largest of the carriers, a few major companies broke ranks and went over to Kahn's side. Seeing deregulation as likely

to come despite their campaigns against it, these defecting airlines decided on a policy of trying to shape the specific package enacted by Congress.

Second, having neutralized the opponents, Kahn moved to consolidate the supporters and mobilize them in favor of new legislation. He did this by a masterly orchestration of diverse forces. Like a gifted conductor managing an intricate composition, he brought every player, every instrument, into the piece at the appropriate time. He made sure that the movements of the symphony did not get out of proper sequence, as they were forever threatening to do. Kahn managed the media, the Congress, the White House, and most important of all the CAB itself, giving each its individual cues, its required up- or downbeat. Exploiting the momentum that had been building since the early 1970s, he said all the right words, called in all the right clichés.[157]

When the act finally passed, the magazine *Aviation Week and Space Technology* editorialized that Kahn had shaken the CAB out of its torpor. In an outburst of mixed metaphors, the magazine praised his tenure at the head of a board that had been "manned largely by political lame ducks and operated by a swollen, legalistic bureaucracy, was still flying through a DC-3 environment that regulated the airlines down to a minimal survival diet and denied the ingredients vital for healthy growth. Surgery always is drastic, and Mr. Kahn's actions induced a certain post-operative shock. But he effectively removed the palsied hand of the CAB from the airlines' growth throttle."[158]

On the day the deregulation act was signed, Kahn wrote several members of Congress to thank them for their support. To one senator he said, "I can't begin to tell you how happy I am that [the Act] has been passed or how confident I am that with it we can create a truly competitive airline industry that is genuinely responsive to the needs of consumers and better, too, for airlines and their employees." He added a handwritten postscript. "This is the last letter I will sign as Chairman of the CAB. As you may have read, the President has persuaded me to try to help him with the problem of inflation. I will try to apply in that job the same philosophy I've tried to apply to air transport." President Carter informed the CAB staff that Kahn was leaving with extreme reluctance and that he himself was taking the chairman away only because even more important tasks beckoned. "You at the Board,"

ST'S HOUR

gainst it, these defecting airlines
pe the specific package enacted

ppponents, Kahn moved to con-
ze them in favor of new legisla-
estration of diverse forces. Like
tricate composition, he brought
to the piece at the appropriate
ments of the symphony did not
were forever threatening to do.
ngress, the White House, and
self, giving each its individual
Exploiting the momentum that
y 1970s, he said all the right
s.[157]

magazine *Aviation Week and*
Kahn had shaken the CAB out
ixed metaphors, the magazine
board that had been "manned
operated by a swollen, legalis-
ough a DC-3 environment that
nimal survival diet and denied
wth. Surgery always is drastic,
certain post-operative shock.
ied hand of the CAB from the

vas signed, Kahn wrote several
for their support. To one sen-
you how happy I am that [the
lent I am that with it we can
industry that is genuinely re-
nd better, too, for airlines and
written postscript. "This is the
of the CAB. As you may have
ne to try to help him with the
pply in that job the same phi-
ansport." President Carter in-
s leaving with extreme reluc-
ing the chairman away only
peckoned. "You at the Board,"

Carter wrote, "have presented my Administration with one of its great success stories."[159]

A number of his friends expressed surprise when Kahn accepted the anti-inflation job as head of the Council on Wage and Price Stability. Indeed he managed to achieve little more success in it than had his predecessor, Robert Strauss. As administrator of a voluntary control program, the chief inflation fighter had few weapons. Within the administration, Kahn had to vie for influence with other heavyweights: Treasury Secretary Michael Blumenthal, White House domestic policy chief Stuart Eizenstat, Council of Economic Advisers Chairman Charles Schultze, and Office of Management and Budget Director James McIntyre, Jr. Several of these rivals had been administration insiders since Carter's inauguration. Even within his own council, Kahn had to compete for influence with Director Barry Bosworth. For months, a freeze on new federal hiring frustrated Kahn's attempts to recruit even a small staff to assist him.[160]

In office, Kahn continued to receive a good press, but the politics of his position made it impossible to avoid costly struggles. Almost immediately, he became embroiled in disagreements with AFL-CIO chief George Meany, who argued that the administration's economic program favored business at the expense of labor. On another occasion, when Kahn suggested that consumers might boycott stores whose price increases exceeded the guidelines, he came under severe attack in the business press. Journalists continued to praise his forthrightness, but noted that this very quality caused "teeth-gnashing" within the administration.[161]

In the 1970s, Kahn had been very good at pointing out areas in which industry-specific protectionist arrangements were promoting inflation and costing consumers money. As chief inflation fighter, he continued his campaign against inappropriate regulation in particular industries; the Carter Administration's program for trucking deregulation, for example, was effectively managed from Kahn's office. But for the country at large, the control of inflation represented a special macroeconomic problem surrounded by a good deal of uncertainty. No rigorous theory of inflation existed, at least none comparable to the long-tested theories of price and entry that guided Kahn's CAB policies. The idea of wage-and-price guidelines seemed to contradict the very reliance on market mechanisms that had formed the basis for his great campaigns at

both the CAB and the New York Public Service Commission. Success for Kahn in the role of chief inflation fighter was just not in the cards. He acquitted himself well, but the odds defeated him as they had defeated Strauss and others.[162]

Kahn's Achievement

Kahn had gone from Cornell to the New York Public Service Commission in 1974 and had remained at that post for a little less than three years. He arrived in Washington in 1977 and held the CAB chairmanship for about a year and a half. The four and a half years of his agency chairmanships is still too recent to allow for mature historical perspective. The utility and the airline stories are far from over. Each industry has been powerfully affected by subsequent events, especially by the absence of general economic prosperity in the years after Kahn's departure.

He had left both the New York PSC and the CAB, for example, prior to the Iranian revolution and the "second oil crisis" of 1979. These events shook the world economy and contributed to the start of another prolonged recession, which in turn struck hard at both electric utilities and airlines. For utilities, the recession meant diminished demand for electricity. This was not entirely bad news, given the industry's upward-shifting marginal-cost curve. But the decline in demand growth was so unusual that utilities actually began canceling construction projects long underway, thereby accepting writeoffs of hundreds of millions of dollars. Around America, abandoned construction sites stood as silent reminders of unrealistic assumptions about growth as the natural order of things—ghost towns of a lost prosperity.

In such a situation, marginal-cost principles assumed even greater significance than they had during Kahn's tenure in New York. The nationwide movement toward the right kinds of price signals had begun even before the second oil shock, and in that movement Kahn's pioneering played an important role. His generic rate hearings were copied in several other states, as intervenors and commissions sought to educate companies and consumers about the potential benefits from marginal-cost rate structures.

Of particular importance was the Public Utility Regulatory Policies Act of 1978, which Congress passed partly to induce com-

ıblic Service Commission. Suc-
nflation fighter was just not in
l, but the odds defeated him as
rs.[162]

the New York Public Service
ned at that post for a little less
shington in 1977 and held the
and a half. The four and a half
is still too recent to allow for
utility and the airline stories
ıs been powerfully affected by
e absence of general economic
departure.
SC and the CAB, for example,
the "second oil crisis" of 1979.
nomy and contributed to the
, which in turn struck hard at
. For utilities, the recession
tricity. This was not entirely
upward-shifting marginal-cost
owth was so unusual that util-
struction projects long under-
hundreds of millions of dollars.
uction sites stood as silent re-
about growth as the natural
ost prosperity.
ost principles assumed even
during Kahn's tenure in New
ward the right kinds of price
second oil shock, and in that
ed an important role. His ge-
several other states, as inter-
to educate companies and
ıefits from marginal-cost rate

Public Utility Regulatory Pol-
passed partly to induce com-

panies and state commissions to use marginalist principles in rate-making. The staff of the New York Public Service Commission directly influenced the content of this legislation, and the commission's own position, of course, derived from Kahn's earlier tenure in Albany. The act imposed a statutory deadline (November 1981) by which all fifty states must follow certain standards in setting rates. One of these standards prohibited a utility from using declining-block rates unless it could demonstrate that its costs declined as its customers' consumption increased, a provision identical to one Kahn had imposed on New York utilities during his PSC chairmanship. Even though national progress toward implementing marginal principles lagged and the 1981 deadline was not universally met, the revolution in utility pricing went forward. All over the country in the early 1980s, rate hearings began to recapitulate the New York experience. Few innovative principles emerged that had not been anticipated by Kahn in the generic rate hearings of 1975–76. Again, the real question was whether state regulatory bodies could adapt quickly enough: whether they could look beyond their own inclination to discipline utilities by reflexive denials of requests for rate increases. As the industry entered one of the most critical stages in its history, some observers began to predict serious power shortages by the late 1980s.[163]

In the airline industry, the post-Kahn era did not appear to be so ominous, but still it was full of uncertainty. The prosperity of the 1977–78 period hardly lasted beyond Kahn's departure, as the oil crisis hit the airlines, doubling fuel prices and making existing fleets uneconomical. The companies had purchased aircraft designed under assumptions of low fuel costs, and the jump in price upset all cost estimates. A new generation of fuel-efficient airplanes would have to be designed and built. But in a period of general economic recession—which always affects this industry in an exaggerated way—where were the airline companies to find the money to pay for the new equipment?[164]

Despite such problems, the early returns on deregulation looked very favorable. Even during the recession, passenger traffic held fairly steady, drawn by special rates. Overall rate levels climbed in response to spiraling costs—particularly of fuel—but this had nothing to do with deregulation. An authoritative early study, produced by the Harvard Faculty Project on Regulation, found that the overall results of deregulation remained strongly positive

through 1980, the cutoff date for the study. Later investigations, including an intensive analysis by the CAB itself (carried through the year 1982), confirmed the same result. In all studies, only two classes of "losers" could be identified: small cities not yet served by new commuter airlines and business travelers who received poorer service than they wanted and were willing to pay for (because of higher load factors, congested airports, and new classes of passengers).[165]

One major airline, Braniff, declared bankruptcy, but by common agreement Braniff's troubles derived from having overextended itself in an attempt to take advantage of deregulation. Under the open-entry provisions of the act of 1978, Braniff attempted to fly dozens of new routes for which it was ill equipped. In its effort to win traffic from competitors, Braniff slashed its rates, only to see other carriers match the cuts. At that point, as one Wall Street analyst remarked, "It became a bleeding contest. And Braniff bled to death faster." Kahn had predicted this kind of outcome. Poor management decisions and bankruptcies, he had argued, were routine occurrences in business. Competitive wars separated good managers from bad and afforded the opportunity for even established companies to make poor decisions. The only way to avoid all business failures was for the government to neutralize all competitive advantages, including superior management, by protecting participants and barring new entrants. Again, Kahn had been one of the few advocates of airline deregulation to engage this question openly. In the years before his arrival in Washington, fear of bankruptcies had imposed a major barrier to deregulation. Even Kahn's free-market-minded predecessor as CAB chairman, John Robson, had been reluctant to liberalize fare competition and route awards to the point that a major company might fail. Kahn himself, when Jimmy Carter's envoys urged him in 1977 to accept the CAB chairmanship, specifically asked, "What if Eastern Airlines goes bankrupt? Is the president going to pretend then that he doesn't know me?" For Kahn knew that, sooner or later, some important companies would perish. Under a system of free competition, mistakes by management could prove fatal; and, even with good management, the heavily unionized trunk carriers would be pressed hard by upstart airlines that employed nonunion labor.[166]

Still, during his tenure as CAB chairman, Kahn never had the

the study. Later investigations,
the CAB itself (carried through
e result. In all studies, only two
fied: small cities not yet served
usiness travelers who received
nd were willing to pay for (be-
sted airports, and new classes of

lared bankruptcy, but by com-
derived from having overex-
ke advantage of deregulation.
of the act of 1978, Braniff at-
s for which it was ill equipped.
competitors, Braniff slashed its
atch the cuts. At that point, as
"It became a bleeding contest.
Kahn had predicted this kind of
ions and bankruptcies, he had
in business. Competitive wars
l and afforded the opportunity
make poor decisions. The only
vas for the government to neu-
, including superior manage-
d barring new entrants. Again,
cates of airline deregulation to
he years before his arrival in
ad imposed a major barrier to
narket-minded predecessor as
een reluctant to liberalize fare
the point that a major com-
n Jimmy Carter's envoys urged
airmanship, specifically asked,
rupt? Is the president going to
v me?" For Kahn knew that,
npanies would perish. Under a
s by management could prove
ement, the heavily unionized
d by upstart airlines that em-

hairman, Kahn never had the

unpleasant duty of explaining such an actual failure. It is clear in retrospect that he came to Washington at an unusually opportune moment. Arriving in the midst of a stalled deregulation movement but with overall economic prosperity still running high, he found the setting ideal for his purposes. No airline would likely plunge into bankruptcy, at least for a while. Kahn himself saw with remarkable clarity how fortunate his timing had been: "I came into the airline situation at precisely the perfect time—with demand recovering sharply relative to capacity, with a large number of empty seats simply begging to be priced at marginal costs, with the whole political atmosphere ripe for a relaxation of unnecessary protectionist and cartelist restrictions, and I have simply been doing what, I think, is the logical thing." The same point might be applied to the timing of his earlier experience in New York. Had Kahn attempted to institute time-of-day pricing in, say, 1954 or 1964 instead of 1974, it is difficult to see how he could have succeeded.[167]

Yet good timing in no sense diminishes the importance of his contributions. In regulatory strategy as in other types of business and political strategy, timing often does determine success or failure. Adams had proved that in the 1870s, and so had James Landis in the 1930s. Kahn's triumph, like theirs, was to see the opportunity clearly and to take advantage of it, acting with full commitment and conviction.

CHAPTER 8

Regulation Reconsidered

IF THIS WERE A BOOK of pictures rather than words, it would contain three different types of photographs. The biographical sections on Adams, Brandeis, Landis, and Kahn would be detailed closeup portraits, each showing a single strategist who succeeded or failed in making regulation work. The short connecting chapters would appear as medium-range photographs depicting a regional or national landscape that was undergoing slow but significant change, such as the change through the seasons from spring to winter. Now, in this final section of the book, I want to add a third perspective: a satellite photograph of the broad regulatory scene, expressed in a few general conclusions about the history of regulation. Most of these have been foreshadowed earlier in the book; others make explicit what has so far been only implied. I bring them together here to form a brief panoramic commentary on the significance of the regulatory experience in America.

As the historical record shows, the regulatory tradition has been adapted to many different ends and purposes. Regulation has served as a versatile tool whose handle has been seized at different times by reformers, business managers, bureaucrats, and lawyers—and manipulated as often for the particular interests of one of these special groups as for the general interest of the American public.

Over time, regulation has performed not only economic tasks, but political, legal, and cultural ones as well. Among the many particular functions that regulation has been used to serve are the functions of:

TER 8

Reconsidered

ures rather than words, it would
f photographs. The biographical
dis, and Kahn would be detailed
single strategist who succeeded
ork. The short connecting chap-
nge photographs depicting a re-
hat was undergoing slow but
hange through the seasons from
al section of the book, I want to
photograph of the broad regula-
neral conclusions about the his-
have been foreshadowed earlier
what has so far been only im-
to form a brief panoramic com-
the regulatory experience in

e regulatory tradition has been
and purposes. Regulation has
ndle has been seized at different
nagers, bureaucrats, and law-
r the particular interests of one
eneral interest of the American

rmed not only economic tasks,
nes as well. Among the many
has been used to serve are the

(a) disclosure and publicity (SEC, Adams' sunshine commission);
(b) protection and cartelization of industries (ICC, FCC, CAB before Kahn);
(c) containment of monopoly and oligopoly (FPC, FTC, Antitrust Division, state utility commissions);
(d) promotion of safety for consumers and workers (Consumer Product Safety Commission, Occupational Safety and Health Administration);
(e) legitimization of the capitalist order (SEC, Environmental Protection Agency).

Obviously such a list could be longer or differently stated. The general point is that regulation has served a variety of purposes, some of which (such as items b and c above) have been mutually inconsistent.

In view of these diverse and sometimes contradictory functions, all overarching theories and heroic generalizations about "Regulation" (with a capital R) run an extremely high risk of being in error. No single theory from any academic discipline can predict precisely which industries will be regulated and which will not. Some industries that in other market economies tend toward cartelization are, in America, regulated—but not all. Some regulated industries in America have social-overhead functions and are "affected with a public interest"—but not all. Some are "natural monopolies," with sharply declining costs to scale—but not all. Much of the American transportation industry is regulated—but not all of it.

Like the application of regulation to industries, the behavior of regulatory commissions, once they have been created, shows no clear single pattern. All agencies do not follow a standard "life cycle," for example, going from youthful exuberance to middle age, then finally to geriatric decrepitude. Instead, young agencies sometimes behave sluggishly (as did the early FTC) and old ones vigorously (the CAB under Alfred Kahn). No stage-by-stage evolution, nor any other assured expectation of agency behavior, can be predicted on the basis of actual regulatory history.[1]

The single constant in the American experience with regulation has been controversy. Here the reason why is not hard to find. Many of the diverse functions assigned to regulatory commissions were regarded by legislatures as essential tasks, but very difficult for existing governmental institutions to perform. Because legisla-

tors did not wish to burden themselves with such duties, they passed the responsibility to specialized agencies. But whatever agency received the assignment was also forced to accept the intrinsic controversy that had created each task in the first place. Almost by definition, therefore, controversy became attached to regulation like a Siamese twin. For the same reason, issues common to regulatory agencies are unlikely ever to be settled, once and for all, so long as the United States remains an open, democratic society.

As in so many other aspects of American politics, the fundamental controversy underlying the history of regulation has been an ongoing need to work out the inevitable tradeoffs between the good of the whole society, on the one hand, and the rights of the individual, on the other. In regulation, these tradeoffs have appeared most clearly as ways of relieving the persistent tension between the forces seeking to implement economic efficiency for the broad benefit of American society, and those dedicated to guaranteeing the observance of legal due process for every individual member of that society. At different times in our history, each party to these fundamental tensions has established a clear advantage over the other. On balance, however, it seems clear that the concern about legal process has controlled the outcome of regulation more often than has the concern about the substance of economic efficiency. In economists' language, this means that the concern for equity has generally triumphed over the quest for efficiency. In lawyers' terms, it means that in regulation the judicial model has usually triumphed over the legislative and administrative model. In cultural terms, it means that the concern for fairness and for the protection of the diverse interests of all affected individuals has most often won out over the concern for overall growth in the national economy. More generally in political terms, it means that regulation is best understood as a political settlement, undertaken in an effort to keep peace within the polity.

Overall, the conclusion appears inescapable that regulation in America has more often functioned as a protective device rather than as a promotional or developmental one. Of course, protection was not always inappropriate. By holding in check socially destructive forms of behavior, protective regulation often cushioned the impact of rapid industrial change. In America, in contrast to older societies, so many other forces consistently acted to promote

pell-mell economic growth that regulation can hardly be condemned for not always doing so.

Because the appropriate balance between economic efficiency and legal due process has seldom been self-evident, individual persons and particular ideas have mattered a great deal in regulatory history. Thus, most of this book has explored the roles of particular persons, and I rest that part of my case on the biographies of Adams, Brandeis, Landis, and Kahn. The more general role of ideas in the history of regulation, however, deserves some additional comment. As the political scientist James Q. Wilson has written, "We must be struck at every turn by the importance of ideas. Regulation itself is such an idea; deregulation is another . . . To the extent [that] an agency can choose, its choices will be importantly shaped by what its executives learned in college a decade or two earlier."[2]

Ideas about regulation, as Wilson implies, vary with time. During the 1930s, national policymakers generally held the powerful conviction that market mechanisms left to themselves would produce widespread injustice and even inefficiency. Hence they believed that an active federal government was essential for the protection of the public interest. So these political activists created a broad portfolio of new, independent agencies: the SEC, the FCC, the CAB, and so on. A few decades later, during the 1960s and 1970s, a new generation of policymakers embraced a very different idea. Rather than applauding the old activism, they became convinced that many of the independent commissions created during the 1930s had since been captured by the very interests that these agencies had been set up to regulate.

Partly as a result of the capture idea, there arose during the 1960s a curious two-pronged reform movement: pointing, on the one hand, toward deregulation and, on the other, toward a new wave of large-scale social and environmental regulation. These new rules were to be enforced not by independent commissions of the 1930s variety, which usually administered brief general statutes designed to give broad discretion to a group of commissioners acting collegially; but rather by an entirely different type of agency, with a single executive at its head (who could be held individually responsible for success or failure) and an agenda set in advance by the explicit provisions of extremely detailed legislation. New laws such as the Clean Air Act and the Occupational Safety

and Health Act, often running to scores of pages in length, were calculated to minimize administrative discretion and to close all possible loopholes. Meanwhile, on the other prong, the deregulation movement—whose basic intellectual premise was that economic markets *do* work well—also advanced, simultaneously but contradictorily, gaining momentum alongside the companion movement toward growth of regulation in the areas of social and environmental policy.

The result, by the 1980s, presented a most peculiar spectacle. In an ironic historical example of the ways in which ideas can move policymakers in opposite directions, significant deregulation had been instituted for such industries as airlines, trucking, railroads, financial markets, and telecommunications. At the same time, additional social and environmental regulation had become firmly embedded in the structure of state and federal government in such a form as to make any capture by regulated interests very difficult, if not impossible.

The movement of ideas alone, of course, had not produced this ironic result. Despite the power of thought in the history of regulation, ideas in themselves could not determine concrete outcomes. Instead, ideas had to interact with particular economic and political circumstances to form a reciprocal relationship in which one or both might be altered. Nor, in any absolute sense, did the ideas themselves have to be demonstrably true in order to exert strong influence. We have seen, for instance, how Louis Brandeis' flawed idea of competition moved the hearts of his contemporaries. To cite a second example, the disparate sets of ideas underlying the initial imposition of regulation in airlines and trucking during the 1930s, and the later deregulation of these same industries in the 1970s and 1980s, could not both have been correct, in the absolute sense. Yet both sets of ideas became institutionalized. What had changed was the historical context in which opposing ideas about the legitimacy and actual performance of economic markets were defined.

During the 1930s, a period not only of depression but also of economic *de*flation, policymakers searched for some way to stabilize prices, as a means of preventing further economic decline. By the 1970s, however, deflation no longer provided the historical filter; indeed, it had become almost inconceivable as a problem for policy to solve. Instead, *in*flation was now the pressing issue, and thus the same protectionist regulations that had been applied in

scores of pages in length, were
rative discretion and to close all
n the other prong, the deregula
ellectual premise was that eco
so advanced, simultaneously but
tum alongside the companion
ulation in the areas of social and

ted a most peculiar spectacle. In
e ways in which ideas can move
ons, significant deregulation had
s as airlines, trucking, railroads,
nications. At the same time, ad
l regulation had become firmly
and federal government in such
regulated interests very difficult,

of course, had not produced this
f thought in the history of regu
d not determine concrete out
ct with particular economic and
reciprocal relationship in which
r, in any absolute sense, did the
onstrably true in order to exert
or instance, how Louis Brandeis'
ved the hearts of his contem
e, the disparate sets of ideas un
of regulation in airlines and
the later deregulation of these
980s, could not both have been
both sets of ideas became insti
was the historical context in
legitimacy and actual perform
efined.
only of depression but also of
searched for some way to stabi
ng further economic decline. By
onger provided the historical fil
inconceivable as a problem for
was now the pressing issue, and
ations that had been applied in

the 1930s to combat deflation now seemed inappropriate to the new economic context. Both ideas remained alive, but a different time meant a different choice for public policy. To state the same point in a more general sense, it is clear that in American history both the producer-oriented protectionist tradition, on the one hand, and the consumer-oriented anticartelist tradition on the other have remained hostage to immediate economic conditions. The strengths of each tradition have ebbed and flowed in response to several external forces: the business cycle, the different degrees of maturity reached by different product markets, and the conditions of international war and peace.

In speculating about the future, it is difficult to foresee with much additional precision what new historical contexts for regulation might develop. But if the past is any guide, a good deal of caution is in order. What I have called in this book the "economist's hour" of the 1970s and 1980s, for example, represents a phenomenon of unpredictable duration. Certainly the economist's hour in the history of regulation came relatively late, long after other notably different hours during which the muckraker and the lawyer alternately held center stage. This history makes it seem unlikely that any single approach to regulation will ever triumph. Therefore, although we may live in the golden years of regulatory economics and its practitioners, we should be in no hurry to crown the economist as permanent king of the regulatory hill.

Economic analysis, however, will always remain directly relevant to regulatory policy. This is true because every industry, whether regulated or not, does possess a certain underlying economic structure: characteristics that make it different from other industries and that help to shape the internal conditions for regulatory opportunities and constraints. *More than any other single factor, this underlying structure of the particular industry being regulated has defined the context in which regulatory agencies have operated.* Sometimes the differences between industry structures can be radical: the railroad industry, with its huge fixed costs and enormous scale economies, could hardly differ more fundamentally from the securities industry, with its paper assets and labor-intensive structure. In other cases the differences can be more subtle, as in the contrast between center and peripheral firms, and the related distinction between tight and loose forms of horizontal combination.

Because the underlying characteristics of the industry so often

shape the limits of governmental action, the industry may be regarded as the dog, the regulatory agency only as the tail. Yet many students of regulation have assumed that tails wag dogs and, further, that one standard type of tail can wag whatever breed of dog may be attached. Such observers, by focusing primarily on the similarities of regulatory commissions (most of which were bipartisan, appointed by the executive, expert in their fields, and so on), have missed a larger truth: the industries that these similarly structured commissions regulated were extremely diverse. Thus these observers have duplicated the errors made historically by many regulators themselves, who often paid more attention to legal processes and administrative procedures than to the greater task of framing strategies appropriate to the particular industries they were regulating. For all parties who seek to understand regulation, the most important single consideration is the appropriateness of the regulatory strategy to the industry involved.[3]

The process of fitting regulatory strategies to particular industries is a difficult task, partly because industrial structures, like regulatory ideas, can change over time. The railroad industry represented a true natural monopoly when the Interstate Commerce Commission first emerged to regulate it during the 1880s. But several decades later, this natural monopoly status of railroads had disappeared before the rise of alternative modes of transportation—the automobile, truck, and airplane. Clearly, by the 1930s, some central assumptions behind the whole scheme of national railroad regulation needed to be revised, and regulatory policy adjusted accordingly. Yet until the 1970s little was done, as assumptions remained those of the 1880s. In the meantime, the Interstate Commerce Commission kept rates inflexible, prevented industry rationalization by blocking truck-rail mergers, and delayed for years the widespread application of unit-train technology. The ICC took so much time to recognize the revolutionary changes in the transportation industries that regulatory policy lagged badly behind market reality, causing unintended injury to the very railroad industry the commission was trying to protect.

Elsewhere, a similar process of unacknowledged change occurred in the communications industries. When Congress created the Federal Communications Commission in 1934, legislators acted on an assumption, valid at the time, that the electromagnetic spectrum (what physicists used to call the "ether") was fi-

action, the industry may be re
gency only as the tail. Yet many
ed that tails wag dogs and, fur
can wag whatever breed of dog
by focusing primarily on the
ons (most of which were bipar
expert in their fields, and so on),
ndustries that these similarly
were extremely diverse. Thus
ne errors made historically by
often paid more attention to
procedures than to the greater
ate to the particular industries
s who seek to understand regu
nsideration is the appropriate-
ne industry involved.[3]
strategies to particular indus-
ause industrial structures, like
me. The railroad industry rep-
vhen the Interstate Commerce
te it during the 1880s. But sev-
nopoly status of railroads had
ernative modes of transporta-
rplane. Clearly, by the 1930s,
he whole scheme of national
evised, and regulatory policy
1970s little was done, as as-
80s. In the meantime, the In-
pt rates inflexible, prevented
g truck-rail mergers, and de-
lication of unit-train technol-
o recognize the revolutionary
ustries that regulatory policy
causing unintended injury to
nission was trying to protect.
unacknowledged change oc-
tries. When Congress created
mmission in 1934, legislators
ne time, that the electromag-
d to call the "ether") was fi-

nite. The lawmakers reasoned that if too many stations used the ether at once, radio signals would become garbled and communication would be impossible. Thus competitors had to be limited in number, through regulatory allocation of different parts of the spectrum to different broadcasting stations. Later on, however, new technology completely altered the situation. Through the use of satellite communications and dish receivers, community television companies in effect created vast new amounts of ether, thereby accommodating dozens of additional broadcast stations. When that occurred, the market power of the three major television networks diminished rapidly, and television broadcasting as an industry began to drift away from center status and toward peripheral.

To cite still another example, a situation of natural monopoly prevailed for many years in long-distance telephoning, based on the once-valid principle that a single set of transcontinental wires could most economically serve consumers' needs. But in the 1960s and 1970s, a technological revolution in microwave communication destroyed that premise and ended the natural monopoly in long-distance telephoning. As the significance of this technological shift became clear, new companies emerged to compete for long-distance telephone business, and just as quickly a movement to deregulate the telephone industry began. By 1984, the giant Bell System, once the most conspicuous natural monopoly in America, had divided itself into a number of much smaller companies. What had been a thoroughgoing center industry (telephoning) now became part center (local telephone service, heavy equipment manufacturing) and part peripheral (long-distance service, high-technology research and development, and light manufacturing). As a result of these seismic changes in industry structure, the regulation of telecommunications became during the 1980s one of the most publicized and controversial problems of state and federal government.[4]

For scholars, the evolution of both the telephone industry and television broadcasting provided outstanding examples of the protean nature of industrial organization. Not since the early emergence of center firms out of the advanced production technology of the late nineteenth century had there appeared such clear instances of the relationship between scientific change and subsequent shifts in the structure of existing industries. Traditionally,

such shifts had been from peripheral to center status, but the communications revolution showed how the movement could work in the opposite direction as well. The implications for regulators were profound, and the task of framing effective regulatory strategies, always one of the most difficult of governmental arts, became in these industries still more challenging.

In thinking about the future of regulation, whether in broadcasting, telephones, or any other industry, it is important to keep in mind the ambiguous record of the past. Even though much of regulatory history is tinged with apparent failure, regulation cannot properly be said either to have "failed" or "succeeded" in an overall historical sense. Instead, individual regulatory experiments and episodes must be judged against a standard true to the particular historical moment. Many observers hold a contrary view and insist on a single overriding verdict of failure. Because of this prevailing opinion in our time, it becomes useful to speculate about ways in which the same judgment might be applied to other parts of government. Can it be said with equal justice, for example, that legislation in general has failed historically, or that the court system has failed, or that the office of the presidency has failed—without specifying exactly which legislation, which court on what case, and which president on what issue? Although the answer might seem self-evident, the fact remains that in popular perceptions over the last three decades regulation has been regarded as a synonym for failure. Even in some of the best scholarship on regulation, failure has often been applied not merely as a conclusion but also as a premise, a tacit assumption hidden behind apparently scholarly explanations presented in theoretical forms: the theories of capture, of public choice, of taxation by regulation, and several others.[5]

To weigh against these multiple theories premised on failure, we have only one premised on success. But it is a very useful one: the theory of "public use of private interest." According to this idea, regulators should always exploit the natural incentives of regulated interests to serve particular goals that the regulators themselves have carefully defined in advance.[6] And, in fact, the historical record suggests that regulation in America has succeeded best when it has respected these incentives instead of ignoring them; when it has based its strategies less on some idealized

ral to center status, but the com-
ow the movement could work in
The implications for regulators
aming effective regulatory strate-
ult of governmental arts, became
lenging.

egulation, whether in broadcast-
ustry, it is important to keep in
e past. Even though much of reg-
parent failure, regulation cannot
iled" or "succeeded" in an over-
dual regulatory experiments and
a standard true to the particular
rs hold a contrary view and insist
ailure. Because of this prevailing
seful to speculate about ways in
be applied to other parts of gov-
l justice, for example, that legis-
rically, or that the court system
e presidency has failed—without
ion, which court on what case,
ue? Although the answer might
ns that in popular perceptions
tion has been regarded as a syn-
f the best scholarship on regula-
d not merely as a conclusion but
otion hidden behind apparently
n theoretical forms: the theories
xation by regulation, and several

le theories premised on failure,
ccess. But it is a very useful one:
ate interest." According to this
xploit the natural incentives of
icular goals that the regulators
l in advance.[6] And, in fact, the
egulation in America has suc-
d these incentives instead of ig-
strategies less on some idealized

vision of what the economy should do and more on a clear understanding of what the economy actually is doing. Regulatory strategies framed in ignorance or disregard of real economic conditions and market incentives—and the number of such attempts has been legion—have usually led only to unfortunate results. By contrast, strategies framed with these conditions and incentives in mind have often produced strikingly successful outcomes, as the stories of Adams, Landis, and Kahn demonstrate so clearly.

Ronald H. Coase

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 1991

Prize Lecture

Lecture to the memory of Alfred Nobel, December 9, 1991

The Institutional Structure of Production

In my long life I have known some great economists but I have never counted myself among their number nor walked in their company. I have made no innovations in high theory. My contribution to economics has been to urge the inclusion in our analysis of features of the economic system so obvious that, like the postman in G.K. Chesterton's Father Brown tale, *The Invisible Man*, they have tended to be overlooked. Nonetheless, once included in the analysis, they will, as I believe, bring about a complete change in the structure of economic theory, at least in what is called price theory or microeconomics. What I have done is to show the importance for the working of the economic system of what may be termed the institutional structure of production. In this lecture I shall explain why, in my view, these features of the economic system were ignored and why their recognition will lead to a change in the way we analyse the working of the economic system and in the way we think about economic policy, changes which are already beginning to occur. I will also speak about the empirical work that needs to be done if this transformation in our approach is to increase our understanding. In speaking about this transformation, I do not wish to suggest that it is the result of my work alone. Oliver Williamson, Harold Demsetz, Steven Cheung, among others, have made outstanding contributions to the subject and without their work and that of many others, I doubt whether the significance of my writings would have been recognized. While it has been a great advantage of the creation of the Prize in Economic Sciences in Memory of Alfred Nobel that, by drawing attention to the significance of particular fields of economics, it encourages further research in them, the highlighting of the work of a few scholars, or, in my case, one scholar, tends to obscure the importance of the contributions of other able scholars whose researches have been crucial to the development of the field.

I will be speaking of that part of economics which has come to be called industrial organization, but to understand its present state it is necessary to say something about the development of economics in general. During the two centuries since the publication of *The Wealth of Nations* the main activity of economists, it seems to me, has been to fill the gaps in Adam Smith's system, to correct his errors and to make his analysis vastly more exact. A principal theme of *The Wealth of Nations* was that government regulation or centralised planning were not necessary to make an economic system function in an orderly way. The economy could be co-ordinated by a system of prices (the "invisible hand") and, furthermore, with beneficial results. A major task of economists since the publication of *The Wealth of Nations*, as Harold Demsetz has explained[1], has been to formalize this proposition of Adam Smith. The given factors are technology and the tastes of consumers, and individuals, who follow their own interest, are governed in their choices by a system of prices. Economists have uncovered the conditions necessary if Adam Smith's results are to be achieved and where, in the real world, such conditions do not appear to be found, they have proposed changes which are designed to bring them about. It is what one finds in the

textbooks. Harold Demsetz has said rightly that what this theory analyses is a system of extreme decentralization. It has been a great intellectual achievement and it throws light on many aspects of the economic system. But it has not been by any means all gain. The concentration on the determination of prices has led to a narrowing of focus which has had as a result the neglect of other aspects of the economic system. Sometimes, indeed, it seems as though economists conceive of their subject as being concerned only with the pricing system and that anything outside this is considered as no part of their business. Thus, my old chief and wonderful human being, Lionel Robbins, wrote, in *The Nature and Significance of Economic Science*, about the "glaring deficiencies" of the old treatment of the theory of production with its discussion of peasant proprietorships and industrial forms: "It suggests that from the point of view of the economist 'organisation' is a matter of internal industrial (or agricultural) arrangement - if not internal to the firm, at any rate internal to 'the' industry. At the same time it tends to leave out completely the governing factor of all productive organisation - the relationship of prices and cost..."[2]. What this comes down to is that, in Robbins' view, an economist does not interest himself in the internal arrangements within organisations but only in what happens on the market, the purchase of factors of production and the sale of the goods that these factors produce. What happens in between the purchase of the factors of production and the sale of the goods that are produced by these factors is largely ignored. I do not know how far economists today share Robbins' attitude but it is undeniable that microeconomics is largely a study of the determination of prices and output, indeed this part of economics is often called price theory.

This neglect of other aspects of the system has been made easier by another feature of modern economic theory - the growing abstraction of the analysis, which does not seem to call for a detailed knowledge of the actual economic system or, at any rate, has managed to proceed without it. Holmstrom and Tirole, writing on *The Theory of the Firm* in the recently published *Handbook of Industrial Organization*, conclude at the end of their article of 63 pages that "the evidence/theory ratio... is currently very low in this field"[3]. Peltzman has written a scathing review of the *Handbook* in which he points out how much of the discussion in it is theory without any empirical basis[4]. What is studied is a system which lives in the minds of economists but not on earth. I have called the result "blackboard economics". The firm and the market appear by name but they lack any substance. The firm in mainstream economic theory has often been described as a "black box". And so it is. This is very extraordinary given that most resources in a modern economic system are employed within firms, with how these resources are used dependent on administrative decisions and not directly on the operation of a market. Consequently, the efficiency of the economic system depends to a very considerable extent on how these organisations conduct their affairs, particularly, of course, the modern corporation. Even more surprising, given their interest in the pricing system, is the neglect of the market or more specifically the institutional arrangements which govern the process of exchange. As these institutional arrangements determine to a large extent what is produced, what we have is a very incomplete theory. All this is beginning to change and in this process I am glad to have played my part. The value of including such institutional factors in the corpus of mainstream economics is made clear by recent events in Eastern Europe. These ex-communist countries are advised to move to a market economy, and their leaders wish to do so, but without the appropriate institutions no market economy of any significance is possible. If we knew more about our own economy we would be in a better position to advise them.

What I endeavoured to do in the two articles cited by the Royal Swedish Academy of Sciences was to attempt to fill these gaps or more exactly to indicate the direction in which we should move if they are ultimately to be filled. Let me start with *The Nature of the Firm*. I went as a student to the London School of Economics in 1929 to study for a Bachelor of Commerce degree, specialising in the Industry group, supposedly designed for people who wished to become works managers, a choice of occupation for which I was singularly ill-suited. However, in 1931, I had a great stroke of luck. Arnold Plant was appointed Professor of Commerce in 1930. He was a wonderful teacher. I began to attend his seminar in 1931, some five months before I took the final examinations. It was a revelation. He quoted Sir Arthur Salter: "The normal economic system works itself'. And he explained how a competitive economic system coordinated by prices would lead to the production of goods and services which consumers valued most highly. Before being exposed to Plant's teaching, my notions on how the economy worked were extremely woolly. After Plant's seminar I had a coherent view of the economic system. He introduced me to Adam Smith's "invisible hand". As I had taken the first year of University work while still at High School, I managed to complete the requirements for a degree in two years. However, university regulations required three years of residence before a degree could be granted. I had therefore a year to spare. I then had another stroke of luck. I was awarded a Cassel travelling scholarship by the University of London. I decided to spend the year in the United States, this being treated as a year's residence at the London School of Economics, the regulations being somewhat loosely interpreted.

I decided to study vertical and lateral integration of industry in the United States. Plant had described in his lectures the different ways in which various industries were organised, but we seemed to lack any theory which would explain these differences. I set out to find it. There was also another puzzle which, in my mind, needed to be solved and which seemed to be related to my main project. The view of the pricing system as a co-ordinating mechanism was clearly right but there were aspects of the argument which troubled me. Plant was opposed to all schemes, then very fashionable during the Great Depression, for the co-ordination of industrial production by some form of planning. Competition, according to Plant, acting through a system of prices, would do all the co-ordination necessary. And yet we had a factor of production, management, whose function was to co-ordinate. Why was it needed if the pricing system provided all the co-ordination necessary? The same problem presented itself to me at that time in another guise. The Russian Revolution had taken place only fourteen years earlier. We knew then very little about how planning would actually be carried out in a communist system. Lenin had said that the economic system in Russia would be run as one big factory. However, many economists in the West maintained that this was an impossibility. And yet there were factories in the West and some of them were extremely large. How did one reconcile the views expressed by economists on the role of the pricing system and the impossibility of successful central economic planning with the existence of management and of these apparently planned societies, firms, operating within our own economy?[5]

I found the answer by the summer of 1932. It was to realise that there were costs of using the pricing mechanism. What the prices are has to be discovered. There are negotiations to be undertaken, contracts have to be drawn up, inspections have to be made, arrangements have to be made to settle disputes, and so on. These costs have come to be known as transaction costs. Their existence implies that methods of co-ordination, alternative to the market, which are themselves costly and in various ways imperfect, may nonetheless be preferable to relying on the pricing

mechanism, the only method of co-ordination normally analysed by economists. It was avoidance of the costs of carrying out transactions through the market that could explain the existence of the firm in which the allocation of factors came about as a result of administrative decisions (and I thought it did). In my 1937 article, I argued that in a competitive system there would be an optimum of planning since a firm, that little planned society, could only continue to exist if it performed its co-ordination function at a lower cost than would be incurred if it were achieved by means of market transactions and also at a lower cost than this same function could be performed by another firm. To have an efficient economic system it is necessary not only to have markets but also areas of planning within organizations of the appropriate size. What this mix should be we find as a result of competition. This is what I said in my article of 1937. However, as we know from a letter I wrote in 1932, which has been preserved, all the essentials of this argument had been presented in a lecture I gave in Dundee at the beginning of October, 1932[6]. I was then twenty-one years of age and the sun never ceased to shine. I could never have imagined that these ideas would become some 60 years later a major justification for the award of a Nobel Prize. And it is a strange experience to be praised in my eighties for work I did in my twenties.

There is no doubt that the recognition by economists of the importance of the role of the firm in the functioning of the economy will prompt them to investigate its activities more closely. The work of Oliver Williamson and others has led to a greater understanding of the factors which govern what a firm does and how it does it. And we can also hope to learn much more in future from the studies of the activities of firms which have recently been initiated by the Center for Economic Studies of the Bureau of the Census of the United States. But it would be wrong to think that the most important consequence for economics of the publication of *The Nature of the Firm* has been to direct attention to the importance of the firm in our modern economy, a result which, in my view, would have come about in any case. What I think will be considered in future to have been the important contribution of this article is the explicit introduction of transaction costs into economic analysis. I argued in *The Nature of the Firm* that the existence of transaction costs leads to the emergence of the firm. But the effects are pervasive in the economy. Businessmen in deciding on their ways of doing business and on what to produce have to take into account transaction costs. If the costs of making an exchange are greater than the gains which that exchange would bring, that exchange would not take place and the greater production that would flow from specialisation would not be realised. In this way, transaction costs affect not only contractual arrangements but also what goods and services are produced. Not to include transaction costs in the theory leaves many aspects of the working of the economic system unexplained, including the emergence of the firm, but much else besides. In fact, a large part of what we think of as economic activity is designed to accomplish what high transaction costs would otherwise prevent or to reduce transaction costs so that individuals can freely negotiate and we can take advantage of that diffused knowledge of which Hayek has told us.

I know of only one part of economics in which transaction costs have been used to explain a major feature of the economic system and that relates to the evolution and use of money. Adam Smith pointed out the hindrances to commerce that would arise in an economic system in which there was a division of labour but in which all exchange had to take the form of barter. No-one would be able to buy anything unless he possessed something that the producer wanted. This difficulty, he explained, could be overcome by the use of money. A person wishing to buy something in a barter system has to find someone who has this product for sale but who also wants some of the

goods possessed by the potential buyer. Similarly, a person wishing to sell something has to find someone who both wants what he has to offer and also possesses something that the potential seller wants. Exchange in a barter system requires what Jevons called "this double coincidence". Clearly the search for partners in exchange with suitable qualifications is likely to be very costly and will prevent many potentially beneficial exchanges from taking place. The benefit brought about by the use of money consists of a reduction in transaction costs. The use of money also reduces transaction costs by facilitating the drawing up of contracts as well as by reducing the quantity of goods that need to be held for purposes of exchange. However, the nature of the benefits secured by the use of money seems to have faded into the background so far as economists are concerned and it does not seem to have been noticed that there are other features of the economic system which exist because of the need to mitigate transaction costs.

I now turn to that other article cited by the Swedish Academy, *The Problem of Social Cost*, published some 30 years ago. I will not say much here about its influence on legal scholarship which has been immense but will mainly consider its influence on economics, which has not been immense, although I believe that in time it will be. It is my view that the approach used in that article will ultimately transform the structure of microeconomics - and I will explain why. I should add that in writing this article I had no such general aim in mind. I thought that I was exposing the weaknesses of Pigou's analysis of the divergence between private and social products, an analysis generally accepted by economists, and that was all. It was only later, and in part as a result of conversations with Steven Cheung in the 1960s that I came to see the general significance for economic theory of what I had written in that article and also to see more clearly what questions needed to be further investigated.

Pigou's conclusion and that of most economists using standard economic theory was, and perhaps still is, that some kind of government action (usually the imposition of taxes) was required to restrain those whose actions had harmful effects on others, often termed negative externalities. What I showed in that article, as I thought, was that in a regime of zero transaction costs, an assumption of standard economic theory, negotiations between the parties would lead to those arrangements being made which would maximise wealth and this irrespective of the initial assignment of rights. This is the infamous *Coase Theorem*, named and formulated by Stigler, although it is based on work of mine. Stigler argues that the Coase Theorem follows from the standard assumptions of economic theory. Its logic cannot be questioned, only its domain[7]. I do not disagree with Stigler. However, I tend to regard the Coase Theorem as a stepping stone on the way to an analysis of an economy with positive transaction costs. The significance to me of the Coase Theorem is that it undermines the Pigovian system. Since standard economic theory assumes transaction costs to be zero, the Coase Theorem demonstrates that the Pigovian solutions are unnecessary in these circumstances. Of course, it does not imply, when transaction costs are positive, that government actions (such as government operation, regulation or taxation, including subsidies) could not produce a better result than relying on negotiations between individuals in the market. Whether this would be so could be discovered not by studying imaginary governments but what real governments actually do. My conclusion; let us study the world of positive transaction costs.

If we move from a regime of zero transaction costs to one of positive transaction costs, what becomes immediately clear is the crucial importance of the legal system in this new world. I explained in *The Problem of Social Cost* that what are traded on the market are not, as is often supposed by economists, physical entities but the rights to

perform certain actions and the rights which individuals possess are established by the legal system. While we can imagine in the hypothetical world of zero transaction costs that the parties to an exchange would negotiate to change any provision of the law which prevents them from taking whatever steps are required to increase the value of production, in the real world of positive transaction costs such a procedure would be extremely costly, and would make unprofitable, even where it was allowed, a great deal of such contracting around the law. Because of this, the rights which individuals possess, with their duties and privileges, will be, to a large extent what the law determines. As a result the legal system will have a profound effect on the working of the economic system and may in certain respects be said to control it. It is obviously desirable that these rights should be assigned to those who can use them most productively and with incentives that lead them to do so and that, to discover and maintain such a distribution of rights, the costs of their transference should be low, through clarity in the law and by making the legal requirements for such transfers less onerous. Since this can come about only if there is an appropriate system of property rights, and they are enforced, it is easy to understand why so many academic lawyers (at least in the United States) have found so attractive the task of uncovering the character of such a property rights system and why the subject of "law and economics" has flourished in American law schools. Indeed, work is going forward at such a pace that I do not consider it over-optimistic to believe that the main outlines of the subject will be drawn within five or ten years.

Until quite recently most economists seem to have been unaware of this relationship between the economic and legal systems except in the most general way. Stock and produce exchanges are often used by economists as examples of perfect or near-perfect competition. But these exchanges regulate in great detail the activities of traders (and this quite apart from any public regulation there may be). What can be traded, when it can be traded, the terms of settlement and so on are all laid down by the authorities of the exchange. There is, in effect, a private law. Without such rules and regulations, the speedy conclusion of trades would not be possible. Of course, when trading takes place outside exchanges (and this is almost all trading) and where the dealers are scattered in space and have very divergent interests, as in retailing and wholesaling, such a private law would be difficult to establish and their activities will be regulated by the laws of the State. It makes little sense for economists to discuss the process of exchange without specifying the institutional setting within which the trading takes place since this affects the incentives to produce and the costs of transacting. I think this is now beginning to be recognized and has been made crystal-clear by what is going on in Eastern Europe today. The time has surely gone in which economists could analyze in great detail two individuals exchanging nuts for berries on the edge of the forest and then feel that their analysis of the process of exchange was complete, illuminating though this analysis may be in certain respects. The process of contracting needs to be studied in a real world setting. We would then learn of the problems that are encountered and of how they are overcome and we would certainly become aware of the richness of the institutional alternatives between which we have to choose.

Oliver Williamson has ascribed the non-use or limited use of my thesis in *The Nature of the Firm* to the fact that it has not been made "operational", by which he means that the concept of transaction costs has not been incorporated into a general theory. I think this is correct. There have been two reasons for this. First, incorporating transaction costs into standard economic theory which has been based on the assumption that they are zero, would be very difficult and economists who, like most scientists, as Thomas Kuhn has told us, are extremely conservative

in their methods, have not been inclined to attempt it. Second, Williamson has also pointed out that although I was correct in making the choice between organization within the firm or through the market the center piece of my analysis, I did not indicate what the factors were that determined the outcome of this choice and thus made it difficult for others to build on what is often described as a "fundamental insight". This also is true. But the interrelationships which govern the mix of market and hierarchy, to use Williamson's terms, are extremely complex and in our present state of ignorance it will not be easy to discover what these factors are. What we need is more empirical work. In a paper written for a conference of the National Bureau of Economic Research I explained why I thought this was so. This is what I said: "An inspired theoretician might do as well without such empirical work, but my own feeling is that the inspiration is most likely to come through the stimulus provided by the patterns, puzzles and anomalies revealed by the systematic gathering of data, particularly when the prime need is to break our existing habits of thought"*. This statement was made in 1970. I still think that in essentials it is true today. Although much interesting and important research was done in the seventies and eighties and we certainly know much more than we did in 1970, there is little doubt that a great deal more empirical work is needed. However, I have come to the conclusion that the main obstacle faced by researchers in industrial organization is the lack of available data on contracts and the activities of firms. I have therefore decided to do something about it. Believing that there is a great deal of data on contracts and the activities of firms in the United States available in government departments and agencies in Washington, D.C., and that this information is largely unknown to economists, I organized a conference at the University of Chicago Law School in the Summer of 1990 at which government officials presented papers in which they described what data was available and how to get access to it and also reported on some of the research being carried out within their departments. The audience consisted of academic economists. It was, as a colleague remarked, a case of supply meeting demand. The proceedings of this conference will be published in a special issue of the *Journal of Law and Economics*. Another development with which I am associated is the establishment of the Center for the Study of Contracts and the Structure of Enterprise at the Business School of the University of Pittsburgh. This Center will make large-scale collections of business contracts and will prepare databases which will be made available to all researchers, whatever their institution. Nor should we forget the work now getting started at the Center for Economic Studies of the Bureau of the Census. This greater availability of data and the encouragement given to all researchers working on the institutional structure of production by the award to me of the Nobel Prize, should result in a reduction in that elegant but sterile theorizing so commonly found in the economics literature on industrial organization and should lead to studies which increase our understanding of how the real economic system works.

My remarks have sometimes been interpreted as implying that I am hostile to the mathematization of economic theory. This is untrue. Indeed, once we begin to uncover the real factors affecting the performance of the economic system, the complicated interrelations between them will clearly necessitate a mathematical treatment, as in the natural sciences, and economists like myself, who write in prose, will take their bow. May this period soon come.

I am very much aware that many economists whom I respect and admire will not agree with the opinions I have expressed and some may even be offended by them. But a scholar must be content with the knowledge that what is false in what he says will soon be exposed and, as for what is true, he can count on ultimately seeing it accepted, if only he lives long enough.

1. Harold Demsetz, *Ownership, Control and the Firm* Volume I, page 145.

2. Lionel Robbins, *The Nature and Significance of Economic Science* (1932), page 70.

3. Richard Schmalensee and Robert D. Willig (editors), *Handbook of Industrial Organization*, page 126.

4. Sam Peltzman, "The Handbook of Industrial Organization: A Review Article", *Journal of Political Economy*, February, 1991, pages 201-217.

5. A fuller account of these events will be found in Oliver E. Williamson and Sidney G. Winter (editors), *The Nature of the Firm, Origins, Evolution and Development*, pages 34-47.

6. Ibid, pages 34-35.

7. George J. Stigler, Two Notes on the Coase Theorem", *Yale Law Journal*, December, 1989, pages 631-633.

8. R.H. Coase, *The Firm, the Market and the Law*, page 71.

From *Nobel Lectures, Economics 1991-1995*, Editor Torsten Persson, World Scientific Publishing Co., Singapore, 1997

Public Choice has demystified and undeified the state.

The Public Choice Revolution

By Pierre Lemieux
Université du Québec en Outaouais

A half-century ago, an orthodox economist would approach the analysis of public policy with the following reasoning: Markets are efficient, or "Pareto-optimal," when perfect competition prevails. Pareto optimality means that there is no way to reallocate inputs or outputs to benefit some individual without harming another individual or, thought of another way, all gains from exchange have been realized. In many cases, such results are precluded by different types of "market failure" like macroeconomic imbalances, natural monopoly, or externalities (positive or negative). Positive externalities can be generated by "public goods," which provide benefits to everybody as long as the goods are produced and consumed by somebody. Government must intervene to correct market failures and maximize social welfare.

That was policy analysis before the public choice revolution. Today, the view is much different and begins with a simple question: How are collective decisions made? The answer, of course, is that the decisions are made by policymakers — politicians and bureaucrats — and by voters. The starting idea of public choice theory is disarmingly simple: Individuals, when acting as voters, politicians, or bureaucrats, continue to be self-interested and try to maximize their utility.

Excluding immediate precursors like Anthony Downs' 1957 book *An Economic Theory of Democracy* and Duncan Black's 1958 book *The Theory of Committees and Elections*, the foundation of the public choice school can probably be dated to the 1962 publication of James Buchanan and Gordon Tullock's *The Calculus of Consent*. Many well-known public choice economists were congregating around Buchanan and Tullock at Virginia Tech at that time: Geoffrey Brennan, Robert D. Tollison, Richard E. Wagner, Winston Bush, and others. For his seminal work in public choice, Buchanan was awarded the 1986 Nobel Prize.

In a narrow sense, public choice analysis is concerned with "state failures." Manned by self-interested actors on a "political market," the state is often incapable of correcting market failures — or, at least, of correcting them at a lower price than the cost of the original market failures themselves. In a wider sense, public choice is, as Dennis Mueller writes in his book *Public Choice III*, "the economic analysis of political institutions." In this broad sense, virtually all economists who study government intervention have now become public choice economists.

THE STATE

Why do we need the state to provide certain goods and services? Why not just have anarchy and let everyone fend for himself either individually or as a member of a private group? The subtitle of James Buchanan's seminal 1975 book *The Limits of Liberty* summarizes where the individuals presumably want to be: "Between Anarchy and Leviathan." In the mainstream public choice perspective, the state is necessary to stop the Hobbesian "war of all against all." As Mancur Olson puts it, a "sedentary bandit," the state, generates more prosperity than the "roving bandits" it puts out of business.

Once it is admitted that the state is necessary, positive public choice analyzes how it assumes its missions of allocative efficiency and redistribution. Normative public choice tries to identify institutions conducive to individuals getting from the state what they want without being exploited by it.

The contractarian approach defended by many public choice theorists is part of the normative leg of public choice theory. It distinguishes a "constitutional stage" in which, conceptually, individuals unanimously accept the rules of the political game, and a "post-constitutional stage" in which the rules of day-to-

Pierre Lemieux is an economist at the Department of Management Sciences of the Université du Québec en Outaouais, and a research fellow at the Independent Institute. He may be contacted by e-mail at pierre.lemieux@uqo.ca.

Why the Majority Wants Both A and Not-A

Assume a society composed of three voters: X, Y, and Z. Further, assume that they are deliberating over three mutually exclusive policy proposals: A, B, and C.

Table 1 shows each voter's order of preference among the three proposals. For example, Voter X's favored proposal is A; his second preference is B, and his least preferred policy is C. The other voters have different preferences. Note that we are only assuming ordinal utility, which means that only the rank of the alternatives matters.

Now, suppose that a referendum asks the voters to choose, under a majority rule, between proposals A and B. Proposal A will win because both X and Y prefer A to B and will vote consequently; Z, who prefers B, will be outvoted. If, instead, the electorate is presented with a choice between B and C, B will win because that is preferred by X and Z. Now, consider what happens if the voters are given the choice between A and C. C will win because voters Y and Z will vote for C. This means that the "social preferences" expressed by the vote are intransitive — A is preferred to B, B is preferred to C, but C is preferred to A.

This happens because the preferences of some voters (Voter Y in this case) are not "single-peaked." To see what this means, imagine a one-dimension continuum from A to B to C (like if they are placed in this order along an axis). Voter X's preferences are single-peaked because, from his most preferred alternative, A, he continuously loses utility as he moves further from A in either direction. In this case, he cannot move "left" from A because he is at the left extreme on the continuum; but the further he goes right, the less utility he gets. Similarly, Voter Z's preferences are single-peaked: his most preferred alternative is B, in the middle of the continuum, and he looses utility as he moves left to A or right to C. But Y's preferences are not single-peaked: he most prefers C and, as he moves left to B, his utility decreases, but then increases again at A. R

TABLE 1

An Illustrative Case of Voting

	First Preference	Second Preference	Third Preference
Voter X	A	B	C
Voter Y	C	A	B
Voter Z	B	C	A

day politics apply. This latter stage typically involves decisions based on majority approval, not unanimous agreement.

CYCLING Why would individuals agree to have collective choices made by majority rule? There is only one way an individual can be sure not to be exploited by a majority: have veto power over any collective choice or, in other words, require that all decisions be approved unanimously. However, unanimous agreement is practically impossible because the decision cost is prohibitively high — a larger number of individuals will have to be persuaded, and it will be in the interest of each one to lie on his preferences in order to manipulate the decision and get the highest benefits and pay the lowest taxes. On the other hand, the lower the required plurality, the higher the risk of an individual being exploited. Depending on the cost of reaching a decision and the lost benefits if the wrong decision is made, the proportion required may be a simple majority (50 percent plus one vote) or some higher, qualified majority.

Here, we meet the first problem uncovered by public choice analysis: majorities — especially simple majorities — are arbitrary. It is far from clear who the majority is and what it wants. Majorities can be inconsistent or, what amounts to the same, can cycle indecisively among alternatives. (See "Why the Majority Wants Both A and Not-A," above." The majority can vote for Proposal A over Proposal B, and B over Proposal C, but then vote for C if asked to choose between A and C. The majority seems intransitive, irrational. This explains why the electorate often appears so inconsistent — for example, when it votes for minimum wages that create unemployment, and then for government programs meant to create jobs.

Cycling can occur when the preferences of some voters are not "single-peaked" — that is, given a broad spectrum of options, the force of the voter's preference does not consistently decline as he considers options further and further from his top choice. There is nothing in individual rationality that implies single-peaked preferences. As Mueller notes in *Public Choice III*, "During the Vietnam War, it was often said that some people favored *either* an immediate pullout or a massive expansion of effort to achieve total victory." Another example: Some people think that either selling tobacco should be forbidden or else smokers should be left alone; middle-of-the-road options are less preferred (by some).

Another way to see the cycling-inconsistency phenomenon is to understand how redistributive coalitions are unstable. If the middle class votes with the poor, the two classes can expropriate the rich. But then it will be in the interest of the rich to bribe the middle class into a new coalition that advocates taking all possible money from the poor and offering a bit more to the middle class. The poor, realizing that they would be better off if they were less exploited, would then offer a new deal to the middle class to exploit the rich and share the loot differently. And so on to a new winning coalition. What, then, does the majority want?

Of course, "the middle class" does not vote like a single person and, in a simple majority system, any coalition comprising a simple majority of voters will do. Logrolling is the way coalitions will often be formed in practice. "Logrolling" describes an informal exchange of votes: Politician X supports a measure he does not much like but Politician Y finds very important in exchange for Y's support for a measure that X wants dearly. If such trade in votes is possible, somebody else can outbid one of the traders and get support for his own pet project. That new coalition can also be overturned. Mueller reports some empirical evidence that such vote trading happens in Congress — for example, lawmakers representing peanut farmers voted for initiatives favored by the sugar-farming industry in return for votes from sugar farmers' congressmen on peanut industry measures.

MEDIAN VOTER THEOREM When no cycling occurs, the median-voter theorem kicks in. In a simple one-dimensional case, the theorem states that, if all voters have single-peaked preferences, the winning alternative in a simple majority vote will be the ideal (or most favored) alternative that is at the median point of the preference distribution. This can be seen in Figure 1, which represents the preferences of voters X_1 to X_5. The horizontal axis represents the different alternatives of a one-dimensional issue — say, a smaller or larger tax rate, with the rate ranging from 0 percent to 100 percent. The strength of each voter's preference for a specific tax rate is shown by the respective inverted-U curve and measured on the vertical axis. (This is measured in an "ordinal" way, i.e., voters prefer alternatives "more" or "less" — consequently the height of each voter's utility curve does not matter.)

Consider Voter X_3. His ideal tax level is t_3, which corresponds to the top of his utility curve. The more the actual tax rate diverges from t_3, the less he likes the alternative, which is just another way to say that his preferences are single-peaked. The same is true for the other four voters. By definition, X_3's ideal alternative, t_3, is the median one among the five voters' ideal alternatives.

The median-voter theorem is now easy to understand. In pair-wise voting, there is no alternative that can win against t_3 (assuming all voters vote). Suppose, for instance, that the electorate is asked to choose between t_2 and t_3. Since t_3 is the median, the majority of voters (i.e., X_3 and the two voters to his right) prefer it to any lower rate. Remember that preferences in this example are single-peaked and that a voter prefers an alternative less the further it is from his ideal one. Note that this is true no matter how the other voters' ideal



MORGAN BALLARD

rates are distributed across the continuum — the median rate will always win.

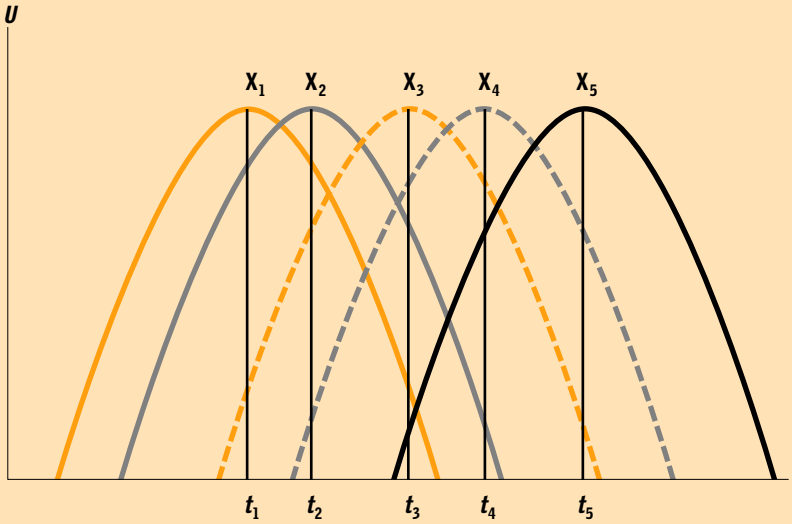
This explains why, especially in two-party systems, political platforms are so similar. If, in the upcoming U.S. presidential election, President Bush promises t_2 and Sen. Kerry promises t_5, Voter X_3 will vote for Bush because Voter X_3 gets more utility from t_2 than from t_5. Hence, t_2 will win in that pairing. If Kerry is self-interested and wants to get elected, he will move to t_4, thereby countering Bush. Bush, if he is self-interested and wants to get elected, will then move even closer to t_3, and so will Kerry. Both politicians will move toward the median point in pursuit of an electoral win.

Except for the median voter, all voters are unhappy with the results. This feature of voting (whether in elections or referenda) is inseparable from collective choices. If a collective choice decided the type of car that everyone would drive, the outcome might be the production of a Ford Taurus for everybody. The median voter would be happy (assuming the Taurus is the preferred car of the median voter), but everybody else would rather be driving another car.

FIGURE 1

Why the Centrist Wins

An illustration of the Median-Voter Theorem.



CYCLES OR TYRANNY? Cycles grow more likely as the number of possible alternatives increases and the individual preferences become more heterogeneous. Qualified majorities can reduce the probability of cycling. An "agenda setter" who decides which alternatives are put on the ballot can also reduce cycling, but he likely will set the agenda to produce an outcome that follows his own preferences. If X is the agenda setter, he will first pair Proposal B with Proposal C and send that choice to the voters. He will then pair the winner with Proposal A, which Voter X prefers. The result of that agenda is that A will win. Likewise, if the agenda setter is Y, he will make sure that the first vote pits Proposal A against Proposal B, and will then run the winner against C so that C wins.

If there is no cycling, another danger looms: a stable majority that would exploit and oppress an identifiable minority. We seem to be facing an uncomfortable choice: either inconsistent majorities with unstable coalitions, or agenda-setter dictatorship, or else a Tocquevillian "tyranny of the majority."

DIFFERENT VOTING RULES Besides the simple majority rule analyzed above, there is a large number of different voting rules meant to find out who exactly is the majority and what it wants, especially when more than two candidates are running. The plurality rule just picks the candidate who gets the highest percentage of the vote. The majority rule with runoff election pairs the two candidates who obtained the highest percentage of the vote in a first round. The Condorcet criterion calls for the winner to be the one who can defeat all the other candidates pairwise. The Hare system and the Coombs system eliminate the least-preferred candidates over many rounds. In the approval voting system, each voter votes for many candidates and the most popular, according to this criterion, wins. The Borda count asks voters to score the candidates, and the winner is the candidate with the highest aggregated score.

All of those systems have different advantages and drawbacks, and often give different results. "For a given set of voters with unchanged preferences," Gordon Tullock concludes, "any outcome can be obtained by at least one voting method." This is not without consequences in the real world: Richard A. Joslyn's research suggests that, in the 1972 Democratic primaries, Edmund Muskie would have won against George McGovern under voting procedures other than the plurality rule.

When the issue to be decided relates to the production and financing of public goods, each individual is incited to understate what he wants, hoping to get the benefits while paying lower taxes. "Preference-revealing" systems have been devised theoretically with the aim of bringing voters to reveal truly what each alternative is worth to them. Mueller is optimistic that this line of research will eventually lead to better collective choice mechanisms. But he also shows how voting with one's feet — exit as opposed to voice — is often a better solution: In a decentralized political system, individuals get the combination of public goods and taxes they want by moving to the jurisdictions that offer their preferred packages. They reveal their true preferences by choosing where to live.

WHY VOTE? In the 2000 presidential election, the voter turnout (proportion of registered voters who did cast a ballot) was 67.5 percent. For public choice theory, the problem is not that one-third of the voters did not vote, but that two-thirds did. Why did they bother? Granted, for the typical individual voter, the cost of voting is low and mainly involves the time spent going to the polls. But the expected benefit (the value of the benefits

Does Your Vote Really Count?

Although the question of the decisiveness of a vote rapidly gets technical, the general ideas can be grasped easily.

Suppose that you are a member of a committee of three persons, X, Y, and Z, including yourself (Y). A simple majority vote is to be taken on Proposal A vs. Proposal B. Assume that there is a 50–50 chance that any of the other two members will vote for A, and likewise for B. There is no abstention or third alternative. What is the probability that your vote will count, in the sense of being decisive? This amounts to asking the probability that there would be a tie if you did not vote.

Because Voters X and Z can each vote for either A or B, there are four possible outcomes of the vote without your voice: A-A (i.e., Voters X and Z both vote for Proposal A), B-B, A-B, and B-A. Only in the last two (A-B and B-A) is there a tie without you; in the other cases, it makes no difference whether you vote for A, for B, or do not vote at all. Therefore, there are two chances in four that your vote will count.

Those who have done some elementary combinatorial analysis and probability theory will recognize that if p is the probability that one of the other voters chooses for A, and $(1\text{-}p)$ the probability that he votes for B instead, then the probability P that there will be a tie is given by the formula

$$P = {}_nC_{n/2} \times p^{n/2} \times (1-p)^{n/2}$$

where ${}_nC_{n/2}$ is the number of combinations of all n voters taken $n/2$ at a time. In fact, with a large number of voters (1,000 or more in the table), an approximation to the above formula has to be used to calculate P, even with computers.

By counting the possible results, calculating their equal probabilities, and adding the probabilities for the tied results, or alternatively by using the above formula, it can be calculated that on a five-person committee (including you), the probability that your vote will count is ⅜ or 0.375.

The larger the number of voters, the lower the probability that your vote will count. With 1,000 voters, the probability is only about 0.02 (or 1:50). With 100 million (the approximate number of voters in a presidential election), it is reduced to 0.00006 (about 1:17,000).

The probability that your vote will be decisive diminishes dramatically as the probability that any voter chooses one alternative or the other diverges from 50–50. This is easy to see in our example of a three-member committee. Assume now that each of the other two members votes for A with probability 0.8 (or, alternatively, votes for B only once in five times). Now, the two results that make your voting worthwhile (AB and BA) only occur 4/25 of the time each, so that the occurrence of one or the other event has a probability of 8/25, or 0.32.

Consider again the upcoming presidential election. Assume that about 100 million other electors will vote, and that the probability that one of those voters taken randomly will vote for Kerry is 0.49 and the probability he will vote for Bush is 0.51. Disregard the complication of the electoral college. The probability that there will be a tie, and that your vote will make a difference, can be calculated: it is eight chances in 10^{8691}. The number 10^{8691} (10 followed by 8,690 zeros) is a more-than-astronomical number. It is much, much larger than the total number of elementary particles in our universe (10^{100}) and the age of the universe in seconds (3×10^{17}). Consequently, the probability that your vote will count is extremely close to zero.

There have been some recent challenges (mentioned by Mueller in *Public Choice III*) to this methodology, and some higher probabilities have been obtained. But, for all practical purposes, one vote — your vote — still has no significant probability of changing the election outcome. R

TABLE 2

Probability that Your Vote Will Count

Number of electors (n) without you	Probability that an elector votes for one of two alternatives (p)	
	$p = 0.5$	$p = 0.49$
2	0.5	0.4998
4	0.375	0.3747
1,000	0.0189	0.0155
10,000,000	0.0002	$5 / 10^{873}$
100,000,000	0.00006	$8 / 10^{8691}$

Note: Probabilities for n = 2 and n = 4 are exact; for other values of n, they are approximated by Owen and Grofman's formula as modified by Mueller (cf. Public Choice III, p. 304–305).

promised by the voter's preferred candidate multiplied by the probability that the voter's action will get the candidate elected) is infinitesimally small. Consequently, the rational voter should abstain. Why so many voters go to the polls is called "the voting paradox."

In his new book, Mueller reviews many solutions that have been proposed to the voting paradox. The common idea is that self-interest has to be conceived in a less narrow fashion: the voter must consider other benefits than the expected gains from his preferred candidate's policies. Those other benefits might be the pleasure of expressing an opinion or of being part of the crowd (like when one cheers or boos at a hockey game), the satisfaction of doing one's moral duty, or the desire to be seen as a "good citizen." Indeed, there is some evidence that people often vote against their own narrow interests. In brief, the benefits individuals get from voting are related to their

expressing their "public preferences."

The voter weighs those benefits, which are independent of his being on the winning or losing side, against the cost of voting. The higher the cost of voting (other things being equal), the lower the turnout. Mueller reports research showing that, before poll taxes were made illegal in Southern states, "a six-dollar poll tax in 1960 reduced the probability of an individual voting by 42 percent." Voters are rational — they vote for other reasons than expected benefits from their preferred candidates or measures, but they weight those reasons against the cost of voting.

This conclusion has massive consequences. One is that, because a voter's expression has little or no impact on the election results, he has no incentive to be well-informed on political issues. Because voters thus remain "rationally ignorant," politicians can, to some extent, cheat on their commitments and indulge in their own personal ideological preferences.

DISTORTING INTERMEDIARIES

In our democracies, voters do not decide most issues directly. In some instances, they vote for representatives who reach decisions in parliamentary assemblies or committees. In other instances, they elect representatives who hire bureaucrats to make decisions. The complexity of the system and the incentives of its actors do not necessarily make collective choices more representative of the citizens' preferences.

POLITICIANS Public choice theory assumes that politicians want to win elections — otherwise they will not be politicians for long. To achieve their goal, the politicians propose measures that they think the majority prefers, and they join political parties. "Parties formulate policies in order to win elections," wrote Anthony Downs in *An Economic Theory of Democracy,* "rather than win elections in order to formulate policies."

The rule used for U.S. congressional elections and parliamentary elections in most British-tradition countries is plurality-based. This encourages the dominance of two political parties because third parties can get few candidates elected even if they can get a good proportion of the vote across all districts. In two-party political systems, as long as issues are one-dimensional and preferences are single-peaked, political parties will try to get as close as possible to the median voter. In multidimensional cases or with multiple-peaked preferences, cycling can occur.

In political systems with proportional representation, more than two political parties will usually thrive and will form coalitions to govern. Different systems of proportional representation exist, but the general idea is that some adjustment procedure brings the proportion of representatives closer to the proportion of the popular vote received by different parties in the whole country. Proportional representation gives better representation to voters with minority preferences or ideologies, but it also leads to more unstable governing coalitions.

BUREAUCRATS When public choice analysis is applied to the bureaucracy, it uncovers other reasons to doubt that the state can efficiently reconcile the individuals' preferences and aggregate their demands for public policies. Once again, public choice assumes that the bureaucrat is an ordinary individual who, like everybody else, tries to maximize his utility.

What does this mean in practice? What does the bureaucrat do to maximize his utility, given the constraints placed upon him? Most attempts to answer that question are adaptations of the systematic model William Niskanen developed in his 1971 book *Bureaucracy and Representative Government.* Bureaucrats are assumed to maximize the size of their bureaus' budgets because they can thereby increase their real remuneration in terms of perks (larger offices, better expense accounts, etc.), lower risk of missing their objectives, recognition, etc. Thus, the bureaucrats will produce more than the politicians (and, presumably, the citizens) want, or at a higher cost.

Bureaucrats can do this because their political sponsors (the politicians or other political masters who determine the agencies' budgets) do not know the real cost of producing what they order from the bureaucracy. Of course, the sponsors will try to control bureaucratic activities, but the monopolistic (single-provider) advantage of the bureaus will defeat at least some of those efforts. This theory is supported by a large number of studies showing that production costs in public bureaucracies are higher than in comparable private enterprises.

Another way the bureaucrat exerts power is by being an agenda setter. As we have seen, the agenda setter can often lead the system toward the results he prefers by deciding which alternatives, in what order, will be voted upon by the politicians or the citizens.

INTEREST GROUPS In order to influence collective choices, citizens have to engage in collective actions: participate in demonstrations, organize lobbying activities, contribute to political parties, etc. The result of a group's collective action (say, tariff protection) is often a public good for the members of the group — each will benefit from it whether or not he has contributed. Moreover, the action of a single individual may not count much in the final success of the collective action. Thus, in participating in a collective action, every individual incurs costs for virtually no benefit and will therefore be tempted to free ride on the efforts of others. That is especially true in large groups where one individual's actions have less impact and shirkers can more easily avoid sanctions like boycott. Thus, small groups with concentrated interests like farmers or steel producers will be better organized and wage more efficient collective action than large groups with diffuse interests like taxpayers or steel consumers. And well-organized groups will exploit less organized ones.

That was a major conclusion of Mancur Olson's highly influential book *The Logic of Collective Action.* For example, Swiss farmers, who are a small part of their country's population, get effective subsidies equivalent to 86 percent of their incomes while farmers in Ghana, who are a large part of their country's population, get an effective negative subsidy amounting to 27 percent of their incomes as they subsidize the small urban class.

Interest groups will engage in what public choice theorists call "rent seeking," i.e., the search for redistributive benefits at the expense of others. The larger the state and the more

benefits it can confer, the more rent-seeking will occur. "The entire federal budget," writes Mueller, "can be viewed as a gigantic rent up for grabs for those who can exert the most political muscle." Rent seeking does not produce a pure transfer; when individuals or groups compete for some advantage from the state (e.g., a subsidy, a monopoly), they will all use real resources (e.g., ink, paper, travel, meals, time) in trying to grab it. As a result, part of the expected rent will be dissipated, creating a net social loss.

GROWTH AND JUSTIFICATIONS OF THE STATE

From about 1870 until around 1913, total government (federal plus state) expenditures in the United States was around 7 percent of GDP. It grew to 12 percent by 1920, and to 20 percent by 1937. By 1980, it had reached 31 percent and has slightly increased since then. The same evolution happened in the public sectors of other countries, though many of them have seen even more growth post-1980, and some now are as high as, or higher than, 50 percent of GDP. Why did the state grow so much?

TWO MODELS Public choice theory oscillates between two models of the state: the citizen-demand model and the Leviathan model. In the citizen-demand model, Mueller explains, "policies are reflections of the preferences of individual voters." In the Leviathan model, he says, "it is the preferences of the state, or of the individuals in the government, that are decisive." Depending on which model is used, analysis of why the state has grown yields different results.

The citizen-demand model claims that the citizens have demanded more public goods, more control of negative externalities (like pollution), and more redistribution of income. One interesting hypothesis, supported by some research, is that the expansion of the voting franchise to women and the poor has fueled the growth of the state. But is this really what the citizens want? And which part of the citizenry?

The Leviathan model stresses factors on the supply side: the state grows because its rulers or beneficiaries want more loot. Bureaucrats comprise a large part of the electorate — 15 percent on average in OECD countries, and 20 percent or more in some countries — and research shows that their voter turnout is higher than other citizens'. Another explanation is that politicians can easily fool rationally ignorant voters.

Mueller, in *Public Choice III,* cites evidence that direct democracy and federalism provide an effective check on the growth of Leviathan. This suggests that the citizen-demand model is more applicable in such cases.

SOCIAL CHOICE THEORIES What are the standards for judging state action? What should the state do, and how? Those questions are the subject matter of normative theories of public choice, which include "constitutional economics" as well as "social choice" theories outside the narrow public choice school.

In social choice theories, we meet the sort of "social welfare functions" that preceded the public choice revolution. A social welfare function is a ranking, in terms of social welfare, of all Pareto-efficient configurations of prices, wages, and income distributions. To choose between them, "society" — which in practice means the state — must weight the utility of different individuals. Before the development of the public choice school of thought, it was already known that social welfare functions require that utility be measurable in a cardinal sense and be comparable between individuals. In other words, somebody must calculate that, say, a 10 percent increase in X's utility is worth "to society" more than a 15 percent drop in Y's utility. There is no scientific basis for such moral judgments or for the social welfare functions based on them.

> The entire federal budget can be viewed as a gigantic rent up for grabs for those who can exert the most political muscle.

In his 1951 book *Social Choice and Individual Values,* Kenneth Arrow attacked the problem with a different methodology. He asked if a social welfare function could be built from a few simple axioms, avoiding any interpersonal comparison of utility. In trying to answer the question, he demonstrated that no social welfare function exists that will, at the same time, respect Pareto optimality, be nondictatorial, and be transitive (i.e., not lead to cycles). In a sense, Arrow generalized the conclusions on cycling and showed that social preferences must be either inconsistent or dictatorial.

TAMING LEVIATHAN

There is an ambivalence in the corpus of theories and empirical evidence called "public choice" that Mueller encapsulates well in *Public Choice III:*

> Some scholars like Brennan, Buchanan, Niskanen, and Usher look at the state and see a grasping beast set upon exploiting its power over citizens to the maximum degree. Others, like Breton and Wittman, when they gaze upon the state, see an institutional equivalent to the market in which democratic competition produces

efficiency levels comparable to those achieved by market competition.

Public choice can be seen as demonstrating either the usefulness of politics or the existence of state failures; as arguing either that the state promotes allocative efficiency or that it is a redistributive machine; as proposing either a demand model of the state where political competitors respond to citizens' demands or a supply model where Leviathan rules. Expressed differently, there is a permanent tension between the libertarian and the interventionist strands in public choice theory.

One thing is sure: public choice has destroyed the naïve view that, in order to justify state intervention, it suffices to show that there exist market failures that an ideal state could correct. After the public choice revolution, political analysts cannot be satisfied with comparing real markets with an ideal state; they must analyze the state as it is before dreaming about what it should be. Public choice has demystified and un-deified the state.

GOING FURTHER Dennis Mueller's *Public Choice III* provides a masterly overview of the public choice literature over the last half century. It will remain as the *Summa Theologica* of public choice economics at the beginning of this century.

In my view, we can go further than Mueller, who perhaps tends to be overly optimistic about politics. The more extended is the state's domain, the more likely its majorities will be either inconsistent or oppressive. The paradigmatic case is perhaps the cycling phenomenon: the more the state intervenes, the more the issues will become multidimensional, the less homogeneous will the voters' preferences be, and the more inconsistent or dictatorial Leviathan must become.

We must question the orthodox approach of weighting costs and benefits in search of optimality. Optimality can be defined in such a broad way that everything is optimal, in the sense that it cannot be otherwise given the present state of the world. Is this not what some economists do when they argue that political competition between interest groups produces optimal results because individuals will get organized when the cost of oppression gets too high? Do we not have a tautological notion of optimality?

Thoughtful defenders of the interventionist strand do have an external criterion of optimality that lies in some sort of social welfare function. What is optimal is what maximizes some function of individual utilities, perhaps represented by monetary costs and benefits. On this basis, it can theoretically be shown that some (competitive) institutions lead to Pareto optimality and the state can use cost-benefit analysis to get us there. But, as public choice analysis has shown, this approach presupposes either interpersonal comparisons of utility or "social preferences" that are inconsistent or dictatorial.

We should take seriously the challenge raised by Anthony de Jasay, himself an (unorthodox) public choice theorist. In his book *The State,* he reminds economists that interpersonal comparisons of utility "are merely a roundabout route all the way back to irreducible arbitrariness, to be exercised by authority." "At the end of the day," he continues, "it is the intuition of the person making the comparison which decides, or there is no comparison.… In an analogous manner, the two statements 'the state found that increasing group *P*'s utility and decreasing that of group *R* would result in a net increase of utility' and 'the state chose to favor group *P* over group *R*' *are descriptions of the same reality.*" Once that is realized, public choice becomes essentially an indictment of the state.

Finally, taking public choice theory seriously implies a serious questioning of the "we" or "they" as political collectives. Statements such as "We, as a society, think or do such and such" and "The French (or the Americans) believe or do this or that" are either rhetorical and logically meaningless or else dictatorial. There is no nondictatorial way to aggregate different individual preferences and fuse them into one set of super-preferences, except if the individuals have identical preferences or if they are unanimous. Examples of similar preferences are found in small and close groups like a couple, a family, or a few friends —although even in those cases, one individual is often the "dictator" or the leader. Examples of unanimous choice include shareholders buying into a company and members joining an association. Unanimity is the only way out of the dilemma between meaningless and dictatorial collectives.

Thus, except in an abstract constitutional perspective (agreement on very basic rules), the political "we" implies that some individuals impose their preferences on others. In this sense, the public choice revolution rings the death knell of the political "we." R

READINGS

- "Dictatorship, Democracy, and Development," by Mancur Olson. *American Political Science Review,* Vol. 87, No. 33 (1993).
- *Government Failure: A Primer in Public Choice,* edited by Gordon Tullock, Arthur Seldon, and Gordon L. Brady. Washington, D.C.: Cato Institute, 2002.
- *The Limits of Liberty: Between Anarchy and Leviathan,* by James M. Buchanan. Chicago, Ill.: University of Chicago Press, 1977.
- *The Logic of Collective Action,* by Mancur Olson. Cambridge, UK: Cambridge University Press, 1965.
- *The Power to Tax: Analytical Foundations of a Fiscal Constitution,* by James M. Buchanan and Geoffrey Brennan. Cambridge, UK: Cambridge University Press, 1980.
- *Public Choice III,* by Dennis C. Mueller. Cambridge, UK: Cambridge University Press, 2003.
- *The State,* by Anthony de Jasay. Oxford, UK: Blackwell Publishing, 1985.

Taxation by regulation

Richard A. Posner

Professor of Law
The University of Chicago

Students of the regulated industries often assume that regulation is designed either to approximate the results of competition or to protect the regulated firms from competition. But neither view explains adequately a number of important phenomena of regulation and regulated industries. Foremost among them is the prevalence of "internal subsidies," whereby unremunerative services are provided, sometimes indefinitely, out of the profits from other services. To understand this and other phenomena, we must assign another important purpose to regulation: we can call it "taxation by regulation." The purpose of this paper is to explore this dimension of the regulatory process, to demonstrate that it explains some otherwise perplexing features of the process and the industries subject to it, and to compare it with other methods of public finance.

■ Two views of the purpose in fact of public utility and common carrier regulation[1] vie in current scholarly debate. One, the more familiar, holds that regulation is a device for protecting the public against the adverse effects of monopoly[2]; the other holds that regulation is procured by politically effective groups, assumed to be composed of the members of the regulated industry itself, for their own protection.[3] In my opinion neither view, at least as thus far formulated, explains an important phenomenon of regulated industries: the deliberate and continued provision of many services at lower rates and in larger quantities than would be offered in an unregulated competitive market or, *a fortiori*, an unregulated monopolistic one.

Richard A. Posner, who received the A.B. degree from Yale University and the LL.B. from Harvard, is Professor of Law at the University of Chicago Law School. Among his major interests is the application of economic theory to law, and he also is the editor of a new journal to be published under the auspices of the University of Chicago. It is entitled *The Journal of Legal Studies*, and its first appearance will be in January 1972. The author would like to thank Edmund W. Kitch and George J. Stigler for their helpful comments on an earlier draft.

[1] By the awkward term "purpose in fact" I mean to distinguish sharply between the reasons ascribed to regulatory laws and decisions by the legislators and regulators themselves and the reasons, whether or not anywhere avowed, that provide a consistent explanation of the actual course and consequences of regulation. It is with the latter that I am concerned here.

Although the emphasis throughout this paper is on the full panoply of rate and entry controls characteristic of public utility and common carrier statutes, where the analysis can be extended to markets having somewhat different regulatory arrangements I have not hesitated to do so.

[2] For example, see Bonbright [3], p. 23.

[3] A rigorous statement of this theory appears in the article by G. J. Stigler in this issue of *The Bell Journal* [59]. The "capture" of regulation by the regulatees is, of course, an old theme in the literature of regulation. Professor Stigler's theory allows for capture by effective political groups other than the regulated firms themselves, and there is accordingly no necessary inconsistency between it and the analysis in this paper.

This phenomenon can be explained, I believe, only if we modify existing views by admitting that one of the functions of regulation is to perform distributive and allocative chores usually associated with the taxing or financial branch of government. And we shall see that an analysis of taxation by regulation explains other perplexing phenomena. But it would be error to think that the analysis compels rejection, as distinct from modification, of the existing views of regulation. I hope to show that any theory that conceives the function of regulation to be to approximate the results of competition, or to enrich the regulated firms, or to do sometimes the one and sometimes the other, is incomplete. But it does not follow that a broadened public-interest approach (one that accommodated certain subsidy elements) or a broadened effective-political-group approach (one that viewed certain customer classes as effective political groups) might not be tenable.

1. Internal subsidies

■ The best known example of a regulated service that an unregulated competitive market would not provide on the same scale is railroad passenger service,[4] but it is far from being an isolated example.[5] Like railroad passenger service, domestic telegraph service would be declining even more rapidly than it is, were it not for the Federal Communications Commission's stubborn rearguard action against further rate increases and service degradation by Western Union.[6] Local airline service provides another example,[7] complicated though by the presence of direct as well as internal subsidies.

The phenomenon is by no means limited to declining industries. The American Telephone and Telegraph Company may soon be required to provide electronic interconnection free of charge (or nearly so) to the National Educational Television network.[8] Broadcasters are required to provide at least some nonremunerative news and public-affairs programming.[9] Liability insurance for high-risk automobile drivers in many states is written at a loss.[10] Producers of natural gas are constrained to sell at a price that does not include scarcity rents, thereby benefiting present consumers of natural gas at the expense of future consumers, who may encounter shortages.[11] Uniform rates, based on averaging together the costs of services whose cost characteristics are in fact very different, are a conspicuous feature of regulated rate structures. Examples are statewide telephone rates, the uniform long-distance rate for telephone calls of the same distance and duration regardless of route, the pegging of airline and rail passenger rates to distance with little or no consideration for cost differences over different routes, and the flat rates characteristic of

[4] See Conant [10], p. 132, Hilton [23], p. 136, and Nelson [40], pp. 286–301.

[5] In at least one respect, we shall see, it is atypical of the general class. See p. 43 and fn 69 *infra*.

[6] See President's Task Force on Communications Policy [50], and U. S. Federal Communications Commission [75].

[7] See Eads [14], pp. 1, 13–14.

[8] See The Communications Act of 1934 [78], sec. 396(h), Interconnection Service [73].

[9] See Public Service Responsibility of Broadcast Licensees [74].

[10] See Keeton and O'Connell [30], pp. 93–94; cf [67], Price Variability in the Automobile Insurance Market.

[11] See MacAvoy [35], p. 271.

urban transit systems.[12] Electrical companies grant discounts, unrelated to any cost savings, to hospitals and other worthy groups,[13] while water companies are often required to furnish water for household use at rates below marginal cost—and to fire departments and schools for nothing.[14]

These practices are very old. In 1827, the State of Illinois required by statute that every keeper of a ferry, toll bridge, or turnpike road "give passage to all public messengers and expresses; to all grand and petit jurors, when going to and returning from court, without any fee or reward whatever" [26].

In all of these examples, as in many others one could cite (including some to be discussed later), a service is provided that does not pay its way in the market. Someone must pay its way, however. Normally it is other customers of the firm rendering the service, who pay a price higher than the cost of serving them. Sometimes it is future generations of consumers (as in the gas example) or the stockholders or creditors of the firm (as, perhaps, in the railroad and telegraph examples). Examples of internal or cross-subsidization, as we may call the practice, lie everywhere at hand in the regulated industries.[15] They are also commonly found among public enterprises here and abroad,[16] the structure of postal rates providing a conspicuous example.[17] They seem much less common in unregulated private markets. The contrast is instructive: it suggests that the provision of internal subsidies is associated with distinctively governmental purposes and functions. We shall return to this point.

Before proceeding, I should caution the reader that identification of our phenomenon—internal subsidization—is often difficult.[18] In practice an internal subsidy is not always easy to distinguish from certain familiar profit-maximizing practices. A monopolist may be able to maximize profit by setting different markups on different sales, depending on the elasticities of demand of particular customers or groups of customers. Like the internal subsidy, this practice involves a departure from cost-based pricing. But it does not involve any unremunerative services. Were the lower-priced services unprofitable, the monopolist could increase his profits still further by terminating them. To take another example, a firm that has a monopoly in some markets can, under certain restricted conditions, increase its profits by selling below cost in other markets where it still faces competition. Unlike internal subsidization, however, such "predatory price discrimination" is strictly a temporary tactic to

---

[12] See Caves [5], p. 369; Davidson [12], p. 217; Johnson [27]; Garfield and Lovejoy [18], pp. 200–203, 243; Meyer, Kain and Wohl [37], pp. 354, 357; Meyer Peck, Stenason and Zwick [38], pp. 166–67; Watkins [82], p. 623.

[13] See Bonbright [3] p. 111.

[14] Hirshleifer, DeHaven & Milliman [24], pp. 109–11; Garfield and Lovejoy [18], p. 225; cf. Keig, Fristoe and Goddard [31].

[15] For additional examples see Bonbright [3], pp. 111–12; Friedlaender [15], pp. 66–68; Meyer, Peck, Stenason and Zwick [38], pp. 194–95; Nelson [40], p. 331; Hearings on Transportation Acts Amendments [66].

[16] See, for example, Coase [8], pp. 1, 13–14; Crew [11], p. 258; Peltzman [44]; Sargent [51], p. 248; Sharp [52], p. 53; Shepherd [53], p. 132. Sargent and Shepherd analyze internal subsidization in the context of public enterprises in terms somewhat parallel to mine in this paper.

[17] See Baratz [2], p. 305; Coase [9], p. 25; Kennedy [29], pp. 93–94.

[18] As emphasized by Nove [42], p. 847.

drive out competitors; prolonging it indefinitely would make no business sense.

Peak-load pricing also involves different prices for the same service, and is also different. One observes resort hotels charging, for identical accommodations, higher prices in peak seasons than in off-seasons. But the reason for the off-season discount is that it attracts customers who would otherwise not buy the service. So long as the price they pay covers the marginal or additional cost of serving them, the discount is profitable. It is therefore not a case of internal subsidization as I use the term.

Temporary below-cost selling might of course stem from errors of various kinds. And even persistent departures from a proper matching of price and cost could reflect simply the difficulty and expense of determining which costs were incurred in making which sales; some averaging of costs over different services is certainly explicable on this ground. Other cases of apparent below-cost selling evaporate on examination. Professor Bonbright gives as an instance of internal subsidization (he calls it "social principles of rate making," but it is the same concept) the granting of special rail fares to clergymen[19]; but one suspects that such a discount is simply a public-relations gesture. In view of the low income of clergymen, which may make their demand for some services relatively elastic, the discount might even be an example of price discrimination.

Finally, some services may appear unremunerative only because an inappropriate conception of cost is employed. As the resort example shows, a price below average cost, but equal to or above marginal cost, may still be remunerative. On the other hand, a price above even long-term marginal cost may be unremunerative.

The last point requires some explanation. In the case of an industry in which average cost decreases with output, a firm that charged a uniform price equal to its marginal cost would not recover its total costs. It could recover them by setting a uniform price equal to average cost. This would force customers willing to pay a price equal to or slightly above marginal cost but not the higher price equal to average cost to turn to more costly substitutes. Neither result is optimal, and the proper solution to the dilemma is a matter of fair debate. One attractive possibility is to charge a price equal to marginal cost for marginal purchases and a sufficiently higher price for inframarginal purchases to cover total costs without losing those sales. Although the proper design of the rate structure is not easy, this approach seems preferable to either the uniform marginal-cost price, which necessitates a government subsidy to make up the deficit in covering total costs, or the uniform average-cost price, which excludes the marginal sale.[20]

It is implicit in this solution, however, that some customers may end up paying a higher price than under average-cost pricing. A simple arithmetical example will illustrate. Suppose the fixed costs of providing a service are $1000 and marginal cost is $1, and suppose further that when price is made equal to average cost 500 units are demanded and supplied at a price of $3 per unit. Therefore, Buyer

---

[19] See Bonbright [3], p. 61.

[20] See Coase [7]; Henderson [22], p. 223; Hirshleifer, DeHaven and Milliman [24], pp. 91–92.

X, who purchases 20 units, pays a total price of $60. Now suppose the industry adopts a two-part pricing system (one version of the preferred solution to the dilemma of proper pricing under conditions of decreasing average cost) under which anyone wishing to purchase service must pay (a) a share of the fixed costs (ideally, the shares should vary inversely with the elasticity of the buyers' demands) plus (b) the marginal cost of each unit he buys. And suppose that *X*'s share of the fixed costs is set at $100. He must now pay $120 for the same 20 units.

To the extent that regulatory agencies are unwilling to visit such consequences upon particular customers of the firms they regulate, output is not carried to the efficient point. The subsidy is indirect but inescapable: the additional price that the rejected marginal customers must pay for substitute products is a cost imposed on them in order to enable inframarginal customers to buy at a cheaper price than if an efficient pricing system were employed. Profit-maximizing price discrimination would also result in the inframarginal customer paying a higher price. The pricing actually employed in decreasing-cost industries appears often to be neither efficient nor profitable.

"Value-of-service" pricing in the railroad industry provides good examples of this point. Before the development of truck transportation, the practice of proportioning rail rates to the price of the commodity transported may have been a roughly adequate method of concentrating the fixed costs of railroad service on those customers whose demands for rail transportation were least elastic. But once there was a good substitute service, the method ceased to have a rational basis, at least under the usual views of regulation, since the price of the commodity shipped bears no necessary relation to the adequacy of trucking as a substitute for transportation by rail. Wristwatches are more expensive per unit of weight than grain, but it does not follow that the demand of the wristwatch shipper for rail service is less elastic; it is probably more elastic, because trucking is a more feasible alternative for the shipper of wristwatches than for the shipper of grain.

Value-of-service pricing may have persisted because it is a convenient method of subsidizing some shippers regardless of the elasticity of demands for rail transportation. The favorable rates at which agricultural commodities continue to be transported seem a case in point.[21] Considering the broad range of subsidies that farmers have managed to obtain for themselves, it is perhaps not surprising that they have obtained internal subsidies as well. And they are not the only group, or railroading the only industry, where we find value-of-service pricing used as a method of internal subsidization.[22] Notice that keeping rates low to customers so favored is not only inefficient but can destroy a decreasing-cost industry.[23] These consequences only underscore the limitations of existing theories of regulation.

A final problem of characterization, one that also arises from the phenomenon of decreasing average costs in some regulated industries, requires brief mention. A pricing system for a decreasing-cost in-

[21] See Friedlaender [15], pp. 18–20; The Hoch-Smith Resolution [80], p. 801; cf Meyer, Peck, Stenason and Zwick [38] pp. 175, 187–88.

[22] See, for example, Johnson [27], pp. 42, 67; Wilson [83], p. 337.

[23] Stigler [60], pp. 213–14.

dustry, to be efficient, must satisfy three criteria: (1) it must enable the total costs of the enterprise to be recovered; (2) it must be so designed that no customer willing to pay at least the marginal cost of serving him is turned away; (3) there should be no sales below marginal cost. Now there are many different formulae for allocating overhead costs consistently with these criteria and the choice of one inescapably affects the distribution of wealth as between various customers. For example, suppose the fixed costs of an enterprise are $100; there are 10 customers who can be made to contribute to those costs; and each of them would pay $20 rather than do without service. The regulatory agency can insist on a pricing scheme under which each pays $10, or on one under which some pay $20 and some less (perhaps zero). Whichever choice is made makes some customers worse off and some better off than they would be under an alternative arrangement.

Because this type of "subsidization" does not involve the maintenance of any uneconomic service, and thereby avoids most of the problems with which this paper is concerned, I exclude it (perhaps arbitrarily) from the category of internal subsidies. That regulation unavoidably involves issues of wealth distribution as between customer classes does, however, emphasize the relevance of notions of taxation and public finance to the theory of public regulation—but on this, more shortly.

2. Internal subsidization and the received views of regulation

■ The existence of the internal subsidy is an embarrassment to proponents of the first view of regulation (at least as usually formulated) mentioned at the beginning of this paper—the view that regulation is imposed in order to bring about results approximating those of competition. As we have seen, the internal subsidy brings about results unthinkable in a competitive market, which is perhaps why a distinguished proponent of the first view, Professor Bonbright, who believes that "public utility services are designed to be sold at cost, or at cost plus a fair profit"[24], considers the internal subsidy aberrational.[25]

Nor is it explicable if regulation is conceived purely as a service demanded by and supplied to the regulated firms themselves, although here the demonstration is more complicated. While an unregulated firm, whether monopolist or competitive, would not profit from engaging in internal subsidization, regulation alters normal business incentives. Under certain conditions, a regulated firm can increase its profits by expanding the size of its plant even if some of the output of the enlarged plant must be sold at a loss.[26] In addition, a regulated firm might provide some unremunerative services as an unintended byproduct of a policy of deliberately not keeping detailed cost information. The purpose of such a policy would be to prevent the regulatory agency, which is normally dependent on the regulated firm for information about the firm and its markets, from acquiring sufficiently detailed knowledge of the firm's operations to enable more stringent regulation. Such a strategy is less risky than it may seem. A regulated monopolist can get by with less information about

[24] See Bonbright [3], p. 23.
[25] *Id.*, Ch. 8.
[26] Averch and Johnson [1], p. 1052.

the costs of particular sales than a nonregulated firm. It may be able to earn the overall profit permitted by the agency, even though many of its sales are unprofitable, by charging supracompetitive prices in markets where its monopoly is secure.

Furthermore, a firm that engages in internal subsidization can argue forcefully to the regulatory agency that the agency should not permit, or at least should strictly limit, the entry of competitors into those markets where the firm makes large profits, because those profits—which new entrants would erode—are necessary in order to cover the losses in the subsidized markets. Finally, regulated monopolists may believe that differential rates, even where cost justified, strike the public as discriminatory and may in consequence invite sterner regulation of the firm's profits.

These considerations do not explain, however, those instances, such as railroad passenger service and natural gas production, where the regulated industry does not appear to be benefiting, directly or indirectly, from internal subsidization. Nor do they explain very convincingly why the practice is ever instituted in the first place. If regulated firms dominate an agency, as the view under examination posits, they can use their position to increase profits directly by getting the agency to fix a high permitted level of profits and forbid the entry of new competitors. A program of internal subsidies seems a needlessly roundabout method for achieving these ends, and also a costly one: it entails smaller profits for the regulated firms since, as mentioned earlier, a profit-maximizing firm is always better off terminating a losing service. If firms do not yet dominate a regulatory agency but seek to do so, a program of internal subsidies is not likely to help them to reach their goal. A regulatory agency that is not a creature of the firms being regulated will not permit plant used for rendering unremunerative services to be figured in the rate base. It will not reason from the fact that the regulated firms are losing money in some markets to the conclusion that it should forbid the entry of new firms into others; rather, it will invite the firms to terminate the losing services. It will insist that the regulated firm adopt accounting procedures that reveal the true costs of serving various classes of customers. These regulatory checks may be far from entirely effective in practice, but it does seem unlikely that, if regulatory agencies had no independent reasons for encouraging internal subsidization, regulated firms could nonetheless engage in the practice on the scale they do.

In sum, existing views of regulation do not explain well the important phenomenon of internal subsidization. A new approach is needed and, in the next part, proposed.

3. Internal subsidization as a branch of public finance

■ Regulation and public finance are ordinarily considered unrelated activities. Occasionally the language of one laps over to the other, and we speak of a monopolist "taxing" his customers, redistributing wealth from them to him, by charging a price above cost. The internal subsidy, it seems to me, is an aspect of public finance in what is at once a more exact and a more natural sense. Taxation in common parlance refers to the use of the powers of the state to extract money from its subjects in order (1) to defray the cost of services that the politically dominant elements of the state wish to provide and

that the market would not provide in the desired quantity and at the desired price, or (2) to transfer money from one group to another, or (3), often, to do both. By this test regulation is in part a system of taxation or public finance. The basic mechanism is the internal subsidy. A firm provides a service below its real cost, and the deficit is made up by (usually) other customers of the firm who pay higher prices than they would otherwise. Were it not for the power of the state, acting through the regulatory agency, to control entry, the system would not be viable. A firm would not institute a losing service. If by mistake it did, it would terminate the service rather than subsidize the losses from profits in other markets. If it foolishly persisted, firms not burdened with the costs of losing services would enter the high-profit markets and their competition would drive down the price; deprived of the necessary supernormal profits, the firm would finally be compelled to terminate the unprofitable service.

Internal subsidization may thus be viewed as an exertion of state power whose purpose, like that of other taxes, is to compel members of the public to support a service that the market would provide at a reduced level, or not at all. It is in fact a form of excise tax, with the burden falling on purchasers of certain goods or services, and the proceeds earmarked for specific uses. As a form of excise tax, it invites comparison with other methods and objects of taxation.

4. Two case studies

■ It may be helpful at this point to illustrate the thesis that regulation is a method of public taxation and expenditure by some cases. I confine myself to two: the wall of protection thrown up around the international telegraph carriers by the Federal Communications Commission, and the methods of regulation used in the cable television industry. These case histories are no more than suggestive, but I believe a broader study of regulated markets would confirm the picture they present.

□ **International telegraph service.** Since this is one of the less well known regulated industries, it may be well to begin with some background.[27] U. S. companies providing international telegraph service (which includes telegrams, teletype, and transmission of computer data) are regulated by the FCC and known as "record carriers," because they transmit communications in the form of a written or otherwise stored record, as distinct from voice-only telephone transmission. There are three principal record carriers: ITT World Com, RCA Global Communications, and Western Union International. Until the 1950's they owned and operated submarine telegraph cables and high-frequency radio transmitting and receiving equipment. AT&T owned high-frequency radio equipment too, with which it supplied overseas telephone service (underwater cables lacked sufficient capacity for such service). AT&T did not offer record services. It seems to have been understood that the FCC would not permit the telephone company to enter the record market in competition with the record carriers.

A technological revolution in the provision of international telecommunication service occurred in the 1950's with the perfection of

[27] A more detailed description of the industry is given by the President's Task Force on Communications Policy in [49] and [50].

the voice-grade submarine cable, developed and installed by AT&T. These cables had sufficient capacity not only to enable telephone communication but also to permit the derivation of telegraph circuits at much lower cost than was possible to the record carriers using their existing, and now largely obsolete, equipment. Overseas telephone calls were cheaper and of much better quality than before, and since a telephone call is a substitute for a telegram or other record service, the new development placed a good deal of pressure on the record carriers. In addition, AT&T began to offer a new service, called "AVD" (alternative voice-data), under which the customer leased a circuit in the AT&T cable which he could use for voice or record service or both. Leased lines represented the most lucrative portion of the record carriers' services, and AT&T's new offering posed a serious threat to those carriers' continued profitability and even to their survival.[28]

They complained to the FCC. The Commission forbade AT&T to offer further AVD service (although AT&T was permitted to continue serving a few markets where the service had already been instituted) and granted the record carriers rights of co-ownership in AT&T's cable.[29]

The record carriers thus weathered this storm, but another was brewing. In 1962 the Communications Satellite Corporation was created to exploit the newly developed communications satellite. The enabling legislation[30] was ambiguous as to whether Comsat was free to deal directly with communications customers in addition to leasing circuits to the record carriers and AT&T. The issue came to a head when Comsat and the Department of Defense negotiated for the lease of thirty circuits in Comsat's Pacific satellite to the Department directly. Comsat offered a rate that was 32-to-40 percent of the rates offered by the record carriers.[31] The disparity was puzzling, since on leased circuits the record carriers do little more than arrange for obtaining a circuit either in one of AT&T's submarine cables, or, in the present case, in one of Comsat's satellites. After the record carriers agreed on a general rate reduction, although not to the level offered by Comsat, the Commission held that as a matter of policy (not law) it would forbid Comsat to deal directly with the communications user.[32]

Bare as this summary is, it indicates the salient fact of regulation in the international communications industry: the insulation of the record carriers from direct competition. AT&T could have competed directly with the record carriers before the invention of the modern submarine cable and especially afterward, as could Comsat, but the regulatory agency prevented such competition. The received views of regulation do not explain well this persistent, indeed dogged, policy of regulatory protection.

If regulation is imposed in order to prevent monopolistic distortions, one must find reasonable grounds for thinking that the entry of AT&T or Comsat into the record business would have reduced

---

[28] See FCC proceedings re the American Telephone and Telegraph Company [68], p. 1159 and [70], p. 433, n 9.

[29] Re The American Telephone & Telegraph Co. [68].

[30] Communications Satellite Act of 1962 [79].

[31] President's Task Force on Communications Policy [49], p. 19.

[32] *Authorized Entities and Authorized Users* [72].

rather than increased competition in that industry. Only if one attributes simplistic, although admittedly common, notions of competition to the FCC is it possible to construct a competitively based rationale for the denial of entry. The argument would be that AT&T controls the domestic leg of most international record traffic, and if permitted to offer international service as well would use its domestic position to monopolize the international service by routing all traffic to itself.

The fallacy in this reasoning lies in the assumption that a monopolist of one stage in the distribution of a good can increase his monopoly return by annexing successive stages as well. The assumption is in general false.[33] The successive stages of distribution are exactly analogous to the sale of complementary products, for example the head and shaft of a hammer. A monopolist of hammer heads could not increase his profits by getting control of the production and sale of the shafts as well. Any monopoly profits to be earned from controlling the manufacture of hammers could be captured by control of one essential component of hammers such as the heads. A monopolist's interests are ordinarily best served by minimizing the cost of complementary products, which will usually require the encouragement of competition in their provision. Thus, if AT&T indeed enjoys an effective monopoly of the domestic leg of international telegraph service, it should be able to extract all possible monopoly profits by means of appropriate charges for that leg. It would have an incentive to enter the international end of the business only if its costs in providing that service would be less than the rates charged by the existing carriers, in which event permitting it to enter would further the ends of competition.

Even if fear of AT&T's monopolizing the international business were plausible, it would not explain why the Commission also forbade Comsat to enter, or why the record carriers' rates at the time of Comsat's attempted entry were more than twice their costs.

The interest of the record carriers in repelling new entry is of course evident, but it seems unlikely that the protective measures taken by the Commission were merely in response to their demand. Although the determinants of political power are unfortunately not clearly understood, one is skeptical that the record carriers by themselves could have induced the Commission to subordinate to their protection the interests of AT&T and the Department of Defense, two of the nation's largest institutions, and of Comsat. The record carriers appear to lack the most important constituents of political influence. Their total revenues in 1968 were $173 million, in contrast to the Bell System's total earnings of $2.4 billion.[34] The record carriers have few employees[35] to go to bat on their behalf,

---

[33] See Bowman [4], p. 19. This statement and the analysis that follows in the text do not hold in a few special cases. Forward integration might facilitate price discrimination, enable the transfer of profits to an unregulated affiliate, forestall new entry, or simply complicate regulation of the primary market. None of these seem likely factors in the case under discussion.

[34] See Federal Communications Commission Annual Report [71].

[35] I am informed by the FCC that as of October 31, 1969, the international record carrier industry (the three major record carriers plus four smaller ones) had a total of 5623 employees in the United States and another 2432 abroad, of whom an undisclosed number are foreigners.

and the carriers' operations are confined to a handful of major cities—principally New York, San Francisco, and Washington[36]—in whose economies they play an insignificant role. One can challenge the comparison with Bell. Since two of the record carriers are parts of very large firms,[37] perhaps the proper comparison is between the parents' revenues (employees, etc.) and Bell's, or between the record carriers' revenues and those revenues of the Bell System that are earned in the international market alone. What this point ignores, however, is that virtually all of the Bell System's operations (even those regulated by state agencies) are directly or indirectly subject to control or influence by the FCC. The entire system, therefore, has a stake in decisions affecting a part. A decision limiting competition by AT&T in the international market could have precedential force in cases involving its right to compete in other markets. The common carrier interests of RCA and ITT, in contrast, are limited to the international telegraph industry.[38] The impact of regulatory decisions in that industry on the parent companies of the record carriers is consequently much smaller than the impact on the Bell System, so one would not expect them to devote comparable efforts to prevailing before the Commission.

A public finance view of regulation provides a clue to the curious success of the international telegraph industry in insulating itself from competition. The record carriers have long provided telegram service at a loss which they recoup by charging supracompetitive prices for other services, principally leased lines, where their costs, as mentioned, are small.[39] As a result, there is a class of customers who receive a service for which they would have to pay much higher prices were it provided in a free market; possibly the service would not be offered at all. These customers would be injured by any policies, such as free entry, which jeopardized the continued provision of international telegram service at the present attractive rates. They constitute allies of the record carriers in seeking the protection of the Commission against new entry and help to explain why such protection has been obtained in the face of strong opposition.

One would like to have a clearer idea of who the class of benefited customers is and how important cheap international telegraph service is to them. We can speculate that they are mostly small firms and individuals—travel agents, some importers, many tourists and their families—rather than large firms: a large user of international telecommunications service would lease circuits or at least subscribe to teletype service. But we know from other industries, notably the retail-drug industry, that small firms (when many in number) are frequently an effective political group in obtaining protective legislation for themselves. And even if the favored telegraph customers do not constitute an effective political group, perhaps they are viewed for one reason or another as particularly deserving, and on that ground favored. A careful empirical study might refute the suggestion that the interests of this customer class are a significant factor in the

[36] In re American Telephone & Telegraph Company [68], p. 1158.

[37] Western Union International is not affiliated with the domestic Western Union Company.

[38] ITT owns some South American telephone companies, but they are, of course, not regulated by the FCC.

[39] See *Authorized Entities and Authorized Users* [72], pp. 432–35.

actions of the Commission, but the important point is that such a study seems clearly indicated.

□ **Cable television.** The cable television industry,[40] in both its economic and its regulatory characteristics, resembles the public utility industries of the late nineteenth century. The industry is in its promotional phase. The systems already installed may be only a small fraction of what we can expect when it is mature. As with the earlier public utility industries, it appears that the returns to scale are large, and it may be inefficient for more than one cable television system to serve any given local area. There is a growing momentum of regulation. All municipalities now require providers of cable television service to obtain a municipal franchise; the franchises are becoming more elaborate; and some states are beginning to regulate the rates charged by cable firms.

On the basis of the received views of regulation, one might hypothesize that regulation had been imposed because the service was thought to be a natural monopoly, or had been demanded by the cable companies themselves in order to create or entrench a monopoly position. But neither hypothesis, separately or in combination, explains the specific types of regulation that have been imposed. On the one hand, the rates charged by the cable companies to their subscribers are rarely regulated (although, as mentioned, there is some movement in that direction); the thrust of regulation thus far has not been to eliminate monopoly pricing. On the other hand, municipal authorities have required cable franchisees to pay substantial fees, in money and kind, to the municipal government. These commonly take the form of the franchisee's agreeing to pay a percentage, sometimes as high as six percent, of gross revenues, coupled with his assuming an obligation to provide several channels in the cable system, without charge, to various municipal bodies such as the schools and the fire department. These exactions both reduce the value of the franchise to the cable firm and raise the price to the subscriber above what it would be if the service were an unregulated monopoly.[41]

Thus neither the subscribers nor the cable companies are clear gainers from the current regulatory policies. But they do generate municipal revenues. A tax is imposed on cable subscribers for the benefit of whoever watches the dedicated channels or partakes (either in reduced other taxes or greater municipal services) of the revenues generated by the franchise. In the latter case, indeed, internal subsidization has become conventional taxation in all but name.[42]

Possibly these burdens have been placed on the cable television industry at the behest of competitors, such as theater owners or local television stations (a hypothesis more plausible in the case of the percentage-of-receipts fee than in the case of the dedicated

---

[40] The discussion that follows draws heavily on Posner [45].

[41] See *id.* at pp. 16–19.

[42] Another interesting example of the interplay of explicit taxation and internal subsidies is found in *Student Educational Group Fares* [62], a decision of the Civil Aeronautics Board dealing with a recent Hawaiian statute that granted a tax rebate to airlines establishing a special group rate for students.

channels). In that event, the analogy would be to a tariff—another exercise of the public finance power.

5. Some general characteristics of regulation and the regulated industries that a public finance approach illuminates

■ We have seen that viewing regulation as a method of taxation or public finance appears to account, better than alternative views, for a major phenomenon found in regulated industries—internal subsidization. But beyond that, the view provides a consistent explanation of many other features of regulation and regulated industries, some of which fit poorly the received views.

□ **Regulatory control over entry.** Control of entry is an essential feature of regulation under the view advanced here because the adoption of a system of internal subsidies creates false pricing signals. Prices in certain markets must exceed costs if the losses sustained in providing the subsidized services are to be recouped. The price-cost spread in the subsidizing markets will naturally attract new entrants. But their costs may actually be higher than those of the existing firms, in which event their entry would produce a misallocation of resources. Entry would also impair or destroy the system of internal subsidies. With free entry, then, both efficiency and the subsidy scheme would be gravely endangered, so the regulatory agency must control entry.

To be sure, were regulation imposed for the sole benefit of the firms regulated, control over entry would also be necessary to prevent the dissipation through competition of the advantages secured to the incumbent firms by regulation. But not all important instances of entry control can be explained on this ground. The Post Office is not a profit-maximizing enterprise—it is in fact run at a deficit—but new entry into postal service is, and must be, barred, in order to preserve the uniform-rate structure and interclass subsidies that are a prominent feature of the Post Office's operation. Given the financial position of the railroads, it is doubtful that the control of entry in that industry is to be explained in terms of the interests of the regulated firms either.

The theory that regulation seeks to approximate the results of competition cannot explain the control of entry at all. If the regulated firm is constrained to sell at a price approximating cost, there will be no incentive for an inefficient firm to enter. If, despite regulation, the firm is charging a higher price, the matter is more complex. In general, however, assuming that differential pricing is feasible, as seems generally the case in decreasing-cost industries, a new entrant will not be attracted into such an industry by monopoly profits unless it is more efficient than the existing seller. The latter can repel entry by fixing a price near marginal cost to any customer solicited by the new entrant and will, because such a policy will not reduce his profits on any other sales (we have assumed he can maintain different prices) and the alternative is to lose a customer whom it is still profitable to serve even at the reduced price. Unless it is a more efficient firm, the new entrant will have higher costs and will not be able to meet the low price. Thus, in the case where public utility regulation is most plausibly explained in terms of an efficiency rationale—where the industry regulated is a decreasing-cost industry—the rationale still will not explain an important feature of

that regulation, the control of entry, because there is no reason to anticipate inefficient or excessive entry in the absence of public control.

□ **Regulatory review of new construction.** Firms subject to public utility or common carrier regulation are commonly required to obtain the permission of the regulatory agency for any major new construction.[43] This control is to be distinguished from control over entry: it applies whether the purpose of the construction is to enable the firm to enter a new market or to serve an existing market. In arguing that regulation is for the exclusive benefit of the regulated firms, one could point out that such control enables an agency to prevent the firms from expanding production in a way that might undermine cartel pricing. Although the provisions of regulatory statutes giving the agencies authority to fix prices and levels of service, requiring the regulated firms to embody all offerings in published tariffs, and forbidding under severe penalty any deviation from the tariff filings, already give an agency broad authority to enforce cartel pricing,[44] control of new construction makes the agency an even more effective enforcer of the cartel.

A consumer-interest view of regulation also provides an explanation, although not a very satisfactory one, of the control of new construction. While there is little solid basis for fear that an unregulated firm, even if a monopolist, would adopt an extravagant construction program,[45] there are, as noted earlier, some reasons for concern that regulated firms would not minimize costs. But it is unlikely that recognition of the side effects of regulation provides a general explanation of the power over new construction. Many regulated firms subject to the power are not monopolists, and the earlier analysis would not apply. If simply a fear of poor management was in the minds of those who framed the various public utility and common carrier statutes, one wonders why such statutes do not give the regulators more direct authority over management.

I suspect that the framers may have been motivated by a somewhat different concern from those previously mentioned—one that arises from the public finance function of regulation. An illustration, again drawn from the international communications industry, will help explain. In 1967, AT&T, acting this time in concert with the record carriers (and several foreign carriers), applied to the FCC for permission to build a fifth voice-grade cable across the Atlantic Ocean ("TAT 5").[46] The cost of constructing the facility was estimated to be $70 million and the planned capacity was 720 voice circuits. Comsat opposed the application. It pointed out that by the time TAT 5 was installed, very large satellites (5000 circuits each) would be in service above the Atlantic and these satellites would provide sufficient capacity to meet all reasonably foreseeable increments of demand for transatlantic telecommunications service at

[43] See, for example, the Communications Act of 1934 [78], sec. 214(a); the Natural Gas Act [81]; and the Transportation Act of 1920 [65].

[44] MacAvoy [36] is a case study of the role of regulatory controls in effectuating cartel pricing.

[45] See Posner [47], pp. 573–77.

[46] The incident is discussed by the staff of the President's Task Force on Communications Policy in [49], pp. 35–49.

a cost per circuit that would be only a small fraction of TAT 5's. AT&T, in reply, noted that the satellites in question might not be in service in time to avoid a shortage. But in that event, judging from subsequent filings and analysis in the proceeding, the economical solution would be to permit queuing, or use peak-load pricing, or launch an additional satellite of an older model. The staff of President Johnson's Task Force on Communications Policy, which analyzed AT&T's application in some depth, concluded that TAT 5 was the least economical alternative.[47]

The cost questions were in fact quite complex and the correctness of the staff's analysis perhaps debatable. The opinions in the case suggest, however, that the FCC itself doubted whether TAT 5 was cost justified. The Commission expressly refused to compare the costs of the cable with those of alternative satellite facilities,[48] adhering to this position in the face of a strong dissenting opinion in which it was urged that the cable was indeed more costly.[49] The majority cannot have been optimistic as to what an analysis of costs would have shown.

In approving the application, the Commission appears to have been strongly influenced by considerations that cannot be understood save in terms of a public finance approach to regulation, such as AT&T's representation that if TAT 5 were approved, it would be able to reduce its transatlantic telephone rates by 27 percent. It is at first glance surprising that the FCC should have been impressed by this offer. If satellites were a cheaper means of meeting demand than the cable, then rates could be reduced by even more than 27 percent if AT&T, rather than building a new cable, leased circuits from Comsat: so why did the Commission refuse to compare cable and alternative satellite costs? The probable explanation lies in Comsat's rate structure. Comsat is wedded, largely it appears for reasons of foreign relations, to a system of uniform global pricing under which the price of a circuit in a Comsat Pacific satellite is roughly the same as the price of a circuit in one of its Atlantic satellites. Because the Atlantic routes are busier, the Atlantic satellites are more fully utilized and the cost per circuit accordingly lower than in the Pacific. But this cost difference is not reflected in the rates, which, as noted, are roughly the same in both markets. Consequently, when AT&T leases circuits from Comsat for transatlantic service, it is forced to pay a considerable premium above the actual cost of the circuits to Comsat, so much so that the price to AT&T (after correcting for certain quality differences) is not clearly lower than the cost to it of circuits in a new cable. It is thus understandable why AT&T should have pushed for approval of TAT 5. But while from its standpoint cable costs may not have been higher than satellite costs, from the broader social standpoint they were (assuming that the staff analysis referred to earlier was correct). It is to prevent unwarranted

[47] See *id.*, pp. 36–41a.

[48] Re American Telephone & Telegraph Company [70], pp. 242–43 and n. 4.

[49] Re American Telephone & Telegraph Company [69], pp. 962–63 and [70], pp. 261–62.

[50] The staff of the President's Task Force on Communications Policy estimated that if the price per circuit in an Atlantic satellite had been determined on the basis of the costs of that satellite (and its associated ground facilities), it would have been $22,400 per half circuit per year in 1970, rather than the actual price of $31,300 set by Comsat.

investments based on divergences between private and social cost calculations caused by internal subsidization that regulatory agencies must have authority over the construction programs of regulated firms even when entry into a new market is not contemplated.

In this case, to be sure, the agency's exercise of its duty was perfunctory at best. Given the circumstances, however, that is not surprising. First of all, the program of internal subsidies that was jeopardized by the grant of the application—Comsat's policy of uniform global pricing of satellite circuits—is not one that the Commission has particularly encouraged. The motives behind it are rooted, as mentioned, in foreign-policy considerations that are the responsibility of other agencies. By granting the application the Commission was able to obtain immediate rate reductions for one of its constituencies, users of transatlantic telecommunications service, and the costs to the Commission cannot have seemed large.

Secondly, the Department of Defense made strong representations to the Commission[51] that the construction of TAT 5 would promote national security—another example of internal subsidization at work. The Department could have requested an appropriation from Congress to contribute to the cost of building a cable not justified by purely civilian demands; prevailing upon the FCC to authorize such a facility was an alternative mode of financing this defense project. The method of obtaining the "appropriation" and the class of "taxpayers" were different, but the object was the same. The subsidization of defense needs appears to be a rather general feature of regulation.[52]

□ **The duty of the regulated firm to serve and regulatory power over the abandonment of service.** Two long-established and complementary features of the regulatory process are the duty of regulated firms to serve all who demand service and the prohibition against such firms' discontinuing a service without the authorization of the regulatory agency. Although the prohibition of arbitrary refusals of service lies close to the heart of the traditional common-law concept of a public utility or common carrier (as the very name, common carrier, suggests) and is a settled feature of regulatory law,[53] it is difficult to explain under existing views of the purpose in fact of regulation. It is not apparent why regulated firms would want to be placed under such a duty or how they might benefit from it; it is only a little less difficult to see why, from the standpoint of consumer interests, the imposition of such a duty would be thought an appropriate part of the regulatory system. To be sure, a monopolist, if he has his way, will establish a schedule of prices under which fewer customers are served than if a competitive price were set; but once the schedule is adopted there is no reason to expect him to refuse service, on any

---

[51] Alluded to in [69], p. 961, the dissenting opinion of Commissioner Johnson.

[52] See, for example, Bonbright [3], p. 113, and National Transportation Policy in [64].

[53] A typical statement of the duty appears in the Interstate Commerce Act [63]: "It shall be the duty of every common carrier subject to this part to provide and furnish transportation upon reasonable request therefor . . . ." Other examples are given in Jones [28], pp. 376–85. See also *id.* at pp. 26–27. The subject is treated exhaustively by Wyman in [84].

but good business grounds, to any customer willing to pay the price.[54] Arbitrary refusals do not make good business sense. There is similarly no reason to expect a commercial enterprise to abandon a profitable service. Yet regulated firms are forbidden to abandon any route without obtaining the permission of the regulatory agency,[55] and bitterly contested abandonment proceedings are a commonplace occurrence, especially in the railroad industry.

Perhaps these controls are designed in many instances to reinforce regulatory control over the profits of the regulated firms: a firm might refuse or terminate service in order to coerce a higher rate from the customer, or as part of a scheme for enhancing its profits by reducing the level of service on which the rates it was permitted to charge had been based. Possibly they were intended to reinforce ordinary contractual remedies for nonperformance of services considered "essential." But these considerations do not provide a complete explanation. They do not explain why regulatory agencies are empowered to require the extension of utility services to new areas and to prevent the discontinuance of manifestly unremunerative services such as long-distance passenger transportation in the railroad industry. These cases can only be explained, I believe, in terms of a public finance view of regulation. Regulated firms, were they not subject to the duty to institute and not to terminate service, could not be relied upon to implement policies of internal subsidization. For reasons to be noted later, they might still offer some unremunerative services and they might still not always discontinue services when they ceased to be remunerative. But there would be no assurance of their cooperation.

□ **Competitive market structures.** The public finance hypothesis also illuminates some of the important characteristics of the regulated industries themselves (as distinct from characteristics of the regulatory schemes). It suggests, for example, why so many regulated industries do not have a monopolistic structure. A program of internal subsidies does not depend on the regulated industry's being a monopoly. So long as the demand for the industry's product is not perfectly elastic, and so long as the obligation to provide internal subsidies is imposed on all the firms in the industry, such a program is feasible just as it is feasible to impose an excise tax on a competitive industry. It is therefore not surprising, under the view advanced here, that many regulated industries are not monopolistic in structure. To be sure, were regulation imposed solely at the behest of regulated firms, one would also expect many regulated industries to be competitive in structure. But one would not expect so many regulated markets (especially in the communications, power, and water-supply industries) to exhibit pronounced characteristics of natural monopoly. These are the least likely settings for firms to seek governmental protection from competition since the existence of a natural monopoly substantially reduces the danger of competition.

---

[54] See Posner [47], p. 584. I except refusals to serve based upon widespread racial prejudice in a community and refusals to serve business customers for monopolistic reasons. These are special cases, and the scope of the duty to serve is considerably broader.

[55] See Jones [28], pp. 385–95.

□ **Regulated industries produce services.** It is a curiosity that public utility and common carrier industries invariably provide services (in the sense of a good that is difficult or impossible to store or transfer) rather than commodities. The public finance view supplies an answer. A subsidized good or service will not in fact be used by those for whom it is intended if they are free to resell it on the free market, which is why direct subsidies are commonly of services rather than commodities.[56]

□ **Regulated industries provide "infrastructure" services.** The specific complex of controls over entry and over the level and structure of rates that is characteristic of public utility and common carrier regulation is confined, for the most part, to the transportation, telecommunication, and power (electricity and gas) industries. Neither of the received views of regulation explains adequately why these particular industries have been singled out. The consumer-interest view of regulation would suggest that these were industries in which monopolistic misallocations of resources were most likely to occur, yet, as mentioned, many of the industries are naturally competitive rather than monopolistic, while a number of important industries, such as computers, drugs, newspapers, and certain nonferrous metals, which appear to have monopoly problems, sometimes quite serious ones, have escaped regulation. On the other hand, if we assume that regulation is imposed primarily for the benefit of the regulated firms, it must be shown why other industries have not obtained the same kind of regulation as public utilities and common carriers.

A partial explanation of the identity of the regulated industries may be that society frequently subjects to the public utility type of control services that it wants provided on the broadest possible basis (in a sense to be defined). The regulated industries are part of the "infrastructure" of economic growth. Adequate transportation, communications, and power (especially electrical) must be in place before the development of modern industry is possible, and most countries, including this one at various periods, have undertaken to subsidize these services or provide them directly in the hope thereby of attracting industrial developers.[57] One can deny the necessity or appropriateness of this state promotional role but hardly its prevalence. And internal subsidization is one method whereby the expansion of the infrastructure services can be promoted.

To be sure, it is not "expansion" in any simple or obvious sense that is involved. In the case of a naturally competitive industry, internal subsidies expand the provision of service to one class of customers, the beneficiaries of the program, but contract it to another: those who must pay a higher price to defray the subsidy and who consequently demand (and are supplied) a smaller quantity. The overall output of the industry is not necessarily larger, and may (as we saw in discussing value-of-service pricing) be smaller. If the industry is monopolistic in structure and it is not feasible to control its monopoly profits directly, a program of internal subsidies may well

[56] Shoup [55], p. 160; Stigler [58], p. 5.

[57] See, for example, Locklin [34], pp. 101–11, and Smead [56]. The theory of social overhead capital (as investment in infrastructure is often called by economists) is discussed by Hirschman [25], pp. 83–97.

bring about a larger output than otherwise. But in either case it would appear that the primary effect of such a program is not to increase the amount of transportation, communication, or power produced but rather to extend the service to classes of customers and geographical areas that might not be served in a free market.

Such a result is nonetheless consistent with the thinking that underlies the desire to force the creation of an adequate infrastructure rather than let the market take its course. The basic assumption, correct or incorrect, is that private enterprise, due to lack of foresight, or imperfections in the capital market, or external economies, will forgo many investments in infrastructure that would be socially profitable.[58] One can argue from this that it is the role of the state to encourage precisely those infrastructure services that are unremunerative.

This view may be reinforced in some cases by another: concern with geographical concentration of population and economic activity. A program of internal subsidies that denies the cost advantages of proximity and density, as is often the case, encourages greater geographic dispersion. Cost advantages based on location, it need hardly be said, are no less real than those based on other factors. But governments, including our own, have frequently followed policies aimed at denying those advantages. Utility regulation is perhaps one of them.

The industries in which we find internal subsidies are commonly also recipients of at least some direct subsidies.[59] This correlation supports the view of regulation as a method of public finance, esespecially where, as in the case of the electrical and telephone subsidies doled out by the Rural Electrification Administration, the recipients of direct subsidies are not members of the industry at all (in the REA case, they are consumer cooperatives). In such a case the established firms in the industry benefit only insofar as the existence of the direct subsidy reduces the pressure on them to provide an internal subsidy, and the subsidy scheme is more convincingly interpreted as a method of obtaining greater service than as a device for enriching corporate treasuries.

To suggest that regulation is a method of promoting the expansion of infrastructure services is not, of course, to explain why it is chosen in preference to alternative methods, such as direct subsidies, or why, with respect to some infrastructure services, such as education, the public utility approach plays a very subordinate role.[60] A framework for answering this question is sketched in part 6, where we look at some of the advantages and disadvantages of regulation in comparison with other methods of taxation, and a highly tentative answer is suggested.

The infrastructure explanation for the identity of the regulated industries is far from being completely satisfactory. It hardly seems applicable when an internal subsidy is used to retard the decline of an old industry, such as railroad passenger service or telegraph service. In addition, the economic case for subsidizing infrastructure

[58] See, for example, Hagen [19], pp. 126–29, and Kindleberger [32].

[59] Cf. Harriss [21], p. 270.

[60] Although not an entirely negligible one, as attested by free cable-television channels for schools, the public-affairs programming obligations of broadcast licensees, and the proposed free interconnection for educational television.

services is often dubious. And internal subsidization seems a somewhat curious way to encourage the expansion of an industry since, as mentioned, the cost of the subsidy is borne by customers of the industry. Indeed, the obligation to provide service to all at a uniform price may retard the undertaking of new extensions of service.[61]

At the least, these considerations suggest that a thoroughgoing justification of internal subsidies on efficiency grounds is impossible. One can easily find examples where an internal subsidy works directly contrary to the dictates of efficient resource allocation. Thus, the subsidization of commuter railroad service aggravates an existing imbalance between private and social costs caused by the fact that individuals who are employed in cities and utilize urban public services can escape the costs of those services by living in a suburb and commuting. It would appear, therefore, that internal subsidies are frequently designed to redistribute wealth rather than to correct imperfections in the market.

6. Internal subsidization compared with other methods of public finance

Limitations of the device. To summarize the discussion at this point, there is persuasive evidence that an important purpose in fact of public utility and common carrier regulation is to compel, by the device of the internal subsidy, the provision of certain services in quantities and at prices that a free market would not offer, much as the conventional taxing-spending power is used to the same end. Serious discussion of the public finance component of regulation has been retarded, however, by a tendency to dismiss it out of hand as an implausible and inappropriate alternative to more conventional exertions of the taxing power. Two objections are usually advanced as conclusive. The first is that internal subsidization distorts the efficient allocation of resources; the second, that it tends to be arbitrary and inequitable. One sometimes hears it said, too, that taxation is the proper business of the legislature and not of regulatory agencies.

1) *Delegation.* To take the last point first, it is difficult to understand why the delegation of a part of the taxing power to appointive agencies, the regulatory commissions, should be thought to offend the principles on which our government is organized. Congress, acting from imperative reasons of practicality, has delegated much of its lawmaking power to appointive agencies. The Federal courts provide a conspicuous example, and the Internal Revenue Service one that is directly in point.

2) *Efficiency.* It is true that internal subsidization, by forcing prices in some markets above cost and prices in others below, distorts the allocation of resources. It creates a secondary inefficiency as well: the entry of new competitors into the high-price markets must be prevented by the regulatory agency lest the source of the internal subsidy be wiped out.[62] Where the high-price market is

[61] Coase [6], p. 139. The effect of uniform price systems with which Professor Coase was concerned—denial of service to high-cost customers willing to pay the full cost of serving them—can be avoided by combining a uniform price system for most customers with a system of surcharges for those who would not be served at the uniform price. This appears on casual observation to be the practice of the telephone industry in this country.

[62] For a good example, see Caves [5], p. 314.

a natural monopoly, this is not an acute problem, but of course not all markets subject to regulation are naturally monopolistic.

The criticism of internal subsidization as inefficient points to a real characteristic of the device; but as a criticism it is superficial. It measures the device against an ideal standard, and of course finds it wanting. The proper comparison is to other exercises of the taxing power. All methods of taxation distort the "optimum" allocation of resources—optimum, that is, without regard to any need or demand to provide certain services publicly—and there are no *a priori* grounds for assuming that excise taxes, such as the internal-subsidy programs imposed by regulatory agencies, produce worse misallocations than income or other taxes.[63] To consider an important example, the exemption from income taxation of the real but not pecuniary income generated by housewives must cause a significant misallocation of resources by inducing many women to stay at home who would be more productive in other employments. The administrative costs of implementing a broader income concept would be so great, however, that this exemption is probably a permanent feature of income taxation. Because of pervasive and ineradicable distortions of this kind, it is not obvious that raising income tax rates would be a more efficient method of providing particular services at below-market prices than internal subsidization.[64] Indeed, insofar as the burden of internal subsidies tends to be borne by customers whose demands are highly inelastic, the allocative effects may be less adverse than those of alternative taxation methods.[65] And in those cases where the regulated firms are obtaining monopoly profits, the adverse allocative effects of the tax will be even fewer.[66]

Internal subsidies are also criticized on the ground that a subsidy in kind is inefficient compared to an unrestricted cash subsidy, because different people have different needs and wants. This is a valid and important point but it is not a criticism of internal subsidies as such, since it applies with equal force to most direct subsidies.

3) *Equity.* Because the determination of the incidence of particular taxes is immensely complex, it is very difficult to gauge the effect of internal subsidies on the distribution of income. At a rough guess, internal subsidization may sometimes benefit the poor[67] but has no general tendency to do so; and as our commuter example shows it may sometimes work in the opposite direction. But poverty is not the only possible justification for the redistribution of income. It is notable that internal subsidization is frequently employed to bolster declining services or sectors; perhaps in these cases it is felt that there are important reliance interests (for example, in location proximate to a railroad line) that deserve protection. And even if no consistent equity justification is possible, that is no special criticism

[63] See Friedman [16], pp. 56–67, and Little [33], p. 608.

[64] Cf. Harberger [20], p. 58.

[65] Cf. Musgrave [39], p. 157.

[66] Cf. Shilling [54], p. 224; Shoup [55], p. 276; Dirlam and Kahn [13], p. 494.

[67] Some rather dubious examples are given in Bonbright [3], pp. 111–12. Lower prices to the poor could in some cases be explained as profit-maximizing price discrimination, and his examples may well be of this type. A better example is provided by the low rail rates for agricultural commodities (see p. 26 and fn. 21 *supra*), but it is possible that the benefits are largely captured by farmers (who may or may not be poor) rather than by consumers. Staudinger [57], p. 259, argues for using public utility pricing to redistribute income to the poor.

of internal subsidies: the redistributive effects of tax-cum-direct-subsidy programs appear in a surprising range of cases to be perverse.[68] If one is to oppose internal subsidies on equity grounds, it must be as part of a broader objection to the redistributive policies of the state.

I turn now to some other, less frequently discussed attributes of regulation as a method of public finance.

4) *Enforcement.* An important characteristic of taxation by regulation is difficulty (and expense) of enforcement. A firm that finds the provision of an unremunerative service irksome may try to terminate it by drastically reducing the quality of the service and then citing the resulting fall in demand as evidence that the public no longer wants the service. This is not so transparent a gambit as it may seem. Since the public is not paying the full cost of the service, it has a natural tendency to demand a very high (and correspondingly costly) level of service. The specification of an appropriate level involves an essentially arbitrary judgment and accordingly gives the firm some room for maneuver. Evidently degradation of service has played an important role in the termination of railroad passenger operations.[69]

The tendency of regulated firms to cheat in providing unremunerative services is probably quite general since, unless regulation is more effective than anyone thinks, a penny saved in skimping on an unremunerative service will not result immediately in a full penny reduction in the rates paid by customers of the firm's lucrative services. The finding in a recent study that the rates set by publicly owned electric utilities[70] are more uniform than those set by privately owned electric utilities supports this suggestion. Uniform rates, we saw, are a common method of internal subsidization; and one would expect a privately owned company to resist providing unremunerative services more energetically than a publicly owned one.

The tendency to cheat is not entirely a bad thing. It may result in a smaller subsidy than if direct subsidization were used, and given the forceful objections to many subsidies, this may be distinctly to the good.

5) *Public scrutiny.* A troubling characteristic of the internal subsidy is its low visibility, which impedes responsible review. The amounts and recipients of direct subsidies are ordinarily specifically stated, but this is not the case with internal subsidies. Since information is not a free good, a subsidy program whose magnitude requires computation is less apt to be challenged than one whose magnitude is patent.

This is a general criticism of hidden subsidies, of which internal subsidies in the regulated industries are only one variety. And it is easily overstated: extravagant subsidy programs sail through Congress with monotonous regularity. Full disclosure is a far from dependable test of whether legislation in the public interest will be adopted, because the public does not vote on specific pieces of legislation, but on representatives, and it is demonstrable that in a repre-

[68] Stigler [58], p. 1.

[69] Doubtless encouraged by the ICC's recent holding that it has no authority over the quality of rail passenger service [76].

[70] Peltzman [44].

sentative system much legislation benefiting special interests at the expense of the larger public will be enacted.[71] This is a basic insight of the effective-political-group theory of regulation. Furthermore, given the size of the Federal budget, the disclosure in an appropriation hearing of the amount of a subsidy may not always be an effective method of assuring a responsible review of the proposal's merits. Suppose that buried in the Defense Department's appropriation request there had been a small item for a contribution to the cost of building a transatlantic cable: can one be confident that it would have received the careful scrutiny of Congress?

Despite the last point, the concern about adequacy of scrutiny has greatest force, I believe, precisely with regard to internal subsidies for national defense. The Defense Department's role in the TAT 5 matter affords a good illustration. Had the Department been forced to include the item in its budgetary request to Congress, it would have had to weigh its importance against that of other national-defense programs. The defense budget is not limitless. The inclusion of the cable item might have compelled the Department to modify some other request. In the context of a regulatory proceeding, however, the cable represented a free good to the Department. The Department had no incentive to evaluate the benefits of TAT 5 to the national defense objectively; indeed, it had an incentive to exaggerate those benefits. The FCC could not exercise a critical scrutiny because it has no competence to deal with military questions. The competent agencies—Congress and the Bureau of the Budget (which reviews all Federal budgetary proposals before submission to Congress)—were bypassed.

6) *Manageability of regulation.* Another problem with internal subsidization is that it complicates an already barely manageable regulatory process. Because there is no objective basis for balancing off distributive benefits against allocative costs,[72] an agency concerned with subsidizing worthy groups is deprived of a clear-cut standard for resolving controversies over pricing and entry. Clear and definite standards are necessary to tolerable regulation.[73] Without a definite standard at the agency level, moreover, judicial review, a potentially

---

[71] See Stigler [59].

[72] In his recent study of the Federal Power Commission [35], pp. 288–89, Paul MacAvoy makes a valiant but, I believe, unsuccessful effort to do so. His position on determining the dollar value of an income redistribution is "that the government should decide, and it indicates value by the amount that consumer X can keep after taxes . . . ." The difficulty with such an approach is that the scheme of Federal taxation does not reflect any consistent or intelligible policy toward the equity of income redistribution—unless we are to assume, for example, that because some dividend income is exempt from tax, stockholders are to be deemed a favored class for purposes of evaluating the redistributions brought about by a monopolist or a regulatory agency. See Pechman [43] for a lively recent discussion. And many of the apparently distributive features of Federal income taxation, including the progressive principle itself, have been justified on grounds (such as benefits received) that have nothing to do with the equity of redistributing income, while other features (such as the nontaxability of real income received in the form of reductions in the prices of consumer goods) reflect purely administrative considerations.

[73] Friendly [17]. For a somewhat different path to the same conclusion, see Posner [46], pp. 84–85.

important check on regulatory excesses,[74] is likely to be ineffectual; the agency can give a plausible justification for any result. Multiple and conflicting standards also breed corruption.[75]

7) *Private demand.* Taxation by regulation, to be feasible, requires that there be sufficient demand in the private market to justify the imposition of the burden of the subsidy on the regulated firms. Where there is not, as in the railroad industry, the results can be disastrous for the industry. One may hazard the guess that regulation has frequently been the principal means of subsidizing infrastructure services for which there is a strong private demand, while in areas like national defense and education, where the market demand is probably small in relation to the amount of service that the state wishes to provide, other methods of subsidization have predominated.

□ **And some advantages.** The balance of advantages is not wholly against the choice of the internal subsidy as a method of public finance. We have indicated several respects in which it may be preferable to other methods, and there are others.

1) *Administrative expense.* Although enforcement of internal subsidization can, as mentioned, be quite costly (railroad abandonment proceedings are a case in point), there are certain offsetting factors. Since no cash transfers are involved in internal subsidization, it is possible to dispense with the frequently elaborate apparatus of a formal transfer program—application forms, disbursement machinery, and the like. Often, too, a program of internal subsidies is implemented simply by the regulated firms' averaging the costs of many customers in setting a rate, and where this is done the firms avoid the expenses that would be incurred in identifying the costs of finer groups of customers and adopting a more complex rate structure tailored to the different costs. This is not to suggest that expenses incurred in implementing the price system are normally wasted; but once it is decided not to allow the price system to control the allocation of resources, a choice implicit in the decision to subsidize, the elimination of some of those expenses may represent a real saving.

2) *Legislative capacity.* By shifting taxing power from Congress (or state legislative bodies) to administrative agencies, internal subsidization economizes on the legislature's time. This is an especially important consideration where the subsidy is of a kind that requires frequent adjustment or review. The ability of a legislature to transact business is obviously limited. Among the ways in which it can be conserved, perhaps the delegation of minor taxing functions to regulatory agencies is relatively efficient.

3) *Protection of expectations.* At least when imposed on a service from the outset, internal subsidies may be less disruptive of public and

---

[74] An expanded role for the courts in the review of regulatory action is urged in Turner [61], p. 386. His position can be defended on the ground that judges are more insulated from political-group pressures than regulatory agencies. Cf. Posner [46], p. 89.

[75] The classic instance is the corruption that beset the FCC, at the highest levels, in the 1950's. The problem revolved around the initial grant of broadcast licenses, where the Commission applied no standard but used a check-list of criteria, enabling any preconceived result to be rationalized. Jones [28], pp. 1081–84, lists 15 criteria.

commercial expectations than other new taxes. An example will illustrate. Suppose a community has pending before it several applications for a cable television franchise and would like to use a few channels in any cable television system that is constructed for municipal functions such as education. And suppose further that the feasible alternative methods of obtaining this service have been narrowed to two: a tax on the gross receipts of the barbers in the community, the proceeds to be used to purchase the channels from the cable franchisee, and a condition in the franchise requiring the franchisee to provide the channels to the school system at no charge. If the first alternative is chosen, the result will be a rise in the cost (and hence presumably price) of barbering, which will lead to a fall in the amount of barbering demanded and supplied. As a result, some of the resources used in barbering in the community will be idle during the period in which they are being redeployed. And there will be an outcry from the barbers. These economic and political costs, incurred by virtue of the change in the economic conditions of the business brought about by a new tax, can be avoided if the second alternative, an internal subsidy by the cable industry, is selected. Since the costs of the cable system are now higher, a smaller system will be built. But the efficient scale (consistent with the obligation to provide free channels to the franchising authority) will be known in advance; there will be no waste in achieving it, as in the barbering example. In fact one observes that public utility and common carrier regulation has typically been imposed upon new services, where it was possible by a system of internal subsidies to finance desired extensions of the service without disturbing settled activities. And perhaps these considerations explain why municipalities have latched onto cable television as an important new source of revenue.[76]

Nonetheless, the explanation is severely limited. The alternatives in our example were too narrow: the municipality could also have placed a gross-receipts or other tax on cable service and raised the money for the free channels that way. It did not have to use internal subsidization, although we have previously discussed some reasons why internal subsidization might sometimes be preferred to alternative forms of excise taxation.

4) *Justice.* There may be some appeal to the notion that it is more "just" for other customers of the same industry to bear the cost of a subsidy of the industry's service than to distribute that cost among the taxpaying public at large. The notion is a little peculiar, however. It is one thing to say that those who benefit from a service should bear its costs, and quite another to impute the cost of a subsidy to those customers who are quite prepared to pay the full cost of serving them.

A final reason for the choice of internal subsidization over alternative methods of public finance has nothing to do with its relative merits. The regulated firms may cast their weight on the side of the internal subsidy, viewing customers who enjoy subsidized rates as useful allies in the maintenance of regulatory barriers to entry. Subsidizing some customers may be the "price" that the franchised monopolist pays for his monopoly. Perhaps careful study would disclose that most regulation is demanded by and supplied to a coalition

[76] See, for example, the Mayor's Advisory Committee [41].

of regulated firms and those of their customers who receive services below cost as a consequence of regulation.

7. Some practical suggestions

■ I trust that the foregoing remarks will not be construed as a "defense" of taxation (and subsidization) by regulation. They may, however, help explain the prevalence and tenacity of the practice, and they do suggest that, short of a thorough overhauling of government subsidy policy, it is less easy to condemn the practice out of hand as inefficient and inequitable than has usually been assumed. Perhaps few subsidies are in the public interest; there may still be cases where, given a decision to subsidize, regulation is the cheapest means of doing so.

But if we are stuck with taxation by regulation, perhaps we are not stuck with its worst features. I propose two modest reforms. The first is that agencies and reviewing courts insist, in proceedings where the maintenance of an internal subsidy is an issue, that the amount and cost of the subsidy, together with the identity of the recipients and of the payors, be calculated and placed in the public record. Perhaps this would eliminate some of the more captious instances of the phenomenon; at least it would bring an important issue of public policy into the open.

Second, more consideration should be given to the most efficient method of attaining the ends of internal subsidization. Accepting the decision to subsidize a specific service and to impose the cost of the subsidy on other customers of the firm providing the service, there may be better ways of achieving this end than control of prices, entry, abandonments, and the like by a regulatory agency. In particular, an explicit excise tax (such as the percentage-of-gross-receipts fee in many cable television franchises), with the proceeds earmarked for the service that the state wants to subsidize, may be preferable to the internal subsidy proper because it entails no limitation on entry into the high-price market; lump-sum fees may be preferable to either.[77] A likely reason why such alternatives are rarely considered is that the usual regulatory agency lacks authority to impose an explicit tax or other fee. In franchise regulation, as the case of cable television suggests, this option is open. Perhaps, therefore, a modest enlargement of the taxing power of regulatory agencies, to permit them to exact a uniform and limited fee from any firm desiring to enter a regulated market in lieu of other regulatory controls, would foster the more efficient use of what appears to be a settled device of public finance.

8. Conclusion

■ This paper merely scratches the surface of an interesting and important question of public policy. I have tried to show that certain views of the purpose in fact of public utility and common carrier regulation—that it is to approximate competitive results, or that it is to benefit the regulated firms—fail to account for a number of significant observed features of the regulatory process and the regulated industries. And I have argued that a consistent and comprehensive explanation of those features requires that we assign an important place to taxation and subsidization among the purposes that regula-

[77] See Posner [45] pp. 19–20.

tion in fact serves. I have attempted further to compare taxation by regulation with other methods of taxation and subsidization, in the hope of assisting evaluation of the pros and cons of alternative methods in particular cases. What I have not attempted to do is explain why some groups are subsidized and others not, or why the same group will receive some internal subsidies but not others (in the case of fire departments, for example, free water but not free telephone service). These fascinating and important questions, which require a better understanding of the magnitude and incidence of taxation by regulation than existing information permits, constitute the agenda for further research into a heretofore rather neglected aspect of public regulation.

References

1. AVERCH, H. and JOHNSON, L. L. "Behavior of the Firm Under Regulatory Constraint," *American Economic Review*, Vol. 52 (December 1962), pp. 1053–69.
2. BARATZ, M. "Cost Behavior and Pricing Policy in the Post Office," *Land Economics*, Vol. 38 (1962).
3. BONBRIGHT, J. C. *Principles of Public Utility Rates*. New York: Columbia University Press, 1961.
4. BOWMAN, W. "Tying Arrangements and the Leverage Problem," *Yale Law Journal*, Vol. 67 (1957).
5. CAVES, R. E. *Air Transport and its Regulators: An Industry Study*. Cambridge, Mass: Harvard University Press (Economic Studies Series, No. 120), 1962.
6. COASE, R. H. "The Economics of Uniform Pricing Systems," *Manchester School of Economic and Social Studies*, Vol. 15 (1947).
7. ———. "The Marginal Cost Controversy," *Economica*, Vol. 13 (1946), pp. 169–82.
8. ———. "The Nationalization of Electricity Supply in Great Britain," *Land Economics*, Vol. 26 (1950).
9. ———. "The Postal Monopoly in Great Britain: An Historical Survey," In *Economic Essays in Commemoration of the Dundee School of Economics, 1931–1955*, J. Eastham, editor, 1955.
10. CONANT, M. *Railroad Mergers and Abandonments*. Berkeley, California: University of California Press, 1965.
11. CREW, M. "Electricity Tariffs," In *Public Enterprise*. Edited by Ralph Turney. 1968.
12. DAVIDSON, R. *Price Discrimination in Selling Gas and Electricity*. 1954.
13. DIRLAM, J. and KAHN, A. "The Merits of Reserving the Cost Savings from Domestic Communications Satellites for the Support of Educational Television," *Yale Law Journal*, Vol. 77 (1968).
14. EADS, G. "The Effect of Regulation on the Cost Performance and Growth Strategies of the Local Service Airlines," unpublished manuscript, Princeton University Department of Economics, 1970.
15. FRIEDLAENDER, A. F. *Dilemma of Freight Transport Regulation*. Washington, D. C.: The Brookings Institution, 1969.
16. FRIEDMAN, M. *Price Theory: A Provisional Text*. Chicago, Ill.: Aldine Press, 1966 (rev. ed).
17. FRIENDLY, H. J. *Federal Administrative Agencies: The Need for Better Definition of Standards*. Cambridge, Mass.: Harvard University Press, 1962.
18. GARFIELD, P. and LOVEJOY, W. *Public Utility Economics*. Englewood Cliffs, New Jersey: Prentice-Hall, 1963.
19. HAGEN, E. E. *The Economics of Development*. Homewood, Ill.: Richard D. Irwin, 1968.
20. HARBERGER, A. C. "Taxation, Resource Allocation, and Welfare," in NATIONAL BUREAU OF ECONOMICS RESEARCH and THE BROOKINGS INSTITUTION, *The Role of Direct and Indirect Taxes in the Federal Revenue System: A Conference Report*. Princeton, N. J.: Princeton University Press, 1964.
21. HARRISS, L. "Subsidies in the United States," *Public Finance*, Vol. 16 (1961).

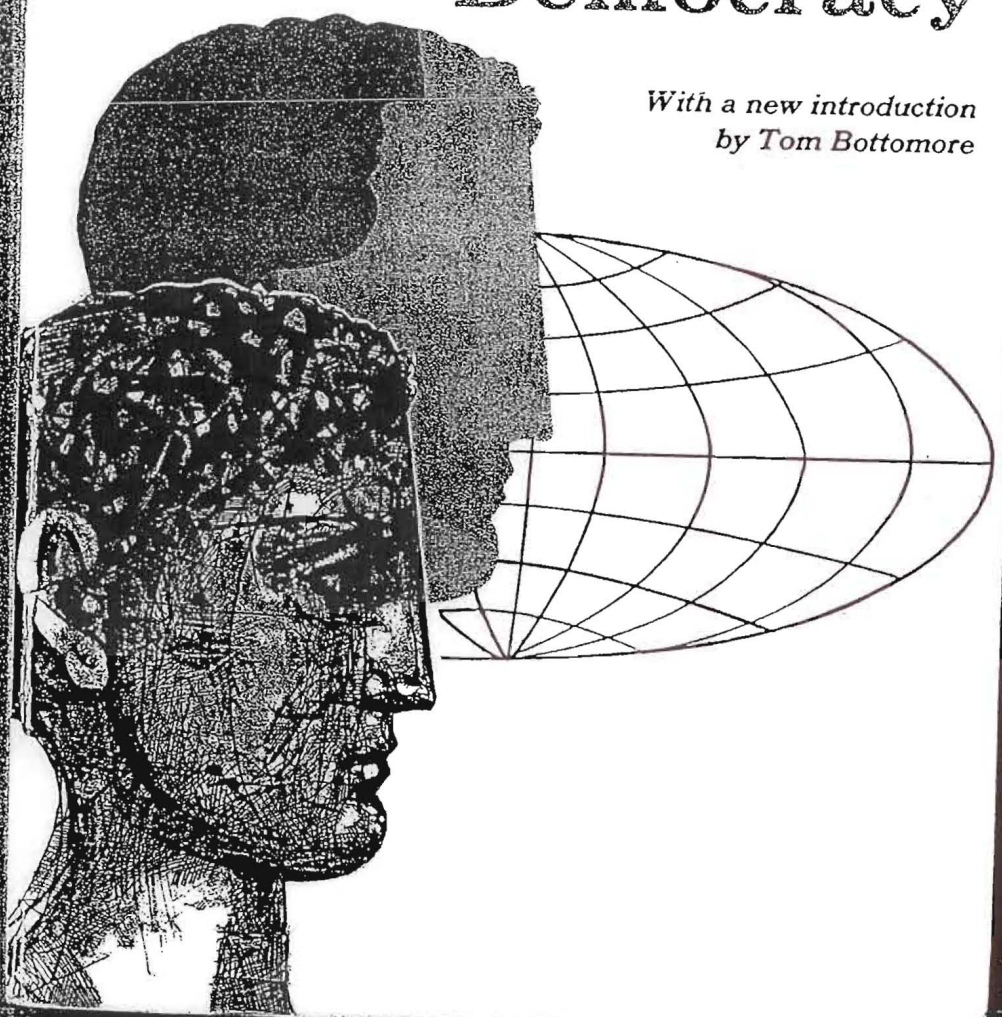
22. HENDERSON, A. "The Pricing of Public Utility Undertakings," *Manchester School of Economic and Social Studies*, Vol. 15 (1947).
23. HILTON, G. W. *The Transportation Act of 1958: A Decade of Experience*. Bloomington, Indiana: Indiana University Press, 1969.
24. HIRSHLEIFER, J., DE HAVEN, J. and MILLIMAN, J. *Water Supply: Economics, Technology, and Policy*. Chicago, Ill.: University of Chicago Press, 1960 (rev. ed, 1969).
25. HIRSCHMAN, A. O. *The Strategy of Economic Development*. New Haven, Conn.: Yale University Press, 1958.
26. ILLINOIS, STATE OF. "An Act to provide for the establishment of ferries, toll bridges, and turnpike roads," *Revised Code of Laws of Illinois*, Sec. 5 (1827).
27. JOHNSON, L. *Communications Satellites and Telephone Rates: Problems of Government Regulation*. Rand Memorandum RM-2845-NASA (October 1961).
28. JONES, W. K. *Cases and Materials on Regulated Industries*. 1967.
29. KENNEDY, J. "Structure and Policy in Postal Rates," *Journal of Political Economy*, Vol. 65 (1957).
30. KEETON, R. E. and O'CONNELL, J. *Basic Protection for the Traffic Victim: A Blueprint for Reforming Automobile Insurance*. Boston, Mass.: Little, Brown & Company, 1966.
31. KEIG, N. G., FRISTOE, C. W. and GODDARD, F. O. "A Critique of the Policy Objectives of Publicly Owned Water Utilities," Unpublished manuscript, University of Florida, 1970.
32. KINDLEBERGER, C. P. *Economic Development*. New York, N. Y.: McGraw Hill, 1958 (2nd ed. 1965).
33. LITTLE, I. M. D. "Direct Versus Indirect Taxes," *Economic Journal*. Vol. 61 (1951). Reprinted in AMERICAN ECONOMICS ASSOCIATION, *Readings in Welfare Economics*, 1969.
34. LOCKLIN, D. P. *Economics of Transportation*. Homewood, Ill.: Richard D. Irwin, 1966 (6th ed.).
35. MACAVOY, P. W. "The Effectiveness of the Federal Power Commission," *Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autumn, 1970), pp. 271–303.
36. ———. *The Economic Effects of Regulation: The Trunk-Line Railroad Cartels and the Interstate Commerce Commission before 1900*. Cambridge, Mass.: MIT Press, 1965.
37. MEYER, J. R., KAIN, J. F. and WOHL, M. *The Urban Transportation Problem*. Cambridge, Mass.: Harvard University Press, 1965.
38. ———, PECK, M., STENASON, J. and ZWICK, C. *The Economics of Competition in the Transportation Industries*. Cambridge, Mass.: Harvard University Press, 1959.
39. MUSGRAVE, R. *The Theory of Public Finance: A Study in Public Economy*. New York: McGraw Hill, 1959.
40. NELSON, J. C. *Railroad Transportation and Public Policy*. Washington, D. C.: The Brookings Institution, 1959.
41. NEW YORK MAYOR'S ADVISORY COMMITTEE ON CATV AND TELECOMMUNICATIONS. *A Report on Cable Television and Cable Telecommunications in New York City*. New York, 1968.
42. NOVE, A. "Internal Economies," *Economic Journal*, Vol. 79 (1969).
43. PECHMAN, J. A. "The Rich, the Poor, and the Taxes They Pay," *The Public Interest*, No. 17 (Fall 1969), pp. 21–43.
44. PELTZMAN, S. "Pricing in Public and Private Enterprises: Electric Utilities in the United States," *Journal of Law and Economics*, forthcoming.
45. POSNER, R. A. *Cable Television: The Problem of Local Monopoly*. Rand Memorandum RM-6309-FF (May 1970).
46. ———. "The Federal Trade Commission," *University of Chicago Law Review*, Vol. 37 (1969).
47. ———. "Natural Monopoly and Its Regulation," *Stanford Law Review*, Vol. 21 (1969).
48. PRESIDENT'S TASK FORCE ON COMMUNICATIONS POLICY. *Final Report*. Washington, D. C.: Government Printing Office, 1968.
49. ———. *Organization of the United States International Communications Industry*. U. S. Department of Commerce, Clearinghouse for Federal Scientific and Technical Information, PB 184 424 (June 1969). Staff Paper 2.

50. ———. *The Western Union Telegraph Company*. U. S. Department of Commerce, Clearinghouse for Federal Scientific and Technical Information, PB 184 418 (June 1969). Staff Paper 5.
51. SARGENT, J. R. "Nationalized Industries and Public Policy," in *The Lessons of Public Enterprise*, edited by Michael Shanks (1963).
52. SHARP, C. *Problems of Urban Passenger Transport with Special Reference to Leicester*. 1967.
53. SHEPHERD, W. G. "Cross-Subsidizing and Allocation in Public Firms," *Oxford Economic Papers*, Vol. 16 (N.S., 1964).
54. SHILLING, N. *Excise Taxation of Monopoly*. New York: Columbia University Press, 1969.
55. SHOUP, C. S. *Public Finance*. Chicago, Ill.: Aldine Press, 1969.
56. SMEAD, E. E. *Governmental Promotion and Regulation of Business*. New York: Appleton-Century Crofts, 1969.
57. STAUDINGER, H. "Social Rates in Electricity," *Social Research*, Vol. 3 (1936).
58. STIGLER, G. J. "Director's Law of Public Income Redistribution, " *Journal of Law and Economics*, Vol. 13 (1970).
59. ———. "Theory of Economic Regulation," *Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971).
60. ———. *The Theory of Price*. New York: The Macmillan Company, 1966 (3rd ed).
61. TURNER, D. "The Scope of Antitrust and Other Economic Regulatory Policies," *Harvard Law Review*, Vol. 82 (1969).
62. U. S., CIVIL AERONAUTICS BOARD. *Student Educational Group Fares*. Docket No. 22402, order of investigation and suspension, July 29, 1970.
63. *U. S. Code*. "Interstate Commerce Act," Vol. 49, Sec. 1(4).
64. ———. "National Transportation Policy," Preceding Vol. 49, Sec. 1.
65. ———. "Transportation Act of 1920," Vol. 49, Secs. 1(18), 1(20), and 1(22).
66. U. S., CONGRESS, HOUSE, COMMITTEE ON INTERSTATE AND FOREIGN COMMERCE. *Transportation Acts Amendments, 1962, Hearings on*. 87th Cong., 2nd Session.
67. U. S. DEPARTMENT OF TRANSPORTATION. *Price Variability in the Automobile Insurance Market*. Automobile and Compensation Study, August 1970.
68. U. S., FEDERAL COMMUNICATIONS COMMISSION. *Re American Telephone and Telegraph Company*, 37 F.C.C. 1151 (1964).
69. ———. *Re American Telephone and Telegraph Company*, 11 F.C.C. 2d 957 (1968).
70. ———. *Re American Telephone and Telegraph Company*, 13 F.C.C. 2d 235 (1968).
71. ———. *Annual Report, 1969*. Washington, D. C.: U. S. Government Printing Office, 1970.
72. ———. *Authorized Entities and Authorized Users Under the Communications Satellite Act of 1962*, 4 F.C.C. 2d 421 (1966).
73. ———. *Interconnection Service*, 14 F.C.C. 2d 599 (1968).
74. ———. *Public Service Responsibility of Broadcast Licensees* (1946).
75. ———. *Re Western Union Telegraph Company*, 12 F.C.C. 2d 980 (1968).
76. U. S., INTERSTATE COMMERCE COMMISSION. *Adequacy—Passenger Service—Southern Pac. Co. Between California and Louisiana*, 335 I.C.C. 415 (1969).
77. ———. *Increased Freight Rates, E.W. & S. Territories, 1956*, 299 I.C.C. 429, 441, 451–59 (1956).
78. *U. S. Statutes at Large*. Communications Act of 1934, as amended by the Public Broadcasting Act of 1967 (U. S. Statutes at Large, vol. 81, p. 365), U. S. Code Vol. 47, Secs. 1 *et seq*.
79. ———. Communications Satellite Act of 1962, Vol. 76; U. S. Code, Vol. 47, Secs. 701 *et seq*.
80. ———. Hoch-Smith Resolution, Vol. 43 (1925).
81. ———. Natural Gas Act, sec. 7(c), as amended. U. S. Code, Vol. 15, sec 717 ff.
82. WATKINS, G. P. "Street-Railway Rates, with Especial Reference to Differentiation," *Quarterly Journal of Economics*, Vol. 25 (1911).
83. WILSON, G. "Effects of Value-of-Service Pricing Upon Motor Common Carriers," *Journal of Political Economy*. Vol. 63 (1965).
84. WYMAN, B. *Public Service Corporations*. Vol. 1 (1911), Parts II–IV.

Joseph A. Schumpeter

Capitalism, Socialism and Democracy

With a new introduction by Tom Bottomore



the equilibrium in perfect competition, and still harder to preserve, but the "beneficial" competition of the classic type seems likely to be replaced by "predatory" or "cutthroat" competition or simply by struggles for control in the financial sphere. These things are so many sources of social waste, and there are many others such as the costs of advertising campaigns, the suppression of new methods of production (buying up of patents in order not to use them) and so on. And most important of all: under the conditions envisaged, equilibrium, even if eventually attained by an extremely costly method, no longer guarantees either full employment or maximum output in the sense of the theory of perfect competition. It *may* exist without full employment; it is *bound* to exist, so it seems, at a level of output below that maximum mark, because profit-conserving strategy, impossible in conditions of perfect competition, now not only becomes possible but imposes itself.

Well, does not this bear out what the man in the street (unless a businessman himself) always thought on the subject of private business? Has not modern analysis completely refuted the classical doctrine and justified the popular view? Is it not quite true after all, that there is little parallelism between producing for profit and producing for the consumer and that private enterprise is little more than a device to curtail production in order to extort profits which then are correctly described as tolls and ransoms?

CHAPTER VII

THE PROCESS OF CREATIVE DESTRUCTION

The theories of monopolistic and oligopolistic competition and their popular variants may in two ways be made to serve the view that capitalist reality is unfavorable to maximum performance in production. One may hold that it always has been so and that all along output has been expanding in spite of the secular sabotage perpetrated by the managing bourgeoisie. Advocates of this proposition would have to produce evidence to the effect that the observed rate of increase can be accounted for by a sequence of favorable circumstances unconnected with the mechanism of private enterprise and strong enough to overcome the latter's resistance. This is precisely the question which we shall discuss in Chapter IX. However, those who espouse this variant at least avoid the trouble about historical fact that the advocates of the alternative proposition have to face. This avers that capitalist reality once tended to favor maximum productive performance, or at all events productive performance so considerable as to constitute a major element in any serious appraisal of the system; but that the later spread of monopolist structures, killing competition, has by now reversed that tendency.

First, this involves the creation of an entirely imaginary golden age of perfect competition that at some time somehow metamorphosed itself into the monopolistic age, whereas it is quite clear that perfect competition has at no time been more of a reality than it is at present. Secondly, it is necessary to point out that the rate of increase in output did not decrease from the nineties from which, I suppose, the prevalence of the largest-size concerns, at least in manufacturing industry, would have to be dated; that there is nothing in the behavior of the time series of total output to suggest a "break in trend"; and, most important of all, that the modern standard of life of the masses evolved during the period of relatively unfettered "big business." If we list the items that enter the modern workman's budget and from 1899 on observe the course of their prices not in terms of money but in terms of the hours of labor that will buy them—i.e., each year's money prices divided by each year's hourly wage rates—we cannot fail to be struck by the rate of the advance which, considering the spectacular improvement in qualities, seems to have been greater and not smaller than it ever was before. If we economists were given less to wishful thinking and more to the observation of facts, doubts would

immediately arise as to the realistic virtues of a theory that would have led us to expect a very different result. Nor is this all. As soon as we go into details and inquire into the individual items in which progress was most conspicuous, the trail leads not to the doors of those firms that work under conditions of comparatively free competition but precisely to the doors of the large concerns—which, as in the case of agricultural machinery, also account for much of the progress in the competitive sector—and a shocking suspicion dawns upon us that big business may have had more to do with creating that standard of life than with keeping it down.

The conclusions alluded to at the end of the preceding chapter are in fact almost completely false. Yet they follow from observations and theorems that are almost completely[1] true. Both economists and popular writers have once more run away with some fragments of reality they happened to grasp. These fragments themselves were mostly seen correctly. Their formal properties were mostly developed correctly. But no conclusions about capitalist reality as a whole follow from such fragmentary analyses. If we draw them nevertheless, we can be right only by accident. That has been done. And the lucky accident did not happen.

The essential point to grasp is that in dealing with capitalism we are dealing with an evolutionary process. It may seem strange that anyone can fail to see so obvious a fact which moreover was long ago emphasized by Karl Marx. Yet that fragmentary analysis which yields the bulk of our propositions about the functioning of modern capitalism persistently neglects it. Let us restate the point and see how it bears upon our problem.

Capitalism, then, is by nature a form or method of economic change and not only never is but never can be stationary. And this evolutionary character of the capitalist process is not merely due to the fact that economic life goes on in a social and natural environment which changes and by its change alters the data of economic action; this fact is important and these changes (wars, revolutions and so on) often condition industrial change, but they are not its prime movers. Nor is this evolutionary character due to a quasi-automatic increase in population and capital or to the vagaries of monetary systems of

[1] As a matter of fact, those observations and theorems are not completely satisfactory. The usual expositions of the doctrine of imperfect competition fail in particular to give due attention to the many and important cases in which, even as a matter of static theory, imperfect competition approximates the results of perfect competition. There are other cases in which it does not do this, but offers compensations which, while not entering any output index, yet contribute to what the output index is in the last resort intended to measure—the cases in which a firm defends its market by establishing a name for quality and service for instance. However, in order to simplify matters, we will not take issue with that doctrine on its own ground.

which exactly the same thing holds true. The fundamental impulse that sets and keeps the capitalist engine in motion comes from the new consumers' goods, the new methods of production or transportation, the new markets, the new forms of industrial organization that capitalist enterprise creates.

As we have seen in the preceding chapter, the contents of the laborer's budget, say from 1760 to 1940, did not simply grow on unchanging lines but they underwent a process of qualitative change. Similarly, the history of the productive apparatus of a typical farm, from the beginnings of the rationalization of crop rotation, plowing and fattening to the mechanized thing of today—linking up with elevators and railroads—is a history of revolutions. So is the history of the productive apparatus of the iron and steel industry from the charcoal furnace to our own type of furnace, or the history of the apparatus of power production from the overshot water wheel to the modern power plant, or the history of transportation from the mail-coach to the airplane. The opening up of new markets, foreign or domestic, and the organizational development from the craft shop and factory to such concerns as U. S. Steel illustrate the same process of industrial mutation—if I may use that biological term—that incessantly revolutionizes[2] the economic structure *from within*, incessantly destroying the old one, incessantly creating a new one. This process of Creative Destruction is the essential fact about capitalism. It is what capitalism consists in and what every capitalist concern has got to live in. This fact bears upon our problem in two ways.

First, since we are dealing with a process whose every element takes considerable time in revealing its true features and ultimate effects, there is no point in appraising the performance of that process *ex visu* of a given point of time; we must judge its performance over time, as it unfolds through decades or centuries. A system—any system, economic or other—that at *every* given point of time fully utilizes its possibilities to the best advantage may yet in the long run be inferior to a system that does so at *no* given point of time, because the latter's failure to do so may be a condition for the level or speed of long-run performance.

Second, since we are dealing with an organic process, analysis of what happens in any particular part of it—say, in an individual concern or industry—may indeed clarify details of mechanism but is inconclusive beyond that. Every piece of business strategy acquires its true significance only against the background of that process and

[2] Those revolutions are not strictly incessant; they occur in discrete rushes which are separated from each other by spans of comparative quiet. The process as a whole works incessantly however, in the sense that there always is either revolution or absorption of the results of revolution, both together forming what are known as business cycles.

within the situation created by it. It must be seen in its role in the perennial gale of creative destruction; it cannot be understood irrespective of it or, in fact, on the hypothesis that there is a perennial lull.

But economists who, *ex visu* of a point of time, look for example at the behavior of an oligopolist industry—an industry which consists of a few big firms—and observe the well-known moves and countermoves within it that seem to aim at nothing but high prices and restrictions of output are making precisely that hypothesis. They accept the data of the momentary situation as if there were no past or future to it and think that they have understood what there is to understand if they interpret the behavior of those firms by means of the principle of maximizing profits with reference to those data. The usual theorist's paper and the usual government commission's report practically never try to see that behavior, on the one hand, as a result of a piece of past history and, on the other hand, as an attempt to deal with a situation that is sure to change presently—as an attempt by those firms to keep on their feet, on ground that is slipping away from under them. In other words, the problem that is usually being visualized is how capitalism administers existing structures, whereas the relevant problem is how it creates and destroys them. As long as this is not recognized, the investigator does a meaningless job. As soon as it is recognized, his outlook on capitalist practice and its social results changes considerably.[3]

The first thing to go is the traditional conception of the *modus operandi* of competition. Economists are at long last emerging from the stage in which price competition was all they saw. As soon as quality competition and sales effort are admitted into the sacred precincts of theory, the price variable is ousted from its dominant position. However, it is still competition within a rigid pattern of invariant conditions, methods of production and forms of industrial organization in particular, that practically monopolizes attention. But in capitalist reality as distinguished from its textbook picture, it is not that kind of competition which counts but the competition from the new commodity, the new technology, the new source of supply, the new type of organization (the largest-scale unit of control for instance)—competition which commands a decisive cost or quality advantage and which strikes not at the margins of the profits and the outputs of the existing firms but at their foundations and their very lives. This kind of competition is as much more effective than the other as a bombardment is in comparison with forcing a door, and

[3] It should be understood that it is only our appraisal of economic performance and not our moral judgment that can be so changed. Owing to its autonomy, moral approval or disapproval is entirely independent of our appraisal of social (or any other) results, unless we happen to adopt a moral system such as utilitarianism which makes moral approval and disapproval turn on them *ex definitione*.

so much more important that it becomes a matter of comparative indifference whether competition in the ordinary sense functions more or less promptly; the powerful lever that in the long run expands output and brings down prices is in any case made of other stuff.

It is hardly necessary to point out that competition of the kind we now have in mind acts not only when in being but also when it is merely an ever-present threat. It disciplines before it attacks. The businessman feels himself to be in a competitive situation even if he is alone in his field or if, though not alone, he holds a position such that investigating government experts fail to see any effective competition between him and any other firms in the same or a neighboring field and in consequence conclude that his talk, under examination, about his competitive sorrows is all make-believe. In many cases, though not in all, this will in the long run enforce behavior very similar to the perfectly competitive pattern.

Many theorists take the opposite view which is best conveyed by an example. Let us assume that there is a certain number of retailers in a neighborhood who try to improve their relative position by service and "atmosphere" but avoid price competition and stick as to methods to the local tradition—a picture of stagnating routine. As others drift into the trade that quasi-equilibrium is indeed upset, but in a manner that does not benefit their customers. The economic space around each of the shops having been narrowed, their owners will no longer be able to make a living and they will try to mend the case by raising prices in tacit agreement. This will further reduce their sales and so, by successive pyramiding, a situation will evolve in which increasing potential supply will be attended by increasing instead of decreasing prices and by decreasing instead of increasing sales.

Such cases do occur, and it is right and proper to work them out. But as the practical instances usually given show, they are fringe-end cases to be found mainly in the sectors furthest removed from all that is most characteristic of capitalist activity.[4] Moreover, they are transient by nature. In the case of retail trade the competition that matters arises not from additional shops of the same type, but from the department store, the chain store, the mail-order house and the supermarket which are bound to destroy those pyramids sooner or later.[5]

[4] This is also shown by a theorem we frequently meet with in expositions of the theory of imperfect competition, viz., the theorem that, under conditions of imperfect competition, producing or trading businesses tend to be irrationally small. Since imperfect competition is at the same time held to be an outstanding characteristic of modern industry we are set to wondering what world these theorists live in, unless, as stated above, fringe-end cases are all they have in mind.

[5] The mere threat of their attack cannot, in the particular conditions, environmental and personal, of small-scale retail trade, have its usual disciplining influence, for the small man is too much hampered by his cost structure and, however

Now a theoretical construction which neglects this essential element of the case neglects all that is most typically capitalist about it; even if correct in logic as well as in fact, it is like *Hamlet* without the Danish prince.

well he may manage within his inescapable limitations, he can never adapt himself to the methods of competitors who can afford to sell at the price at which he buys.

CHAPTER VIII

MONOPOLISTIC PRACTICES

WHAT has been said so far is really sufficient to enable the reader to deal with the large majority of the practical cases he is likely to meet and to realize the inadequacy of most of those criticisms of the profit economy which, directly or indirectly, rely on the absence of perfect competition. Since, however, the bearing of our argument on some of those criticisms may not be obvious at a glance, it will be worth our while to elaborate a little in order to make a few points more explicit.

1. We have just seen that, both as a fact and as a threat, the impact of new things—new technologies for instance—on the existing structure of an industry considerably reduces the long-run scope and importance of practices that aim, through restricting output, at conserving established positions and at maximizing the profits accruing from them. We must now recognize the further fact that restrictive practices of this kind, as far as they are effective, acquire a new significance in the perennial gale of creative destruction, a significance which they would not have in a stationary state or in a state of slow and balanced growth. In either of these cases restrictive strategy would produce no result other than an increase in profits at the expense of buyers except that, in the case of balanced advance, it might still prove to be the easiest and most effective way of collecting the means by which to finance additional investment.[1] But in the process of creative destruction, restrictive practices may do much to steady the ship and to alleviate temporary difficulties. This is in fact a very familiar argument which always turns up in times of depression and, as everyone knows, has become very popular with governments and their economic advisers—witness the NRA. While it has been so much misused and so faultily acted upon that most economists heartily despise it, those

[1] Theorists are apt to look upon anyone who admits this possibility as guilty of gross error, and to prove immediately that financing by borrowing from banks or from private savers or, in the case of public enterprise, financing from the proceeds of an income tax is much more rational than is financing from surplus profits collected through a restrictive policy. For some patterns of behavior they are quite right. For others they are quite wrong. I believe that both capitalism and communism of the Russian type belong in the latter category. But the point is that theoretical considerations, especially theoretical considerations of the short-run kind, cannot solve, although they contribute to the solution of, the problem which we shall meet again in the next part.

The Use of Knowledge in Society

F.A. Hayek

American Economic Review, XXXV, No. 4; September, 1945, pp. 519-30.

I

What is the problem we wish to solve when we try to construct a rational economic order? On certain familiar assumptions the answer is simple enough. *If* we possess all the relevant information, *if* we can start out from a given system of preferences, and *if* we command complete knowledge of available means, the problem which remains is purely one of logic. That is, the answer to the question of what is the best use of the available means is implicit in our assumptions. The conditions which the solution of this optimum problem must satisfy have been fully worked out and can be stated best in mathematical form: put at their briefest, they are that the marginal rates of substitution between any two commodities or factors must be the same in all their different uses. H.1

This, however, is emphatically *not* the economic problem which society faces. And the economic calculus which we have developed to solve this logical problem, though an important step toward the solution of the economic problem of society, does not yet provide an answer to it. The reason for this is that the "data" from which the economic calculus starts are never for the whole society "given" to a single mind which could work out the implications and can never be so given. H.2

The peculiar character of the problem of a rational economic order is determined precisely by the fact that the knowledge of the circumstances of which we must make use never exists in concentrated or integrated form but solely as the dispersed bits of incomplete and frequently contradictory knowledge which all the separate individuals possess. The economic problem of society is thus not merely a problem of how to allocate "given" resources—if "given" is taken to mean given to a single mind which deliberately solves the problem set by these "data." It is rather a problem of how to secure the best use of resources known to any of the members of society, for ends whose relative importance only these individuals know. Or, to put it briefly, it is a problem of the utilization of knowledge which is not given to anyone in its totality. H.3

This character of the fundamental problem has, I am afraid, been obscured rather than illuminated by many of the recent refinements of economic theory, particularly by many of the uses made of mathematics. Though the problem with which I want primarily to deal in this paper is the problem of a

rational economic organization, I shall in its course be led again and again to point to its close connections with certain methodological questions. Many of the points I wish to make are indeed conclusions toward which diverse paths of reasoning have unexpectedly converged. But, as I now see these problems, this is no accident. It seems to me that many of the current disputes with regard to both economic theory and economic policy have their common origin in a misconception about the nature of the economic problem of society. This misconception in turn is due to an erroneous transfer to social phenomena of the habits of thought we have developed in dealing with the phenomena of nature. H.4

II

In ordinary language we describe by the word "planning" the complex of interrelated decisions about the allocation of our available resources. All economic activity is in this sense planning; and in any society in which many people collaborate, this planning, whoever does it, will in some measure have to be based on knowledge which, in the first instance, is not given to the planner but to somebody else, which somehow will have to be conveyed to the planner. The various ways in which the knowledge on which people base their plans is communicated to them is the crucial problem for any theory explaining the economic process, and the problem of what is the best way of utilizing knowledge initially dispersed among all the people is at least one of the main problems of economic policy—or of designing an efficient economic system. H.5

The answer to this question is closely connected with that other question which arises here, that of *who* is to do the planning. It is about this question that all the dispute about "economic planning" centers. This is not a dispute about whether planning is to be done or not. It is a dispute as to whether planning is to be done centrally, by one authority for the whole economic system, or is to be divided among many individuals. Planning in the specific sense in which the term is used in contemporary controversy necessarily means central planning—direction of the whole economic system according to one unified plan. Competition, on the other hand, means decentralized planning by many separate persons. The halfway house between the two, about which many people talk but which few like when they see it, is the delegation of planning to organized industries, or, in other words, monopoly. H.6

Which of these systems is likely to be more efficient depends mainly on the question under which of them we can expect that fuller use will be made of the existing knowledge. And this, in turn, depends on whether we are more likely to succeed in putting at the disposal of a single central authority all the knowledge which ought to be used but which is initially dispersed among many different individuals, or in conveying to the individuals such additional knowledge as they need in order to enable them to fit their plans with those of others. H.7

III

It will at once be evident that on this point the position will be different with respect to different kinds of knowledge; and the answer to our question will therefore largely turn on the relative importance of the different kinds of knowledge; those more likely to be at the disposal of particular individuals and those which we should with greater confidence expect to find in the possession of an authority made up of suitably chosen experts. If it is today so widely assumed that the latter will be in a better position, this is because one kind of knowledge, namely, scientific knowledge, occupies now so prominent a place in public imagination that we tend to forget that it is not the only kind that is relevant. It may be admitted that, as far as scientific knowledge is concerned, a body of suitably chosen experts may be in the best position to command all the best knowledge available—though this is of course merely shifting the difficulty to the problem of selecting the experts. What I wish to point out is that, even assuming that this problem can be readily solved, it is only a small part of the wider problem. H.8

Today it is almost heresy to suggest that scientific knowledge is not the sum of all knowledge. But a little reflection will show that there is beyond question a body of very important but unorganized knowledge which cannot possibly be called scientific in the sense of knowledge of general rules: the knowledge of the particular circumstances of time and place. It is with respect to this that practically every individual has some advantage over all others because he possesses unique information of which beneficial use might be made, but of which use can be made only if the decisions depending on it are left to him or are made with his active coöperation. We need to remember only how much we have to learn in any occupation after we have completed our theoretical training, how big a part of our working life we spend learning particular jobs, and how valuable an asset in all walks of life is knowledge of people, of local conditions, and of special circumstances. To know of and put to use a machine not fully employed, or somebody's skill which could be better utilized, or to be aware of a surplus stock which can be drawn upon during an interruption of supplies, is socially quite as useful as the knowledge of better alternative techniques. And the shipper who earns his living from using otherwise empty or half-filled journeys of tramp-steamers, or the estate agent whose whole knowledge is almost exclusively one of temporary opportunities, or the *arbitrageur* who gains from local differences of commodity prices, are all performing eminently useful functions based on special knowledge of circumstances of the fleeting moment not known to others. H.9

It is a curious fact that this sort of knowledge should today be generally regarded with a kind of contempt and that anyone who by such knowledge gains an advantage over somebody better equipped with theoretical or technical knowledge is thought to have acted almost disreputably. To gain an advantage from better knowledge of facilities of communication or transport is sometimes regarded as almost dishonest, although it is quite as important that society make use of the best opportunities in this respect as in using the latest scientific discoveries. This prejudice has in a considerable measure affected the attitude toward commerce in general compared with that toward production. Even economists who regard themselves as definitely immune to the crude materialist fallacies of the past constantly commit the same mistake where activities directed toward the acquisition of such practical knowledge are concerned—apparently because in their scheme of things all such knowledge is supposed to be "given." The common idea now seems to be that all such knowledge should as a matter of course be readily at the command of everybody, and the reproach of irrationality leveled against the existing economic order is frequently based on the fact that it is not so available. This view disregards the fact that the method by which such knowledge can be made as widely available as possible is precisely the problem to which we have to find an answer. H.10

IV

If it is fashionable today to minimize the importance of the knowledge of the particular circumstances of time and place, this is closely connected with the smaller importance which is now attached to change as such. Indeed, there are few points on which the assumptions made (usually only implicitly) by the "planners" differ from those of their opponents as much as with regard to the significance and frequency of changes which will make substantial alterations of production plans necessary. Of course, if detailed economic plans could be laid down for fairly long periods in advance and then closely adhered to, so that no further economic decisions of importance would be required, the task of drawing up a comprehensive plan governing all economic activity would be much less formidable. H.11

It is, perhaps, worth stressing that economic problems arise always and only in consequence of change. So long as things continue as before, or at least as they were expected to, there arise no new problems requiring a decision, no need to form a new plan. The belief that changes, or at least day-to-day adjustments, have become less important in modern times implies the contention that economic problems

also have become less important. This belief in the decreasing importance of change is, for that reason, usually held by the same people who argue that the importance of economic considerations has been driven into the background by the growing importance of technological knowledge. H.12

Is it true that, with the elaborate apparatus of modern production, economic decisions are required only at long intervals, as when a new factory is to be erected or a new process to be introduced? Is it true that, once a plant has been built, the rest is all more or less mechanical, determined by the character of the plant, and leaving little to be changed in adapting to the ever-changing circumstances of the moment? H.13

The fairly widespread belief in the affirmative is not, as far as I can ascertain, borne out by the practical experience of the businessman. In a competitive industry at any rate—and such an industry alone can serve as a test—the task of keeping cost from rising requires constant struggle, absorbing a great part of the energy of the manager. How easy it is for an inefficient manager to dissipate the differentials on which profitability rests, and that it is possible, with the same technical facilities, to produce with a great variety of costs, are among the commonplaces of business experience which do not seem to be equally familiar in the study of the economist. The very strength of the desire, constantly voiced by producers and engineers, to be allowed to proceed untrammeled by considerations of money costs, is eloquent testimony to the extent to which these factors enter into their daily work. H.14

One reason why economists are increasingly apt to forget about the constant small changes which make up the whole economic picture is probably their growing preoccupation with statistical aggregates, which show a very much greater stability than the movements of the detail. The comparative stability of the aggregates cannot, however, be accounted for—as the statisticians occasionally seem to be inclined to do—by the "law of large numbers" or the mutual compensation of random changes. The number of elements with which we have to deal is not large enough for such accidental forces to produce stability. The continuous flow of goods and services is maintained by constant deliberate adjustments, by new dispositions made every day in the light of circumstances not known the day before, by *B* stepping in at once when *A* fails to deliver. Even the large and highly mechanized plant keeps going largely because of an environment upon which it can draw for all sorts of unexpected needs; tiles for its roof, stationery for its forms, and all the thousand and one kinds of equipment in which it cannot be self-contained and which the plans for the operation of the plant require to be readily available in the market. H.15

This is, perhaps, also the point where I should briefly mention the fact that the sort of knowledge with which I have been concerned is knowledge of the kind which by its nature cannot enter into statistics and therefore cannot be conveyed to any central authority in statistical form. The statistics which such a central authority would have to use would have to be arrived at precisely by abstracting from minor differences between the things, by lumping together, as resources of one kind, items which differ as regards location, quality, and other particulars, in a way which may be very significant for the specific decision. It follows from this that central planning based on statistical information by its nature cannot take direct account of these circumstances of time and place and that the central planner will have to find some way or other in which the decisions depending on them can be left to the "man on the spot." H.16

V

If we can agree that the economic problem of society is mainly one of rapid adaptation to changes in the particular circumstances of time and place, it would seem to follow that the ultimate decisions must be left to the people who are familiar with these circumstances, who know directly of the relevant changes and of the resources immediately available to meet them. We cannot expect that this problem will be solved by first communicating all this knowledge to a central board which, after integrating *all*

knowledge, issues its orders. We must solve it by some form of decentralization. But this answers only part of our problem. We need decentralization because only thus can we insure that the knowledge of the particular circumstances of time and place will be promptly used. But the "man on the spot" cannot decide solely on the basis of his limited but intimate knowledge of the facts of his immediate surroundings. There still remains the problem of communicating to him such further information as he needs to fit his decisions into the whole pattern of changes of the larger economic system. H.17

How much knowledge does he need to do so successfully? Which of the events which happen beyond the horizon of his immediate knowledge are of relevance to his immediate decision, and how much of them need he know? H.18

There is hardly anything that happens anywhere in the world that *might* not have an effect on the decision he ought to make. But he need not know of these events as such, nor of *all* their effects. It does not matter for him *why* at the particular moment more screws of one size than of another are wanted, *why* paper bags are more readily available than canvas bags, or *why* skilled labor, or particular machine tools, have for the moment become more difficult to obtain. All that is significant for him is *how much more or less* difficult to procure they have become compared with other things with which he is also concerned, or how much more or less urgently wanted are the alternative things he produces or uses. It is always a question of the relative importance of the particular things with which he is concerned, and the causes which alter their relative importance are of no interest to him beyond the effect on those concrete things of his own environment. H.19

It is in this connection that what I have called the "economic calculus" proper helps us, at least by analogy, to see how this problem can be solved, and in fact is being solved, by the price system. Even the single controlling mind, in possession of all the data for some small, self-contained economic system, would not—every time some small adjustment in the allocation of resources had to be made—go explicitly through all the relations between ends and means which might possibly be affected. It is indeed the great contribution of the pure logic of choice that it has demonstrated conclusively that even such a single mind could solve this kind of problem only by constructing and constantly using rates of equivalence (or "values," or "marginal rates of substitution"), *i.e.*, by attaching to each kind of scarce resource a numerical index which cannot be derived from any property possessed by that particular thing, but which reflects, or in which is condensed, its significance in view of the whole means-end structure. In any small change he will have to consider only these quantitative indices (or "values") in which all the relevant information is concentrated; and, by adjusting the quantities one by one, he can appropriately rearrange his dispositions without having to solve the whole puzzle *ab initio* or without needing at any stage to survey it at once in all its ramifications. H.20

Fundamentally, in a system in which the knowledge of the relevant facts is dispersed among many people, prices can act to coördinate the separate actions of different people in the same way as subjective values help the individual to coördinate the parts of his plan. It is worth contemplating for a moment a very simple and commonplace instance of the action of the price system to see what precisely it accomplishes. Assume that somewhere in the world a new opportunity for the use of some raw material, say, tin, has arisen, or that one of the sources of supply of tin has been eliminated. It does not matter for our purpose—and it is very significant that it does not matter—which of these two causes has made tin more scarce. All that the users of tin need to know is that some of the tin they used to consume is now more profitably employed elsewhere and that, in consequence, they must economize tin. There is no need for the great majority of them even to know where the more urgent need has arisen, or in favor of what other needs they ought to husband the supply. If only some of them know directly of the new demand, and switch resources over to it, and if the people who are aware of the new gap thus created in turn fill it from still other sources, the effect will rapidly spread throughout the whole economic system and influence not

only all the uses of tin but also those of its substitutes and the substitutes of these substitutes, the supply of all the things made of tin, and their substitutes, and so on; and all his without the great majority of those instrumental in bringing about these substitutions knowing anything at all about the original cause of these changes. The whole acts as one market, not because any of its members survey the whole field, but because their limited individual fields of vision sufficiently overlap so that through many intermediaries the relevant information is communicated to all. The mere fact that there is one price for any commodity—or rather that local prices are connected in a manner determined by the cost of transport, etc.—brings about the solution which (it is just conceptually possible) might have been arrived at by one single mind possessing all the information which is in fact dispersed among all the people involved in the process. H.21

VI

We must look at the price system as such a mechanism for communicating information if we want to understand its real function—a function which, of course, it fulfils less perfectly as prices grow more rigid. (Even when quoted prices have become quite rigid, however, the forces which would operate through changes in price still operate to a considerable extent through changes in the other terms of the contract.) The most significant fact about this system is the economy of knowledge with which it operates, or how little the individual participants need to know in order to be able to take the right action. In abbreviated form, by a kind of symbol, only the most essential information is passed on and passed on only to those concerned. It is more than a metaphor to describe the price system as a kind of machinery for registering change, or a system of telecommunications which enables individual producers to watch merely the movement of a few pointers, as an engineer might watch the hands of a few dials, in order to adjust their activities to changes of which they may never know more than is reflected in the price movement. H.22

Of course, these adjustments are probably never "perfect" in the sense in which the economist conceives of them in his equilibrium analysis. But I fear that our theoretical habits of approaching the problem with the assumption of more or less perfect knowledge on the part of almost everyone has made us somewhat blind to the true function of the price mechanism and led us to apply rather misleading standards in judging its efficiency. The marvel is that in a case like that of a scarcity of one raw material, without an order being issued, without more than perhaps a handful of people knowing the cause, tens of thousands of people whose identity could not be ascertained by months of investigation, are made to use the material or its products more sparingly; *i.e.*, they move in the right direction. This is enough of a marvel even if, in a constantly changing world, not all will hit it off so perfectly that their profit rates will always be maintained at the same constant or "normal" level. H.23

I have deliberately used the word "marvel" to shock the reader out of the complacency with which we often take the working of this mechanism for granted. I am convinced that if it were the result of deliberate human design, and if the people guided by the price changes understood that their decisions have significance far beyond their immediate aim, this mechanism would have been acclaimed as one of the greatest triumphs of the human mind. Its misfortune is the double one that it is not the product of human design and that the people guided by it usually do not know why they are made to do what they do. But those who clamor for "conscious direction"—and who cannot believe that anything which has evolved without design (and even without our understanding it) should solve problems which we should not be able to solve consciously—should remember this: The problem is precisely how to extend the span of out utilization of resources beyond the span of the control of any one mind; and therefore, how to dispense with the need of conscious control, and how to provide inducements which will make the individuals do the desirable things without anyone having to tell them what to do. H.24

The problem which we meet here is by no means peculiar to economics but arises in connection with nearly all truly social phenomena, with language and with most of our cultural inheritance, and constitutes really the central theoretical problem of all social science. As Alfred Whitehead has said in another connection, "It is a profoundly erroneous truism, repeated by all copy-books and by eminent people when they are making speeches, that we should cultivate the habit of thinking what we are doing. The precise opposite is the case. Civilization advances by extending the number of important operations which we can perform without thinking about them." This is of profound significance in the social field. We make constant use of formulas, symbols, and rules whose meaning we do not understand and through the use of which we avail ourselves of the assistance of knowledge which individually we do not possess. We have developed these practices and institutions by building upon habits and institutions which have proved successful in their own sphere and which have in turn become the foundation of the civilization we have built up. H.25

The price system is just one of those formations which man has learned to use (though he is still very far from having learned to make the best use of it) after he had stumbled upon it without understanding it. Through it not only a division of labor but also a coördinated utilization of resources based on an equally divided knowledge has become possible. The people who like to deride any suggestion that this may be so usually distort the argument by insinuating that it asserts that by some miracle just that sort of system has spontaneously grown up which is best suited to modern civilization. It is the other way round: man has been able to develop that division of labor on which our civilization is based because he happened to stumble upon a method which made it possible. Had he not done so, he might still have developed some other, altogether different, type of civilization, something like the "state" of the termite ants, or some other altogether unimaginable type. All that we can say is that nobody has yet succeeded in designing an alternative system in which certain features of the existing one can be preserved which are dear even to those who most violently assail it—such as particularly the extent to which the individual can choose his pursuits and consequently freely use his own knowledge and skill. H.26

VII

It is in many ways fortunate that the dispute about the indispensability of the price system for any rational calculation in a complex society is now no longer conducted entirely between camps holding different political views. The thesis that without the price system we could not preserve a society based on such extensive division of labor as ours was greeted with a howl of derision when it was first advanced by von Mises twenty-five years ago. Today the difficulties which some still find in accepting it are no longer mainly political, and this makes for an atmosphere much more conducive to reasonable discussion. When we find Leon Trotsky arguing that "economic accounting is unthinkable without market relations"; when Professor Oscar Lange promises Professor von Mises a statue in the marble halls of the future Central Planning Board; and when Professor Abba P. Lerner rediscovers Adam Smith and emphasizes that the essential utility of the price system consists in inducing the individual, while seeking his own interest, to do what is in the general interest, the differences can indeed no longer be ascribed to political prejudice. The remaining dissent seems clearly to be due to purely intellectual, and more particularly methodological, differences. H.27

A recent statement by Professor Joseph Schumpeter in his *Capitalism, Socialism, and Democracy* provides a clear illustration of one of the methodological differences which I have in mind. Its author is pre-eminent among those economists who approach economic phenomena in the light of a certain branch of positivism. To him these phenomena accordingly appear as objectively given quantities of commodities impinging directly upon each other, almost, it would seem, without any intervention of human minds. Only against this background can I account for the following (to me startling) pronouncement. Professor Schumpeter argues that the possibility of a rational calculation in the absence of markets for the factors of

production follows for the theorist "from the elementary proposition that consumers in evaluating ('demanding') consumers' goods *ipso facto* also evaluate the means of production which enter into the production of these goods."*1 H.28

Taken literally, this statement is simply untrue. The consumers do nothing of the kind. What Professor Schumpeter's *"ipso facto"* presumably means is that the valuation of the factors of production is implied in, or follows necessarily from, the valuation of consumers' goods. But this, too, is not correct. Implication is a logical relationship which can be meaningfully asserted only of propositions simultaneously present to one and the same mind. It is evident, however, that the values of the factors of production do not depend solely on the valuation of the consumers' goods but also on the conditions of supply of the various factors of production. Only to a mind to which all these facts were simultaneously known would the answer necessarily follow from the facts given to it. The practical problem, however, arises precisely because these facts are never so given to a single mind, and because, in consequence, it is necessary that in the solution of the problem knowledge should be used that is dispersed among many people. H.29

The problem is thus in no way solved if we can show that all the facts, *if* they were known to a single mind (as we hypothetically assume them to be given to the observing economist), would uniquely determine the solution; instead we must show how a solution is produced by the interactions of people each of whom possesses only partial knowledge. To assume all the knowledge to be given to a single mind in the same manner in which we assume it to be given to us as the explaining economists is to assume the problem away and to disregard everything that is important and significant in the real world. H.30

That an economist of Professor Schumpeter's standing should thus have fallen into a trap which the ambiguity of the term "datum" sets to the unwary can hardly be explained as a simple error. It suggests rather that there is something fundamentally wrong with an approach which habitually disregards an essential part of the phenomena with which we have to deal: the unavoidable imperfection of man's knowledge and the consequent need for a process by which knowledge is constantly communicated and acquired. Any approach, such as that of much of mathematical economics with its simultaneous equations, which in effect starts from the assumption that people's *knowledge* corresponds with the objective *facts* of the situation, systematically leaves out what is our main task to explain. I am far from denying that in our system equilibrium analysis has a useful function to perform. But when it comes to the point where it misleads some of our leading thinkers into believing that the situation which it describes has direct relevance to the solution of practical problems, it is high time that we remember that it does not deal with the social process at all and that it is no more than a useful preliminary to the study of the main problem.

H.31

1. J. Schumpeter, *Capitalism, Socialism, and Democracy* (New York; Harper, 1942), p. 175. Professor Schumpeter is, I believe, also the original author of the myth that Pareto and Barone have "solved" the problem of socialist calculation. What they, and many others, did was merely to state the conditions which a rational allocation of resources would have to satisfy and to point out that these were essentially the same as the conditions of equilibrium of a competitive market. This is something altogether different from knowing how the allocation of resources satisfying these conditions can be found in practice. Pareto himself (from whom Barone has taken practically everything he has to say), far from claiming to have solved the practical problem, in fact explicitly denies that it can be solved without the help of the market. See his *Manuel d'économie pure* (2d ed., 1927), pp. 233-34. The relevant passage is quoted in an English translation at the beginning of my article on "Socialist Calculation: The Competitive 'Solution,' " in *Economica,* New Series, Vol. VIII, No. 26 (May, 1940), p. 125.

American Economic Review, XXXV, No. 4; September, 1945, pp. 519-30.

Retour page d'accueil

© Institut Hayek

Annu. Rev. Law Soc. Sci. 2005. 1:369–96
doi: 10.1146/annurev.lawsocsci.1.031805.111122
Copyright © 2005 by Annual Reviews. All rights reserved
First published online as a Review in Advance on August 19, 2005

WHY LAW, ECONOMICS, AND ORGANIZATION?

Oliver E. Williamson
Walter A. Haas School of Business, University of California, Berkeley, California 94720; email: owilliam@haas.berkeley.edu

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

Key Words governance, transaction costs, law school teaching

■ **Abstract** This review shows that a combined law, economics, and organization theory approach leads to different and deeper understandings of the purposes served by complex contract and economic organization. The business firm for these purposes is described not in technological terms (as a production function) but in organizational terms (as an alternative mode of governance). Firm and market are thus examined comparatively with respect to their capacities to organize transactions, which differ in their complexity, so as to economize on transaction costs. The predictive theory of economic organization that results has numerous ramifications for public policy toward business and for teaching and research in the law schools.

INTRODUCTION

Whereas law and economics began as the application of economic reasoning to antitrust and regulation, it has since been expanded to bring economic analysis to bear (in varying degree) on every facet of the law school curriculum. Occasional dissents notwithstanding, law and economics is widely regarded as a success story.

I concur with this favorable assessment but would observe that economic analysis comes in more than one flavor. As between the two main branches—the science of choice and the science of contract—law and economics scholarship mainly works out of the science of choice tradition. All well and good for many purposes but not, I contend, for all. Specifically, those parts of the law and economics enterprise that are centrally concerned with issues of economic organization ought to be informed, additionally or instead, by the science of contract perspective.[1]

This involves, among other things, supplanting the neoclassical theory of the firm-as-production function (which is a technological construction) with the theory of the firm-as-governance structure (which is an organizational construction). A critical concession, which many law and economics scholars are loathe to make,

[1] Areas of the law that are most in need of a more veridical theory of economic organization include antitrust, regulation, corporations, labor law, corporate governance, agency, administrative law, property, contract, secured transactions, and torts. But the science of contract branch of economics has ramifications to the law quite generally.

is that the orthodox theory of the firm was never designed with reference to (and, hence, is often poorly suited to interpret) nonstandard and unfamiliar contractual practices and organizational structures.

I begin with a brief discussion of the sciences of choice and of contract and of the differing needs of each for a theory of the firm. I then turn in the next section to what I regard as the chief lessons of organization theory for a theory of the firm-as-governance structure. The comparative contractual approach to economic organization, of which the theory of the firm-as-governance structure is a part, is then sketched in the section on Comparative Contractual Analysis. Applications to public policy analysis are set out in the next section, and the lessons of the comparative contractual approach to economic organization for the teaching of contract law are developed in the final section. Concluding remarks follow.

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

CONCEPTUAL FRAMEWORKS

Choice and Contract

Although orthodox economic theory, with its emphasis on scarcity and efficient resource allocation, is widely regarded as an all-purpose theory, it is more properly regarded as the "dominant paradigm" (Reder 1999, p. 43). Plainly, dominant paradigms command more respect. Often, however, their uses are much more apt for some types of problems than they are for others.

Lionel Robbins captured the emerging consensus of what economics was all about in his description of economics as "the science which studies human behavior as a relation between ends and scarce means which have alternative uses" (Robbins 1932, p. 16)—or as Reder puts it, economics deals with "the allocation of scarce resources among alternative uses for the maximization of want satisfactions" (Reder 1999, p. 43). The theory of consumer behavior and the theory of the firm are the two key building blocks upon which this science rests: The consumer seeks to maximize utility subject to a budget constraint, and the firm is a production function that transforms inputs into outputs, with efficiency realized through the choice of optimal factor proportions. All well and good for the study of supply and demand, prices and output. The economist working out of such a setup decidedly does not, however, address himself to issues of firm and market organization except in narrowly delimited ways.[2] The firm, for all intents and purposes, is a "black box."

James Buchanan has declared this science of choice perspective as a "wrong turn" (Buchanan 1975, p. 225), but I put it somewhat differently. Economics

[2] As Ronald Coase has put it, in the Robbins conception of economics, the economist "does not interest himself in the internal arrangements within organizations but only in what happens on the market" (Coase 1992, p. 714).

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

became unduly preoccupied with the science of choice to the neglect of the science of contract. Rather than deal with contract and exchange, economics became the science of constrained optimization.

As perceived by Buchanan, the principal needs for a science of contract are found in the field of public finance and take the form of social ordering: "Politics is a structure of complex exchange among individuals, a structure within which persons seek to secure *collectively* their own privately defined objectives that cannot be efficiently secured through simple market exchanges" (Buchanan 1987, p. 296; emphasis added). By contrast, I see the needs for a science of contract primarily with reference to the field of industrial organization and in the context of private ordering.

Compared with the politics of collective action, private ordering is accomplished through the individual efforts of the immediate parties to an exchange. Out of awareness of the limitations of spot-market contracting and the impossibility of comprehensive contracting, the immediate parties to an exchange craft governance structures that permit them to realize mutual gains.

The role of the courts, for such a purpose, is very different from that projected under the science of choice perspective.

Firms

"Any standard economic theory, not just neoclassical, starts with the existence of firms. Usually, the firm is a point or at any rate a black box. . . . But firms are not points. They have internal structure. This internal structure must arise for some reason" (Arrow 1999, p. vii). The contrast between the science of choice and the science of contract in this respect is fundamental. As Harold Demsetz has put it, "It is a mistake to confuse the firm of economic theory with its real-world namesake. The chief mission of neoclassical economics is to understand how the price system coordinates the use of resources, not to understand the inner workings of real firms" (Demsetz 1983, p. 377). By contrast, the science of contract is expressly concerned with the attributes of firms, especially in relation to the attributes of alternative modes of governance, as these bear on the management of transactions. As against a technological view of the firm, the firm (and other modes of governance) are described as governance structures. John R. Commons's prescient conception of economics is broadly congruent with the science of contract perspective: "[T]he ultimate unit of activity. . . must contain in itself the three principles of conflict, mutuality, and order. This unit is a transaction" (Commons 1932, p. 4). Not only does transaction cost economics (TCE) concur that the transaction is the basic unit of analysis, but it views governance as the means by which to infuse order, thereby mitigating conflict and realizing mutual gains.

As developed below, implementing the private ordering branch of the science of contract is a much more microanalytic project than is the science of choice approach to economics. Hitherto neglected attributes of both transactions and governance structures now need to be uncovered and their ramifications worked

out. In the process, a whole series of public policy differences between the choice and contract perspectives emerge.

ORGANIZATION THEORY

Organization theory is a vast field to which sociology, psychology (cognitive, social, evolutionary), aspects of political science, economics, and cultural anthropology all relate. W. Richard Scott's (1987) influential text distinguishes three main branches: rational, natural, and open systems—where the rational systems approach places primary emphasis on formal structure, the natural systems approach features informal organization, and the open systems approach examines shifting coalitions within the organization and in relation to their environment. All have a role to play in understanding complex organization. Of these three, I place primary emphasis on the contributions of rational systems theory, although provision is also made for the spontaneous forces of informal organization and the intertemporal transformations that relate thereto. Chester Barnard [1962 (1938)], Herbert Simon (1947), and March & Simon (1958) are especially prominent to the rational systems tradition. Of the many contributions that originate with this tradition, the five that I regard as most relevant to the science of contract approach to economic organization are (*a*) human actors, (*b*) adaptation, (*c*) intertemporal transformations, (*d*) choice of the unit of analysis, and (*e*) discrete structural features.[3]

If, as I contend, organization theory is important to the study of economic organization in these five and other respects, the puzzle is why organization theory has not been more fully incorporated within economics. The chief reasons, I think, are these: (*a*) organization theory has less relevance to the science of choice than to the science of contract, and most economists have been content to work out of the "dominant paradigm"; (*b*) organization theorists mainly deliver a negative message (the science of choice is wrongheaded) rather than relate to the opportunities opened up by the incipient science of contract; and (*c*) leaders of the law and economics movement, such as Richard Posner, who were votaries of orthodoxy, were dismissive of organization theory: "[O]rganization theory. . . [adds] nothing to. . .economics that the literature on information costs had not added much earlier" (Posner 1993, p. 84).

Be that as it may, my sense is that a theory of economic organization that aspires to deal with real firms and, more generally, with economic organization in an uncontrived way cannot ignore or dismiss the contributions of organization theory named above—which are not the main issues with which the "literature on information costs" has been concerned (even now, to say nothing of "much earlier").

[3]Other important contributions include (*f*) weak form selection, (*g*) informal organization, (*h*) cognitive specialization, and (*i*) bureaucracy.

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

Human Actors

Simon advised social scientists that "[n]othing is more fundamental in setting our research agenda and informing our research methods than our view of the nature of the human beings whose behavior we are studying" (Simon 1985, p. 303). The two attributes of human actors that are especially relevant to the economics of governance are cognition and self-interestedness.

Simon took early exception with the idea that human actors are supremely rational, and he proposed instead that human actors be described as boundedly rational, by which he meant that they are "intendedly rational, but only limitedly so" (Simon 1957a, p. xxiv). Human actors are thus neither nonrational nor irrational but are attempting effectively to cope.

TCE agrees that scholars' view of the human beings whose behavior they are studying has profound ramifications for the research agenda. It also concurs that human actors are subject to bounded rationality. Rather than dwell on the lessons of bounded rationality for the science of choice (where the use of maximizing apparatus was contested), however, TCE turns to the science of contract and takes the chief lesson of bounded rationality for the study of economic organization to be that all complex contracts are unavoidably incomplete.

Contractual incompleteness by itself, however, does not a serious problem of contracting make. Governance problems are posed when incomplete contracts (to include unforeseen contingencies) are combined with opportunism. The conflicts to which Commons referred now appear, especially during contract execution and at the contract renewal interval.

Note that TCE does not dispute that most people will do what they say (and some will do more) without self-consciously asking whether the effort is justified by expected discounted net gains. But while accurate descriptions of what is going on "most of the time" are important, much of what is interesting about human behavior in general and contract in particular has reference not to routines but to exceptions. Faced with unanticipated disturbances for which an incomplete contract makes inadequate or incorrect provision (by reason of gaps, errors, and omissions), such disturbances will push the parties to an incomplete contract off of the contract curve. Strategic considerations now come into play if, rather than frailty of motive, opportunism is the operative condition.[4] Contractual breakdowns by reason of defection from the spirit of cooperation and reliance on the letter of the contract are now in prospect.

Inefficiencies of all kinds nevertheless invite relief. Out of awareness of prospective hazards, parties to a contract have incentives to craft ex ante safeguards in a

[4]Interestingly, opportunism makes an appearance in the natural system treatment of sociologists. As Scott puts it, "there is frequently a disparity between. . .the professed or official goals that are announced and the actual or operative goals that can be observed to govern the activities of participants" (Scott 1987, p. 52). Whereas rational system theorists emphasize the normative structure of ex ante decisions, natural system theorists stress the behavioral structure of ex post outcomes (Scott 1987, p. 53).

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

cost-effective degree. Rather than postulate either myopia or omniscience, TCE assumes that human actors have the capacity for "feasible foresight," which is a rational spirit construction. George Schultz speaks to the point as follows: "my training in economics has had a major influence on the way I think about public policy tasks, even when they have no particular relationship to economics. Our discipline makes one think ahead, ask about indirect consequences, take note of variables that may not be directly under consideration" (Schultz 1995, p. 1). But economists do not have a lock on this. As the evolutionary biologist Richard Dawkins (1976) observes, the "capacity to simulate the future in imagination. . . [saves] us from the worst consequences of the blind replicators" (p. 200). Practitioners, consultants, and public policy analysts who possess the skills for and practice the art of feasible foresight will look ahead, discern potential hazards, and fold these into the ex ante design.

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

Adaptation

Interestingly, both the economist Hayek (1945) and the organization theorist Barnard [1962 (1938)] are in agreement that adaptation is the central problem of economic organization. The adaptations to which they have reference, however, differ. Hayek had reference to the adaptations of autonomous economic actors who adjust spontaneously to changes in the market (mainly as signaled by changes in relative prices). By contrast, Barnard appealed to intentionality. He featured cooperative adaptation made by economic actors with the assistance of hierarchy within firms. Although adaptation of each type is important and can be studied separately, TCE is interested in markets and hierarchies (rather than markets alone, or hierarchies alone). TCE therefore deals with adaptations of both kinds (and mixtures thereof). Specifically, TCE holds that choice of contractual mode should be derived by recognizing that the adaptive needs of transactions (in autonomous and cooperative respects) vary with the attributes of transactions and that the adaptive capacities of alternative modes of governance also differ. The upshot is that efficiency gains are realized by aligning transactions with governance structures so as to effect an economizing outcome. Pushing the logic of autonomous and cooperative adaptation to completion thereby leads to a predictive theory of comparative economic organization (Williamson 1991).

Intertemporal Transformations

That internal organization has a life of its own has been evident to sociologists of organization for a long time. There is more to it, moreover, than simply being alerted to hitherto neglected regularities. Once disclosed, the ex ante organizational design ramifications of these regularities need to be worked out.

Robert Michels' 1911 book on *Political Parties* focused on the intertemporal transformations that regularly attended democratic efforts at political organization. The most important such intertemporal transformation is summarized by the

famous Iron Law of Oligarchy: "It is [hierarchical] organization which gives birth to the dominion of the elected over the electors, of the mandatories over the mandators, of the delegates over the delegators. Who says organization, say oligarchy" (Michels 1962, p. 365). Michels traced the source of these oligarchical tendencies to "the nature of the human individual,. . . the nature of the political struggle,. . . and the nature of organization" (p. 6).

Michels, moreover, had a very farsighted view of his findings: "The sociologist should aim. . . at the dispassionate exposition of tendencies and counter-operating forces, of reasons and opposing reasons, at the display, in a word, of the warp and the woof of social life" (Michels 1962, p. 6). Unless we are alert to the intertemporal propensities of organization, we will be needlessly victimized by them: "[N]othing but a serene and frank examination of the oligarchical dangers of democracy will enable us to minimize these dangers" (Michels 1962, p. 370). Thus, although the oligarchical propensities of democratic organization may have been poorly understood by academics and some practitioners until Michels clarified the issue, the lurking hazards of oligarchy should no longer come as a surprise. Today's organizational designers presumably take the Iron Law of Oligarchy into account in the initial design calculus.

Selznick characterized "Michels' theory about democratic organization. . . as a *special case* of the general recalcitrance of the human tools of action. The tendency for goals to be subverted through the creation of new centers of interest and motivation *inheres in all organizations*" (Selznick 1950, p. 162; emphasis added). The study of unanticipated consequences of all kinds—of which oligarchy is but one example—thus describes the larger research agenda.

Akin to the discussion of feasible foresight in the section on Human Actors, above, TCE responds in a three-part way. First, be alert to all the significant, unanticipated consequences and bureaucratic propensities that students of internal organization uncover. Second, take the logic to completion. For each unanticipated effect, ask from where it arises, what are the mechanisms through which it operates, what are the effects on contract and organization, and what are the ramifications for ex ante design (thereby mitigating unwanted consequences and enhancing beneficial effects). Third, upon taking a farsighted view of contract and organization, do not rely entirely on the reports by organization theorists of unanticipated consequences. Given contractual incompleteness (by reason of bounded rationality) and the possibility of defection from agreements (by reason of opportunism), practitioners of TCE look ahead to ascertain whether and when predictable contractual hazards will accrue. If and as such hazards can be projected, the governance ramifications need to be worked out. [An illustration is the Fundamental Transformation, by which a large numbers bidding competition is (sometimes) transformed into a small numbers supply relation during contract execution and at the contract renewal interval (Williamson 1985, pp. 61–63). As developed in the section on Applications to Public Policy, below, contractual safeguards and (possibly) vertical integration arise to mitigate such hazards.]

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

Unit of Analysis

TCE adopts the purposive perspective of John R. Commons by naming the transaction as the unit of analysis. But that is merely the first step. Naming a unit of analysis needs to be followed by providing operational content. The proponents of many would-be units of analysis never undertake this second step or founder upon reaching it.

Identifying the critical dimensions with respect to which transactions differ is facilitated by asking which attributes, among the countless ways in which transactions differ, have consequential transaction cost effects. Some transactions are simple while others are complex. What are the distinguishing features? Older style institutional economics never asked, hence never answered, this question.

The obvious place to begin is with the ideal transaction in law and economics—namely, contracts that take place between faceless economic actors, where continuity is unimportant because the identity of the parties does not matter. Then ask the question, "What attributes of transactions are responsible for the breakdown of this contractual ideal?" Relevant attributes for describing transactions between parties where identity does matter include asset specificity in its various forms (which gives rise to bilateral dependency), uncertainty (for which consciously coordinated adaptations to disturbances may be needed), and frequency (which has a bearing on the future value of preserving a continuing relation and on the incentive to incur the cost of specialized governance).

Discrete Structural

If alternative modes of organization differ in discrete structural ways, then marginal analysis can be supplanted by discrete structural analysis, which is purportedly easier to implement (Simon 1978, pp. 6–7).[5] As a comparative contractual matter, however, the real import of the proposition that moving from one generic form of organization to another is attended by discontinuities is that alternative modes of governance have different strengths and weaknesses by reason of these discontinuities. As with the transaction, moreover, there is a need to go beyond this first step to ascertain the critical attributes with respect to which governance structures differ. The question to be asked and answered here is this: How do alternative modes of governance differ in contract implementation and enforcement respects?

One device for getting at this is to pose the puzzle of selective intervention: Can a firm replicate the market mode for all state realizations for which market procurement works well and intervene always but only when expected net gains can

[5]Because marginal analysis is actually easy to implement, economists can be thought of as analytical satisficers: They use workable apparatus that (often) is "good enough." Also note that the use of marginal analysis and discrete structural analysis can be joined, as in Riordan & Williamson (1985), where discrete structural differences give rise to first-order effects and marginal analysis introduces second-order refinements.

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

be projected. If feasible, then large firms will always do as well as a collection of small firms (through replication) and will sometimes do better (by selective intervention). As I have developed elsewhere (Williamson 1985, chapter 6), such efforts are not only impossible but, if attempted, are attended by a series of unwanted effects. This is because efforts to preserve the high-powered incentives of markets within hierarchies give rise to asset dissipation losses and strategic distortions. The upshot is that the move from market to hierarchy is attended by a weakening of incentive intensity and, as a consequence, by an increase in administrative oversight and control.

A third discrete structural difference arises in contract law respects. The idea that each generic mode of governance is supported by a distinctive form of contract law can be traced to Karl Llewellyn's (1931) early distinction between contract as framework and contract as legal rules; to Ian Macneil's (1974) further distinctions among classical, neoclassical, and relational contract laws; to later treatments of private ordering (Galanter 1981, Klein & Leffler 1981); and to credible contracting (Williamson 1983, Gilson 1984).

Classical contract law of a legal rules kind applies to the ideal transaction in both law and economics, where large numbers of informed and "faceless buyers and sellers. . . [meet] for an instant to exchange standardized goods at equilibrium prices" (Ben-Porath 1980, p. 4). Such a legal rules regime gives way to contract as framework when long-term contracting with dependency relations sets in. The parties here have an interest in promoting continuity in the face of unforeseen disturbances, and hence move to a more cooperative and adaptable contracting form. Such neoclassical contracts are not, however, indefinitely elastic. When push comes to shove, the letter of the contract becomes the basis for "ultimate appeal" to the courts (Llewellyn 1931, p. 737)—wherein the written contract serves to delimit threat positions.

What then is the contract law of internal organization? As developed elsewhere (Williamson 1991), the implicit law of internal organization is that of forbearance. Thus, whereas courts routinely grant standing to interfirm disputes over prices, damages ascribed to delays, failures of quality, and the like, courts will refuse to hear disputes between one internal division and another over identical technical issues. If access to the courts is denied, hierarchy is its own court of ultimate appeal, whereupon firms have access to fiat that interfirm contracting does not.

Taken together, the lessons of organization theory for the science of contract (private ordering branch) are these:

1. All complex contracts are unavoidably incomplete (by reason of bounded rationality), and hence comprehensive contingent claims contracting is infeasible and once-and-for-all auctions (competition for the market) are often fraught with hazards.
2. Farsighted players to an incomplete contract have the incentive to look ahead, identify potential hazards, and attempt to provide ex ante relief for these hazards through the judicious choice of governance.

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

3. Adaptation is the central problem of economic organization, and autonomous and cooperative types of adaptation need to be distinguished and, as appropriate, provided for.
4. Because organizations have a life of their own, all significant intertemporal regularities need to be uncovered and the ramifications for economic organization worked out.
5. The key attributes of the transaction (which is taken to be the basic unit of analysis for the science of contract) need to be named and their ramifications worked out.
6. Because alternative modes of governance differ in discrete structural ways, the syndrome of attributes that defines each mode needs to be named and the comparative strengths and weaknesses of each generic form worked out.

The upshot is that, upon moving from the science of choice to the science of contract perspective, the contributions of organization theory for the study of economic organization come to life. The burgeoning study of the economics of organization thus holds that organizations matter (in the above-described way, as well as others) and that organizations are susceptible to analysis (especially when viewed through a comparative contractual lens in which economizing on transaction costs is featured).

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

COMPARATIVE CONTRACTUAL ANALYSIS

Discriminating Alignment

The discriminating alignment hypothesis out of which TCE works holds that transactions, which differ in their attributes, are aligned with governance structures, which differ in their costs and competence, so as to effect a (mainly) transaction cost economizing result. As indicated above, this requires that the attributes of both transactions and governance structures be identified and the relations between them worked out.

Going beyond the proposition that the transaction is the basic unit of analysis, TCE takes the next step and names asset specificity (in its various forms), uncertainty, and frequency as key attributes. Of these three, asset specificity is the most important and distinctive to the TCE enterprise.

As developed elsewhere, asset specificity is a measure of the degree to which the assets needed to produce a good or service can be redeployed to alternative uses and users without loss of productive value. Whereas identity is unimportant for generic goods and services, the identity of the immediate parties to an exchange are critical as asset specificity (of physical, human, site, dedicated, brand name, or temporal kinds) builds up. In that event, a bilateral dependency condition sets in and the parties are subject to opportunistic defection from the spirit of a contract

to insist on the letter where large gains are at stake. Maladaptation costs attended by costly bargaining are the result.

Put differently, contractual hazards arise when incomplete contracts that are supported by nontrivial investments in specific assets are beset by disturbances (uncertainty). Out of awareness of these hazards, parties to such contracts have incentives to take hazard-mitigating actions, such as by devising safeguards that serve to infuse order and thereby reduce conflict and realize mutual gains.

As discussed above in conjunction with discrete structural analysis, alternative modes of governance are defined as internally consistent syndromes with respect to the following attributes: incentive intensity, administrative controls, and contract law regimes. Because different modes of governance combine these attributes differently, alternative modes differ in their capacities to implement autonomous and cooperative adaptations. The details are developed elsewhere (Williamson 1991). By way of summary, the discrete structural differences by which firm and market are distinguished are

1. incentive intensity: the high-powered incentives of markets give way to low-powered incentives in firms;
2. administrative controls: compared with markets, firms are supported by a more extensive array of administrative rules and procedures, including accounting and auditing, as well as the supports of informal organization;
3. contract law: the contract law of markets is legalistic and relies on court ordering, whereas, as described above, the contract law of internal organization is that of forbearance.

Because of these differences, markets enjoy the advantage in effecting autonomous adaptations, whereas the advantage accrues to firms in effecting cooperative adaptations.

The Simple Contractual Schema

Upon adopting a comparative contractual approach to economic organization in which (*a*) the transaction is made the basic unit of analysis, (*b*) alternative modes of organization are described as governance structures to which discrete structural differences accrue, and (*c*) economizing on transaction costs is taken to be the main case, a very different concept of the firm and of the purposes served by nonstandard and unfamiliar contractual practices and organizational structures results. Note that the firm, in this scheme of things, is not a stand-alone concept but is examined in relation to alternative modes of governance. Always and everywhere the action resides in the microanalytics of transactions and governance structures.

Thus, assume that a firm can make or buy a component and assume further that the component can be supplied by either of two technologies. One is a general-purpose technology and the other a special-purpose technology. The special-purpose technology requires greater investment in transaction-specific durable assets and is more efficient for servicing steady-state demands. Steady-state,

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

however, is an analytical convenience: Most contracts are implemented under conditions of uncertainty for which adaptation to disturbances is needed. Because an incomplete contract between bilaterally dependent parties (that is, those for which continuity has value) is often silent on or makes incorrect or inadequate provision for some of these adaptations, contractual conflicts prospectively arise. Thus although mutual gains will always be realized upon costlessly restoring a position on the contract curve, each party may posture and make opportunistic representations over the division of gains. Costly delays and imperfect adaptations result.

Using h as a measure of contractual hazards, the transactions in Figure 1 that use the general-purpose technology are ones for which $h = 0$. Autonomous adaptation in a competitive market suffices because the parties are faceless. If instead transactions use the special-purpose technology, an $h > 0$ condition exists. Assets here are specialized, whence productive values would be sacrificed if $h > 0$ transactions were to be prematurely terminated. Such bilaterally dependent parties have incentives to promote continuity and safeguard investments. Cooperative adaptation thus comes to the fore.

Let s denote the magnitude of any such safeguards. An $s = 0$ condition is one in which no safeguards are provided; a decision to provide safeguards is reflected by an $s > 0$ result.

Safeguards can take either of two forms. One form is to provide interfirm contracts with added support: Penalties to deter breach are introduced, added information disclosure is provided, and specialized dispute settlement machinery (e.g., arbitration) is devised. This safeguard is the credible interfirm commitment option. A second form is to take transactions out of markets and organize them under unified ownership where hierarchy (to include fiat) is used to effect coordination.

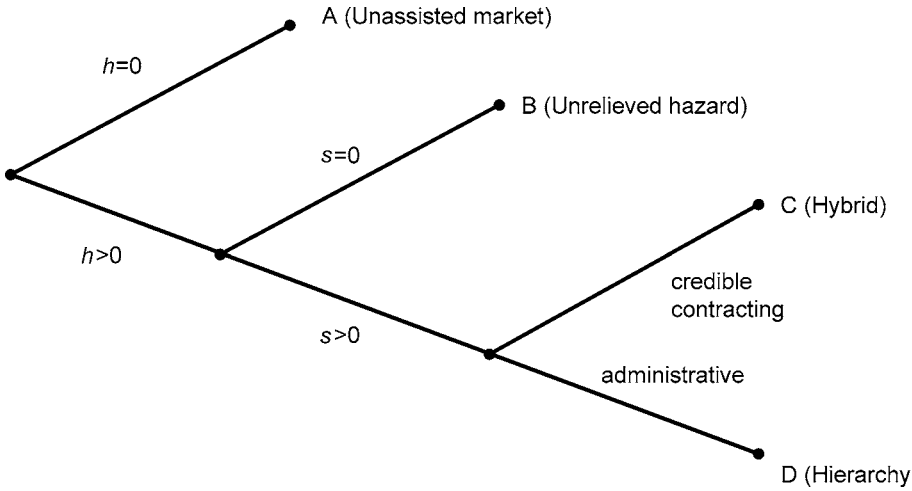


Figure 1 The simple contractual schema.

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

Node A corresponds to the ideal transaction in law and economics: With an absence of dependency ($h = 0$), prices are set competitively in the market (by supply and demand), and, in the event of contractual breakdown, the courts award damages. Node B poses unrelieved contractual hazards in that specialized investments are exposed ($h > 0$) for which no safeguards ($s = 0$) have been provided. Such hazards will be recognized by farsighted players, who will price out the implied risks. Nodes C and D are those for which additional contractual support has been provided ($s > 0$), either in the form of contractual safeguards (node C) or unified ownership (node D).

In the event that costly breakdowns continue in the face of best bilateral efforts to craft safeguards at node C, the transaction may be taken out of the market and organized under unified ownership (vertical integration) instead. Inasmuch, however, as added bureaucratic costs accrue upon taking a transaction out of the market and organizing it internally, internal organization is usefully thought of as the organization form of last resort: Try markets, try hybrids, and have recourse to the firm only when all else fails. Node D, the firm, thus comes in only as higher degrees of asset specificity and added uncertainty pose greater needs for cooperative adaptation.

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

APPLICATIONS TO PUBLIC POLICY

Node A excepted, which is the ideal transaction in law and economics to which I referred previously, the neoclassical and transaction cost approaches to firm and market organization plainly differ. These differences are due to the broader conception of economic organization out of which TCE works (where alternative modes of organization are described as governance structures, to which the lessons of organization theory apply), and these differences have ramifications for public policy toward business. Neoclassical and transaction cost interpretations of nonstandard and unfamiliar contracting practices and organizational structures are compared and contrasted here. The overarching difference is this: Orthodox economics is more imperial in that it imposes a price theoretic interpretation on the phenomena in question, whereas TCE is more curious and asks the question "What's going on here?" The TCE action is in the details of transactions on the one hand and governance structures on the other, which is closer in spirit to organization theory.

Vertical Integration/Vertical Market Restraints

Orthodox explanations for integration (backward, forward, or lateral) of the firm-as-production function kind invoke considerations of technology, inefficient factor proportions that result from double-marginalization (McKenzie 1951), and/or distortions that arise from government-imposed quotas or sales taxes.

Joe Bain's treatment of thermal economies, recently repeated by Daniel Spulber (1999, p. 270), is illustrative of technological reasoning:

> [T]he cases of clear economies of integration generally involve a physical or technical aspect of the processes in a single plant. A classic case is that of integrating iron-making and steel-making to effect a saving in full costs by eliminating a reheating of iron before it is fed to a steel furnace. Where integration does not have this physical or technical aspect—as it does not, for example, in integrating the production of assorted components with the assembly of those components—the case for cost savings from integration is much less clear (Bain 1968, p. 381).

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

As a technological matter, however, the thermal economies to which Bain and Spulber refer actually require only that the two stages be located in close proximity to one another. That the two stages be placed under unified ownership is not implied. If, therefore, economies somehow accrue to the unified ownership of these two stages (that is, the relation between the two stages is better mediated by hierarchy rather than by market), this must be due to other, possibly transactional rather than technological, reasons.

TCE thus looks behind apparent explanations (such as price discovery or thermal economies) to see if they withstand comparative institutional scrutiny. It also asks whether outside procurement poses interfirm contractual hazards for which cost-effective relief will be realized upon taking the transaction in question into the firm (added bureaucratic costs notwithstanding). Specifically, the progressive buildup of contractual complications, as discussed in conjunction with the simple contractual schema in Figure 1, is mainly what explains successive moves from ideal market to hybrid to hierarchy.

So what about vertical market restrictions? How are these to be understood? For starters, vertical market restrictions can be interpreted as a decision to remain at node C rather than move to node D. The transaction in question is one to which hazards accrue ($h > 0$) for which cost-effective safeguards are needed ($s > 0$). If most of the hazards can be relieved at node C without incurring the added bureaucratic cost burdens (weakening of incentive intensity, added administrative costs) of unified ownership, then hybrid modes, of which franchising is an example, will be employed (provided that the contractual restrictions that accrue thereto are not treated as unlawful).

Vertical market restrictions often arise in the support of brand name capital (Klein 1980), where the concern is that such capital will be devalued by subgoal pursuit among independent or quasi-independent distributors (often franchisees), with the result that the integrity of the system is placed at risk. Depending on the particulars of the transaction, customer and territorial restrictions, exclusive dealing, or other franchise restrictions may be imposed. Absent strategic purpose, for which pre-existing monopoly power is a requisite, the choice of instruments for imposing vertical restraints will be discerned by examining where and how the contractual hazards originate.

Price theoretic explanations for nonstandard modes of contracting include the efficiency benefits that purportedly accrue to price discrimination, the benefits of

efficient risk bearing in the face of differential risk aversion, and the attenuation of free-rider hazards through the use of vertical market restrictions. The allocative efficiency benefits that accrue to price discrimination in a zero transaction cost world (which can be readily displayed in price theoretic terms) are much more problematic, however, if the costs of discovering customer preferences and of preventing arbitrage are positive. Invoking risk aversion to explain contracting practices among firms, moreover, is often second order in relation to more basic concerns with contractual hazards. Finally, unspecific free-rider claims are too often used as a shibboleth. The action, always and everywhere, resides in the details.[6]

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

The "New Economy"

Is there really a new economy? Yes and no. On the one hand, there is nothing new under the sun: real time responsiveness, innovation, outsourcing, and predatory behavior are not novel issues. Each of these has been magnified, however, by the deployment of new information technologies, by an increasing appreciation for relational contracting, and by the races for the commercialization and control of information age and biotechnology developments. A change in kind seems to describe competition in many high technology sectors.

Orthodox microtheory bears on some of these issues, but often in limited ways. TCE makes limited yet productive contact in the following respects: (*a*) Express provision for cooperative adaptation is congruent with the need for real time responsiveness; (*b*) innovation is examined in a systems context—in which firm size, incentives, and intertemporal transformations are featured (Williamson 1975, pp. 196–207); (*c*) crafting credible commitments to support outsourcing and the bureaucratic advantages of outsourcing over internal procurement are both TCE themes; and (*d*) tests for predation that exonerate behavior directed at less efficient competitors (Posner 1976, p. 193) are too static in that they fail to make provision for contingent predation—"now it's there, now it isn't, depending on whether an entrant has appeared or vanished" (Williamson 1977, p. 339), which introduces intertemporal considerations.

To be sure, new economy issues pose strategic and knowledge creation challenges that go beyond TCE (Shapiro & Varian 1999). Also, concepts such as "disequilibrium contracting" (Williamson 1991) boggle the mind. That TCE is more responsive to many of the pressing needs of public policy in the new economy than is received price theory is noteworthy but scarcely grounds for complacency.

[6]Although Posner (1979) contends that "the proper lens for viewing antitrust problems is price theory" (p. 932), Alan Meese (1997) observes that, "[d]espite references by Chicagoans to 'price theory,' Chicago's approach to vertical restraints has never rested upon. . .price theory. Instead, the Chicago approach to vertical restraints is an application of [NIE/TCE reasoning]" (p. 203). Also see Joskow (1991, pp. 567–57).

Regulation/Deregulation

FRANCHISE BIDDING[7] Posner's sanguine assessment of the efficacy of franchise bidding for natural monopoly begins with the claim that to "expound the details of particular regulations and proposals. . . would serve only to obscure the basic issues" (Posner 1972, p. 98). In the imperial tradition, all of the relevant action is concentrated in the ex ante bidding competition for the contract. This is consonant with Posner's dismissive view of organization theory, to which I referred at the outset, and illustrates the pitfalls of doing public policy analysis heedless of process transformations. Upon going beyond ex ante bidding competition to include ex post contract implementation, the attributes of the good or service to be franchised turn out to be crucial to an informed assessment. Specifically, if the good or service is to be supplied under conditions of uncertainty and if nontrivial investments in specific assets are involved, the efficacy of franchise bidding is highly problematic. The upshot is that franchise bidding for natural monopoly is not an all-purpose but rather a conditional solution.[8]

RESTRUCTURING ELECTRICITY SUPPLY IN CALIFORNIA Efforts to promote efficiency by creating markets for electric power have been implemented in a number of countries with varying degrees of success. California is a recent example where the efforts to restructure have been incompletely worked through. Again, the imperial view (this is the law here) trumps the process view (what's going on here?). This shows up in two respects. First, "good theories" were naively expected to be implemented without making provision for the realities of the political and regulatory process. Failing to make ex ante provision for these realities, politics and regulation are conveniently made the ex post scapegoats for behaving in perverse or unanticipated ways that, in large measure, were foreseeable and should have been factored into the calculus (Williamson 1996, chapter 8). Such lapses in realpolitik aside, Paul Joskow (2000) observes that too much deference was given to the (assumed) efficacy of smoothly functioning markets and insufficient attention was given to potential investment and contractual hazards and appropriate governance responses thereto. As Joskow puts it:

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

[7]This subsection is elaborated in Williamson (1996, pp. 84–85).

[8]Examples in which franchise bidding for goods and services supplied under decreasing cost conditions can possibly supplant extant regulation or public ownership with expected net gains include local service airlines and, possibly, postal delivery. The winning bidder for each base plant (terminals, post office, warehouses, and so on) can be owned by the government, and other assets (planes, trucks, and the like) will have an active secondhand market. Franchise bidding is not totally lacking in merit, therefore; on the contrary, it is a very imaginative proposal. TCE maintains, however, that all contracting schemes—of which franchise bidding for natural monopoly is one—need to be examined microanalytically and assessed in a comparative institutional manner.

> Many policy makers and fellow travellers have been surprised by how difficult it has been to create wholesale electricity markets. . . . Had policy makers viewed the restructuring challenge using a TCE framework, these potential problems are more likely to have been identified and mechanisms adopted ex ante to fix them (Joskow 2000, p. 51).

THE INSTITUTIONAL ENVIRONMENT The New Institutional Economics operates at two interrelated levels: the institutional environment (or rules of the game) and the institutions of governance (or play of the game). The study of privatizing telecommunications by Levy & Spiller (1994, 1996) examines the institutional environment (rules of the game) in five countries through a comparative contractual lens in which contractual hazards and credible contracting, or the lack thereof, are featured. This bottom-up approach reveals that the decision to privatize and the nature of privatization vary with the condition and quality of judicial independence, the division of powers between the executive and legislative branches, the competence of the regulatory bureaucracy, and contractual safeguards. Whether and how to privatize telecommunications should therefore be made conditional on these features.

Similar considerations arise in privatizing socialist economies. The "big bang" approach pays little heed to differences among industries, whereas those who are more concerned with cultivating institutions and the mechanisms of governance advise that a more gradual program be adopted in which the "easy cases" are privatized first. Because natural monopolies pose strains on deregulation and privatization alike (Arrow 2000, Williamson 2000), these are candidates to be privatized late (if at all) and then with the support of a fall-back regulatory apparatus.

Corporate Governance/Debt and Equity[9]

Price theory was long silent on the matter of corporate governance. Firms were simply assumed to maximize profits. The idea that managers might engage in subgoal pursuit that is contrary to profit maximization was inimical to the orthodox construction [although it can be and has been addressed in nearly orthodox terms by reformulating the objective function (Baumol 1959, Williamson 1964)].

TCE interprets the board of directors mainly as a security feature that arises in support of the contract for equity finance. Specifically, debt and equity are viewed not merely as alternative modes of finance, which is the law and economics construction (Easterbrook & Fischel 1986, Posner 1986), but also as alternative modes of governance. Thus, suppose that a firm is seeking cost-effective finance for the following series of projects: general-purpose mobile equipment; a general-purpose office building located in a population center; a general-purpose plant located in a manufacturing center; distribution facilities located somewhat more

[9]This subsection is based on Williamson (1996, pp. 184–85).

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

remotely; special-purpose equipment; market and product development expenses; and the like.

Suppose further that debt is a governance structure that works almost entirely out of rules. Specifically, assume that debt financing requires the debtor to observe the following: (*a*) Stipulated interest payments will be made at regular intervals; (*b*) the business will continuously meet certain liquidity tests; (*c*) sinking funds will be set up and principal repaid at the loan-expiration date; and (*d*) in the event of default, the debt-holders will exercise preemptive claims against the assets in question. If everything goes well, interest and principal will be paid on schedule. But debt is unforgiving if things go poorly. Failure to make scheduled payments thus results in liquidation. The various debt-holders will then realize differential recovery in the degree to which the assets in question are redeployable.

Because the value of a pre-emptive claim declines as the degree of asset specificity deepens, the terms of debt financing will be adjusted adversely. Confronted with the prospect that specialized investments will be financed on adverse terms, the firm might respond by sacrificing some of the specialized investment features in favor of greater redeployability. But then a lower cost of capital comes at an added production cost. Might it be possible to relieve the trade-off by inventing a new governance structure to which suppliers of finance would attach added confidence? In the degree to which this is feasible, value-enhancing investments in specific assets could thereby be preserved.

Suppose arguendo that a financial instrument called equity is invented, and assume that equity has the following governance properties: (1) It bears a residual-claimant status to the firm in both earnings and asset-liquidation respects; (2) it contracts for the duration of the life of the firm; and (3) a board of directors is created and awarded to equity, a board of directors that (*a*) is elected by the pro rata votes of those who hold tradeable shares, (*b*) has the power to replace the management, (*c*) decides on management compensation, (*d*) has access to internal performance measures on a timely basis, (*e*) can authorize audits in depth for special follow-up purposes, (*f*) is apprised of important investment and operating proposals before they are implemented, and (*g*) in other respects bears a decision-review and monitoring relation to the firm's management (Fama & Jensen 1983).

The board of directors thus evolves as a way to reduce the cost of capital for projects that involve limited redeployability. Not only do the added controls to which equity has access have better assurance properties, but equity is more forgiving than debt. Efforts are therefore made to work things out and preserve the values of a going concern when maladaptation occurs. Thus, whereas the governance structure associated with debt is of a very market-like kind, that associated with equity is much more intrusive and is akin to administration. The correspondence to which I referred earlier between outside procurement/debt and vertical integration/equity therefore obtains. In effect, debt is the market form of finance, and equity (the administrative form) appears as contractual hazards build up. Equity is the financial instrument of last resort.

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

Other Variations on a Theme

TCE maintains that any issue that arises as or can be posed as a contracting problem can be examined to advantage in transaction cost economizing terms. Accordingly, the reach of transaction cost reasoning is virtually endless. I briefly sketch two additional applications here (without bothering with price theoretic explanations).

PUBLIC BUREAUS According to Douglass North, "Political markets are. . . prone to inefficiency" (North 1990, p. 365) and "high transaction cost issues gravitate to the polity" (p. 372). That is worse than a paradox. That is perverse. Bad enough that political markets are inefficient. But surely the appropriate lesson is for high transaction cost issues to flee from rather than be attracted to the polity?

Maybe, but then again, maybe not. High transaction cost issues, after all, are ones that are inherently difficult to organize. As set out in Figure 1, such transactions are ones for which node A governance (in the market) is poorly suited compared with node D governance (in the firm). If still additional contractual hazards build up, might some of these transactions be candidates for governance in the public bureau? That is precisely the argument that I advance elsewhere (Williamson 1999). Specifically, the many disabilities of the public bureau notwithstanding—very low-powered incentives, very costly administrative procedures, very protective employment relations—there are some transactions (of which foreign affairs is an example) for which the public bureau comes off best judged, as it should be, comparatively. There is a place for each generic form of organization, yet each needs to be kept in its place.

LABOR ORGANIZATION The organization of labor reflects many purposes, monopsony power and political purposes included. What about efficiency? Again, the action resides in the details. Those labor transactions that pose greater contractual hazards ($h > 0$) will benefit from governance efforts to mitigate the hazards ($s > 0$), whereas it will be less cost-effective to supply these same safeguards to generic labor (of a node A kind), which is a recurrent theme. As developed elsewhere (Williamson et al. 1975; Williamson 1985, chapter 10), the observed organization of labor tracks an efficiency rationale.

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

CONTRACT AND ECONOMIC ORGANIZATION

Alternative Approaches

If the contractual approach to economic organization has the reach that I ascribe to it, then the systematic application of TCE to legal education and to legal and economic research on contracting holds out considerable promise. This will entail going beyond the "sort of contract law that has flourished in American law schools: the law embodied n judicial decisions and studied by analyzing these decisions" (Rubin 1995, p. 109). What Edward Rubin (1995) recommends instead is that the

law schools (and students of contract more generally) need a "theory of contract. . . that addresses the *contracting process itself*, rather than the judicial adjudication of that process," whereupon a "nonjudicial domain of contracting behavior" will be given prominence (p. 108, emphasis added).

In principle, law and economics could have been applied to that purpose. That project, however, took a "massive wrong turn" by the argument advanced by Posner and others that "the contract law goal [of] economic efficiency. . . [was] achieved through common-law adjudication" (Rubin 1995, p. 113). By drawing attention away from contracts and the contracting process toward judicial adjudication, "law and economics became just another tool for analyzing judicial decisions" (p. 113). Rubin is nevertheless heartened that although the "law school curriculum continues to be relatively resistant to a transactional theory of contract,. . . legal scholarship has gradually begun to shift its focus as a result of the economic and sociological analysis of transactions" (p. 114).

So what does a combined law, economics, and organizations approach to the study of contract, broadly conceived, entail? As I see it, the overarching move is to bring the lens of transaction cost economizing assiduously to bear. The examination of incomplete contracting in its entirety will be facilitated by supplanting the academic concept of contract as legal rules by that of private ordering and by inquiring into the mechanisms through which transaction cost economizing is accomplished. Interestingly, Ronald Gilson (1984) made many of these same arguments earlier in his examination of corporate finance transactions.

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

The Economizing Perspective

The economizing perspective holds that, subject to the remediableness criterion, inefficiency invites its own demise—where inefficiency is assessed in relation to feasible alternatives (rather than a hypothetical ideal) and provision is made for implementation costs. Because joint gains will always be realized by moving from a less to a more efficient mode, provided that implementation costs do not dissipate the gains, farsighted businessmen and their lawyers will eschew inferior outcomes (such as node B in the schema). In contrast to Machiavelli's myopic advice to "get them before they get us," the farsighted view of contracting is to "give and receive credible commitments" (Williamson 1983, 1993b)—by providing better information and added security features that serve to infuse confidence and realize mutual gains.

Gilson's (1984, p. 255) description of business lawyers as transaction cost engineers is very much in this spirit. He thus urges that transactions be examined not in a one-sided way but "from the perspective of *both* clients" (p. 245; emphasis in original), whence mutual gain is the object. He furthermore adopts a transaction cost economizing approach to private ordering (Gilson 1984, p. 255), including express reference to credible commitments (p. 281). Also, he views departures from the assumptions of the (ideal) capital asset pricing model—namely, common time horizon, identical expectations, no transaction costs, and costless information

(Gilson 1984, p. 252)—as grist for the TCE mill: "[T]he unreality of these [ideal]. . . assumptions is not cause for despair. Rather, it is the very failure of these assumptions to describe the real world that I find the *potential for value creation* by lawyers" (Gilson 1984, p. 253; emphasis added). The institutions of governance arise precisely on account of these disparities (Arrow 1963).

Private Ordering

THE CONCEPT Marc Galanter (1981) takes exception with the usual academic/legal centralist approach to contract in which disputes purportedly "require 'access' to a forum external to the original social setting of the dispute [whereby] remedies will be provided as prescribed in some body of authoritative learning and dispensed by experts who operate under the auspices of the state" (p. 1). The facts disclose otherwise: Most disputes, including many that under current rules could be brought to a court, are resolved by avoidance, self-help, and the like (p. 2). This is because in "many instances the participants can devise more satisfactory solutions to their disputes than can professionals constrained to apply general rules on the basis of limited knowledge of the dispute" (p. 4). Gilson (1984) concurs: When business lawyers play the role of transaction cost engineer well, "the courts, and formal law generally, shrink dramatically in importance" (p. 294).

CONTRACT LAWS (PLURAL) Karl Llewellyn's (1931) earlier dissent from the legal rules approach to contract introduces the concept of contract as framework:

> [T]he major importance of legal contract is to provide a framework for well-nigh every type of group organization and for well-nigh every type of passing or permanent relation between individuals and groups. . .—a framework highly adjustable, a framework which almost never accurately indicates real working relations, but which affords a rough indication around which such relations vary, an occasional guide in cases of doubt, and a norm of ultimate appeal when the relations cease in fact to work (Llewellyn 1931, pp. 736–37).

This last point is important in that the prospect of ultimate appeal to the courts serves to delimit threat positions.

Related ideas have been advanced by others, including Clyde Summers (1969) who distinguishes between "black letter law" (which bears a likeness to black box economics) and a more circumstantial approach to contract. The former employs the counterfactual "illusion that contract rules can be stated without reference to surrounding circumstances and are therefore generally applicable to all contractual relations" (p. 566).

The TCE argument that each generic mode of governance is supported by a distinctive form of contract law is broadly in this circumstantial spirit. The ideal (node A) transaction in both law and economics is that of spot markets to which identity is unimportant and legal rules apply (Macneil 1974). This legal rules approach gives way to Llewellyn's concept of contract-as-framework as the importance of continuity builds up and incomplete long-term contracting is adopted

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

(node C). That in turn undergoes change when transactions are taken out of the market and organized internally (node D), where the implicit law of contract now becomes that of forbearance. As previously noted, courts routinely grant standing to firms engaged in interfirm contracting should there be disputes over prices, the damages to be ascribed to delays, failures of quality, and the like, yet courts will refuse to hear disputes between one internal division and another over identical technical issues. Access to the courts being denied, the parties must resolve their differences internally (Rubin 1995, p. 117). Accordingly, hierarchy is its own court of ultimate appeal. That firms and markets differ in their access to fiat is partly explained by these contract law differences (Williamson 1991).

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

Mechanisms

CORPORATE ACQUISITION TRANSACTIONS TCE subscribes to the dictum that "explanations in the social sciences should be organized around (partial) *mechanisms* rather than general *theories*" (Elster 2000, p. 75; emphasis in original). That is evident in the way by which TCE examines the canonical make-or-buy decision and of contracting more generally. It is also evident in Gilson's (1984) examination of efforts by business lawyers to perfect the acquisition agreement in the face of "deviations" from the ideal assumptions of the capital asset pricing model:

> Earnout or contingent-pricing techniques respond to the failure of the homogeneous expectations assumption; controls over operation of the seller's business during the period in which the determinants of the contingent price are measured respond to failure of the common-time-horizon assumption; and the panopoly of representations and warranties, together with provisions for indemnification and other verification techniques, respond to the failure of the costless-information (Gilson 1984, p. 293).

CONTRACT LAW DOCTRINE A microanalytic examination of the mechanisms that arise in conjunction with contract law doctrines would also be illuminating. Ian Macneil (1974) describes the legal system's "less than total commitment to the keeping of promises" as follows:

> Contract remedies are generally among the weakest of those the legal system can deliver. But a host of doctrines and techniques lies in the way of even those remedies: impossibility, frustration, mistake, manipulative interpretation, jury discretion, consideration, illegality, duress, undue influence, unconscionability, capacity, forfeiture and penalty rules, doctrines of substantial performance, severability, bankruptcy laws, statutes of fraud, to name a few; almost any contract doctrine can and does serve to make the commitment of the legal system to promise keeping less than complete (Macneil 1974, p. 730).

The refusal by the courts to enforce stipulated damages clauses is especially puzzling. Because the parties to a contract can be presumed to know best what

contractual terms serve their interests, why should the courts refuse to enforce stipulated damages in the event of breach?

One possibility is that contract is a devious thing. Thus, although such a clause may frequently be the efficient way to settle a breach, it could also serve strategic purposes, of which induced breach is one.

The issue of contrived cancellation has been addressed by Kenneth Clarkson, Roger Miller, and Timothy Muris in their discussion of refusal of the courts to enforce stipulated damage clauses where breach has been deliberately induced (Clarkson et al. 1978, pp. 366–72). Induced breach could arise when a party intentionally withholds relevant information yet complies with the letter of the contract. Or it might involve perfunctory fulfillment of obligations where more resourceful cooperation is needed (pp. 371–72). In either case, induced breach is costly to detect and/or prove (p. 371). Transaction cost considerations are plainly operative.

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

Ramifications for Legal Education

Gilson (1984) advises that my observation that the legal centralism approach to contract relieves "lawyers and economists... of the need to examine the variety of ways by which individual parties to exchange 'contract out of or away from' the governance structures of the state by devising private orderings" (Williamson 1983, p. 520) is too sweeping. It should be restricted to academic lawyers and economists (Gilson 1984, p. 295). That is because "business lawyers have done an awfully good job at something the law schools did not and, for the most part, still do not teach: helping people arrange their relationships in the absence of governmental intervention: facilitating *private ordering*" (Gilson 1984, p. 303; emphasis in original). But then "why have law schools done so bad a job training business lawyers?" (p. 303). Gilson's answer is that "There has been no theory... that dealt with private ordering" (p. 304) prior to the appearance of "two areas in economics—finance and transaction cost economics" (p. 305).

Twenty years later we find that the teaching of contract law has changed very little. What explains this continuing neglect?

One explanation is that mainline law and economics has remained comfortably ascendant. The relation between law and economics thus continues to be one in which textbook economic orthodoxy is the fount. The predilection to work out of a theory of the firm-as-production function setup is thus reaffirmed and the subject of organization remains disjunct. Reservations about the efficiency of common law adjudication notwithstanding, contract law teaching stays predominantly focused on legal rules and adjudication.

A second explanation is that the world of private ordering is impossibly complex. As good lawyers are quick studies, better that they learn about private ordering on the job rather than in the classroom.

The first of these arguments is a lame excuse for complacency, whereas the second overlooks the possibility that the economics of organization involves

variations on a few key themes. In that event, attention can be focused on canonical cases—of which credible interfirm contracting is one and vertical integration is another. The buzzing, blooming confusion of private ordering is thereby reduced to more manageable proportions. Because the classroom is the place to lay out the intuition, merits, and mechanisms of credible contracting (node C) and to examine the comparative strengths and weaknesses of the firm-as-governance structure (node D), to relegate the study of private ordering to on-the-job training is anachronistic.

Even, moreover, if the basic law school curriculum is unmoved by these arguments, it is noteworthy that a number of leading law schools have begun to offer an elective course on complex deals, many of them modeled after the course offered by Gilson and Victor Goldberg at Columbia Law School on "Deals: The Economic Structure of Transactions and Contracting." If the demand for transaction cost engineers cannot be met by the law schools, the business schools could end up eating that lunch (Rubin 1995, p. 114).

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

CONCLUSIONS

There is growing agreement that "the objectives of firms, the reason for their existence and the manner of their decision taking. . . will require modes of analysis quite different from those which have dominated in this century" (Hahn 1991, p. 49). Not only does TCE hold that the way to think about contract and organization is to bring the purposive and farsighted lens of economizing to bear,[10] but the existence and governance of firms are both the key TCE issues.

As developed herein, organization theory has massive ramifications for the TCE theory of the firm. Salient contributions from organization theory include the description of human actors in more veridical terms, the importance of intertemporal process transformations, choice of the unit of analysis, and the description of alternative modes of governance as syndromes of complementary attributes. The resulting theory of the firm differs greatly from the neoclassical (Kreps 1990, p. 96). Because "[a]ny standard theory, not just neoclassical, starts from the existence of firms" (Arrow 1999, p. vii), that is very basic.

To be sure, the proximate lessons (as advanced by organization theorists) and the ultimate lessons (as viewed from an economizing perspective) often differ—and that is consequential. But the more basic point is this: Someone needed to

[10]Farsighted contracting is more plausible in intermediate product market contracts than in final goods markets. Still, farsighted firms that are selling to consumers who lack the relevant expertise and foresight nevertheless can and do take steps to alleviate the hazards—through branding, warranties, guarantees, and the like. I do not mean to suggest, however, that there is never an occasion to craft additional relief (possibly with the aid of public policy) against residual hazards.

step up and offer trenchant critiques and identify relevant phenomena. Organization theorists were prepared to do that when others were complacent or held back.

The theory of the firm-as-governance structure that is sketched herein is an ongoing rather than finished construction.[11] Its evolving status notwithstanding, it has already served to deepen our understanding of many complex contractual and organizational phenomena and it operates as a check against overuses and misuses of orthodoxy. In that spirit, I suggest that mainstream law and economics stands to benefit by incorporating the lessons and some of the methods of law, economics, and organization—both as these bear on public policy and in relation to the law school curriculum.[12]

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

ACKNOWLEDGMENTS

This paper was originally prepared for the ceremony inaugurating the Program in Law and Economics at the University of Chile in August 2000. That program is a joint effort between the Faculty of Economics and Business and the Faculty of Law to "promote and develop interdisciplinary theoretical and applied research in the area." It was subsequently presented at the opening session of the fourteenth annual conference of the International Society for New Institutional Economics in Tuebingen, Germany, at the Law and Economics Seminar at George Mason University, the Law, Economics, and Organization Seminar at the University of Southern California, and at the First Annual Symposium at the Center for Legal Dynamics of Advanced Market Societies at Kobe University. Useful questions and comments at all five sessions are gratefully acknowledged. An earlier version of this paper was published in the *Kobe University Law Review* (2004) 38:59–95.

[11] Full formalization is the ultimate objective. The Grossman-Hart-Moore model (Hart 1995) qualifies as a fully formal model but is lacking in plausibility (Kreps 1999). The treatment of procurement by Bajari & Tadelis (2001), which focuses on the incentive and ex post adaptation differences between fixed price and cost plus contracting, is much closer in spirit to TCE.

[12] One of the comments that I have received on this review is that the basic message has not only been heard, but that it has registered and taken effect. That is gratifying, yet other readers remark that much of this is unfamiliar terrain and needs to be more fully spelled out. I come out somewhere in between. Thus, although many of the firm-as-governance structure ideas have taken hold, private ordering remains underdeveloped and organization theory is scanted by mainline law and economics—witness the leading textbooks (Cooter & Ulen 2000, Polinsky 1989, Posner 1998). Public policy inroads notwithstanding, the basic contract law course remains immune to the arguments in this paper.

The *Annual Review of Law and Social Science* is online at http://lawsocsci.annualreviews.org

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org by University of Colorado - Boulder on 04/12/10. For personal use only.

LITERATURE CITED

Arrow KJ. 1963. Uncertainty and the welfare economics of medical care. *Am. Econ. Rev.* 53(December):941–73

Arrow KJ. 1999. Forward. In *Firms, Markets, and Hierarchies*, ed. G Carroll, D Teece, pp. vii–viii. New York: Oxford Univ. Press

Arrow KJ. 2000. Economic transition: speed and scope. *J. Inst. Theor. Econ.* 156(March): 9–18

Bain J. 1968. *Industrial Organization*. New York: Wiley. 2nd ed.

Bajari P, Tadelis S. 2001. Incentives versus transaction costs: a theory of procurement contracts. *RAND J. Econ.* 32(3):387–407

Barnard C. 1962 (1938). *The Functions of the Executive*. Cambridge: Harvard Univ. Press

Baumol WJ. 1959. *Business Behavior, Value and Growth*. New York: Macmillan

Ben-Porath Y. 1980. The F-connection: family, friends, and firms and the organization of exchange. *Popul. Dev. Rev.* 6:1–30

Buchanan JM. 1975. A contractarian paradigm for applying economic theory. *Am. Econ. Rev.* 65(2):225–30

Buchanan JM. 1987. The constitution of economic policy. *Am. Econ. Rev.* 77(3):243–50

Clarkson KW, Miller RL, Muris TJ. 1978. Liquidated damages v. penalties. *Wis. Law Rev.* 1978:351–90

Coase RH. 1992. The institutional structure of production. *Am. Econ. Rev.* 82(September): 713–19

Commons JR. 1932. The problem of correlating law, economics, and ethics. *Wis. Law Rev.* 8:3–26

Cooter R, Ulen T. 2000. *Law and Economics*. Reading, MA: Addison-Wesley. 3rd ed.

Dawkins R. 1976. *The Selfish Gene*. New York: Oxford Univ. Press

Demsetz H. 1983. The structure of ownership and the theory of the firm/comment. *J. Law Econ.* 26(June):375–93

Easterbrook F, Fischel D. 1986. Close corporations and agency costs. *Stanford Law Rev.* 38(January):271–301

Elster J. 2000. Arguing and bargaining in two constituent assemblies. *Univ. Penn. J. Const. Law* 2:345

Fama EF, Jensen MC. 1983. Separation of ownership and control. *J. Law Econ.* 26 (June):301–26

Galanter M. 1981. Justice in many rooms: courts, private ordering, and indigenous law. *J. Leg. Plur.* 19:1–47

Gilson R. 1984. Value creation by business lawyers: legal skills and asset pricing. *Yale Law J.* 94(December):239–313

Hahn F. 1991. The next hundred years. *Econ. J.* 101(January):47–50

Hart O. 1995. *Firms, Contracts, and Financial Structure*. New York: Oxford Univ. Press

Hayek F. 1945. The use of knowledge in society. *Am. Econ. Rev.* 35(September):519–30

Joskow P. 1991. The role of transaction cost economics in antitrust and public utility regulatory policies. *J. Law Econ. Org.* 7(Special Issue):53–83

Joskow PL. 2000. *Transaction cost economics and competition policy.* Work. Pap., MIT

Klein B. 1980. Transaction cost determinants of 'unfair' contractual arrangements. *Am. Econ. Rev.* 70(May):356–62

Klein B, Leffler K. 1981. The role of market forces in assuring contractual performance. *J. Polit. Econ.* 89(May):615–41

Kreps DM. 1990. Corporate culture and economic theory. In *Perspectives on Positive Political Economy*, ed. J Alt, K Shepsle, pp. 90–143. New York: Cambridge Univ. Press

Kreps DM. 1999. Markets and hierarchies and (mathematical) economic theory. In *Firms, Markets, and Hierarchies*, ed. G Carroll, D Teece, pp. 121–55. New York: Oxford Univ. Press

Levy B, Spiller P. 1994. The institutional foundations of regulatory commitment: a comparative analysis of telecommunications regulation. *J. Law Econ. Org.* 10(October):201–46

Levy B, Spiller P. 1996. *Regulations, Institutions, and Commitment: Comparative Studies of Telecommunications*. Cambridge, UK: Cambridge Univ. Press

Llewellyn KN. 1931. What price contract? An essay in perspective. *Yale Law J.* 40:704–51

Macneil IR. 1974. The many futures of contracts. *South. Calif. Law Rev.* 47(May):691–816

March JG, Simon HA. 1958. *Organizations*. New York: Wiley

McKenzie L. 1951. Ideal output and the interdependence of firms. *Econ. J.* 61(December):785–803

Meese AJ. 1997. Price theory and vertical restraints: a misunderstood relation. *UCLA Law Rev.* 45(October):143–204

Michels R. 1962. *Political Parties*. Glencoe, IL: Free

North D. 1990. A transaction cost theory of politics. *J. Theor. Polit.* 2(4):355–67

Polinsky AM. 1989. *An Introduction to Law and Economics*. Boston: Little, Brown. 2nd ed.

Posner RA. 1972. The appropriate scope of regulation in the cable television industry. *Bell J. Econ. Manag. Sci.* 3(1):98–129

Posner RA. 1976. *Antitrust Law*. Chicago: Univ. Chicago Press

Posner RA. 1979. The Chicago School of antitrust analysis. *Univ. Penn. Law Rev.* 127(April):925–48

Posner RA. 1986. *Economic Analysis of Law*. Boston: Little, Brown. 3rd ed.

Posner RA. 1993. The new institutional economics meets law and economics. *J. Inst. Theor. Econ.* 149(March):73–87

Posner RA. 1998. *Economic Analysis of Law*. New York: Aspen Law & Business. 5th ed.

Rabin M. 1998. Psychology and economics. *J. Econ. Lit.* 36(March):11–46

Reder M. 1999. *The Culture of a Controversial Science*. Chicago: Univ. Chicago Press

Riordan M, Williamson O. 1985. Asset specificity and economic organization. *Int. J. Ind. Org.* 3:365–78

Robbins L. 1932. *Essay on the Nature and Significance of Economic Science*. London: Macmillan

Rubin E. 1995. The non-judicial life of contract: beyond the shadow of the law. *Northwestern Univ. Law Rev.* 90(Fall):107–31

Schultz G. 1995. Economics in action: ideas, institutions, policies. *Am. Econ. Rev. Papers Proc.* 85(May):1–8

Scott WR. 1987. *Organizations: Rational, Natural, and Open Systems*. Upper Saddle River, NJ: Prentice Hall

Selznick P. 1950. The iron law of bureaucracy. *Modern Rev.* 3:157–65

Shapiro C, Varian HR. 1999. *Information Rules: A Strategic Guide to the Network Economy*. Boston, MA: Harvard Bus. Sch.

Simon H. 1947. *Administrative Behavior*. New York: Macmillan

Simon H. 1957a. *Administrative Behavior*. New York: Macmillan. 2nd ed.

Simon H. 1957b. *Models of Man*. New York: Wiley

Simon H. 1978. Rationality as process and as product of thought. *Am. Econ. Rev.* 68(May):1–16

Simon H. 1985. Human nature in politics: the dialogue of psychology with political science. *Am. Polit. Sci. Rev.* 79:293–304

Spulber D. 1999. *Market Microstructure: Intermediaries and the Theory of the Firm*. Cambridge, UK: Cambridge Univ. Press

Summers C. 1969. Collective agreements and the law of contracts. *Yale Law J.* 78(March):537–75

Williamson OE. 1964. *The Economics of Discretionary Behavior: Managerial Objectives in a Theory of the Firm*. Englewood Cliffs, NJ: Prentice-Hall

Williamson OE. 1975. *Markets and Hierarchies: Analysis and Antitrust Implications*. New York: Free

Williamson OE. 1977. Predatory pricing: a strategic and welfare analysis. *Yale Law J.* 87(December):284–340

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

Williamson OE. 1983. Credible commitments: using hostages to support exchange. *Am. Econ. Rev.* 73 (September):519–40

Williamson OE. 1985. *The Economic Institutions of Capitalism*. New York: Free

Williamson OE. 1991. Comparative economic organization: the analysis of discrete structural alternatives. *Admin. Sci. Q.* 36(June): 269–96

Williamson OE. 1993b. Calculativeness, trust, and economic organization. *J. Law Econ.* 36(April):453–86

Williamson OE. 1996. *The Mechanisms of Governance*. New York: Oxford Univ. Press

Williamson OE. 1999. Public and private bureaucracies. *J. Law Econ. Org.* 15(April): 306–42

Williamson OE. 2000. *Empirical Microeconomics*: *Another Perspective.* Work. Pap., Inst. Manag. Innov. Org., Univ. Calif., Berkeley

Williamson OE, Wachter ML, Harris JE. 1975. Understanding the employment relation: the analysis of idiosyncratic exchange. *Bell J. Econ.* 6(Spring):250–80

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

Annual Review of Law and Social Science
Volume 1, 2005

CONTENTS

COMING OF AGE: LAW AND SOCIETY ENTERS AN EXCLUSIVE CLUB, *Lawrence M. Friedman* 1

THE COMPARATIVE STUDY OF CRIMINAL PUNISHMENT, *James Q. Whitman* 17

ECONOMIC THEORIES OF SETTLEMENT BARGAINING, *Andrew F. Daughety and Jennifer F. Reinganum* 35

LAW AND CORPORATE GOVERNANCE, *Neil Fligstein and Jennifer Choo* 61

TRANSNATIONAL HUMAN RIGHTS: EXPLORING THE PERSISTENCE AND GLOBALIZATION OF HUMAN RIGHTS, *Heinz Klug* 85

EXPERT EVIDENCE AFTER *DAUBERT*, *Michael J. Saks and David L. Faigman* 105

PLEA BARGAINING AND THE ECLIPSE OF THE JURY, *Bruce P. Smith* 131

THE DEATH PENALTY MEETS SOCIAL SCIENCE: DETERRENCE AND JURY BEHAVIOR UNDER NEW SCRUTINY, *Robert Weisberg* 151

VOICE, CONTROL, AND BELONGING: THE DOUBLE-EDGED SWORD OF PROCEDURAL FAIRNESS, *Robert J. MacCoun* 171

LAW, RACE, AND EDUCATION IN THE UNITED STATES, *Samuel R. Lucas and Marcel Paret* 203

LAW FACTS, *Arthur L. Stinchcombe* 233

REAL JURIES, *Shari Seidman Diamond and Mary R. Rose* 255

FEMINISM, FAIRNESS, AND WELFARE: AN INVITATION TO FEMINIST LAW AND ECONOMICS, *Gillian K. Hadfield* 285

CRIMINAL DISENFRANCHISEMENT, *Christopher Uggen, Angela Behrens, and Jeff Manza* 307

AFTER LEGAL CONSCIOUSNESS, *Susan S. Silbey* 323

WHY LAW, ECONOMICS, AND ORGANIZATION? *Oliver E. Williamson* 369

REVERSAL OF FORTUNE: THE RESURGENCE OF INDIVIDUAL RISK ASSESSMENT IN CRIMINAL JUSTICE, *Jonathan Simon* 397

INDEX
Subject Index 423

Annu. Rev. Law. Soc. Sci. 2005.1:369-396. Downloaded from arjournals.annualreviews.org
by University of Colorado - Boulder on 04/12/10. For personal use only.

Creative Destruction's Reconstruction: Joseph Schumpeter Revisited

By J. BRADFORD DELONG

My guess is that average literate Americans know of three 20th-century economists: John Maynard Keynes, Milton Friedman, and Alan Greenspan. Perhaps they also know of Paul Samuelson (but as textbook author, not economic theorist), of Friedrich Hayek (but think that he is the father of an actress), and of John Kenneth Galbraith (as William F. Buckley Jr.'s friend who appeared on TV). The rest of us disappear into a blur of gray suits, spectacles, and, usually, baldness — an assemblage of personalities too bland to be successful accountants.

In Prophet of Innovation: Joseph Schumpeter and Creative Destruction (Belknap Press/Harvard University Press, 2007), Thomas K. McCraw, an emeritus professor of business history at Harvard Business School, tries to add another name to the list — Joseph Schumpeter. From the start, you should know that I am partial to Schumpeter; McCraw quotes me and Larry Summers saying that if Keynes was the most important economist of the 20th century, then Schumpeter may well be the most important of the 21st. McCraw's fascinating book details why that is so, but McCraw, too, is so partial to Schumpeter that he fails to fully explain how Schumpeter tripped himself up over a political understanding as clumsy as his economic understanding was brilliant.

Schumpeter was born in 1883 in what is now the Czech Republic, and died in America in 1950. When he was four, his father died. When he was 10, his mother remarried and moved to Vienna, where his aristocratic stepfather helped him enter elite schools. He was a star: youngest professor in the empire at 26, finance minister of Austria (briefly) at 36, a bank president, and then a professor again. His mother, wife, and newborn son died in rapid succession in 1926 — an awful and tragic shock. He was important enough for Harvard to pluck him from Europe in 1932 to make him one of its superstars. And, after his move to America, he was one of the most famous American economists.

But as fascinating as his life was, it is Schumpeter's economics that sing to me, because he tried to set long-term economic growth — entrepreneurship and enterprise — at the top of the discipline's agenda.

Over the previous two and a half centuries, three different economic worldviews, in succession, reigned. In the late 18th and early 19th centuries, Adam Smith's was the key economic perspective, focusing on domestic and international trade and growth, the division of labor, the power of the market, and the minimal security of property and tolerable administration of justice that were

needed to carry a country to prosperity. You could agree or you could disagree with Smith's conclusions and judgments, but his was the proper topical agenda.

The second reign was that of David Ricardo and Karl Marx. Their preoccupations dominated the late 19th and early 20th centuries. They worried most about the distribution of income and the laws of the market that made it so unequal. They were uneasy about the extraordinary pace of technological, organizational, and sociological change, and about whether an ungoverned market economy could produce a distribution of income — both relative and absolute — fit for a livable world. Again, you could agree or disagree with their judgments about trade, rent, capitalism, and machinery, but they asked the right questions.

The third reign was that of John Maynard Keynes. His agenda dominated the middle and late 20th century. Keynes's theories centered on what economists call Say's Law — the claim that except in truly exceptional conditions, production inevitably creates the demand to buy what is produced. Say's Law supposedly guaranteed something like full employment, except in truly exceptional conditions, if the market was allowed to work. Keynes argued that Say's Law was false in theory, but that the government could, if it acted skillfully, make it true in practice. Agree or disagree with his conclusions, Keynes was in any case right to focus on the central bank and the tax-and-spend government to supplement the market's somewhat-palsied invisible hand to achieve stable and full employment.

B ut there ought to have been a fourth reign, for there was a set of themes not sufficiently explored. That missing reign was Schumpeter's, for he had insights into the nature of markets and growth that escaped other observers. It is in that sense that the late 20th and early 21st centuries in economics ought to have been his: He asked the right questions for our era.

He asked those questions in a book he wrote while working at the University of Czernowitz in his mid-20s: the Theory of Economic Development. Previous first-rank economists (with the partial exception of Marx) had concentrated on situations of equilibrium. In that model, development is a gradual process, in which competition keeps goods high-quality and affordable, and the abstemious owners of capital await the long-term rewards of deferred gratification.

Schumpeter pointed out that that wasn't how market economies really worked. The essence of capitalist economies was, as Marx had recognized before him, the entrepreneur and the innovator: the risk taker who sets in motion new and more-efficient ways of making old or new products, and so produces an economy in constant change. Marx saw that the coming of capitalist economies destroyed all feudal, traditional, and patriarchal relationships and orders. Schumpeter saw farther:

that market capitalism destroys its own earlier generations. There is, he wrote, a constant "process of industrial mutation — if I may use that biological term — that incessantly revolutionizes the economic structure from within, incessantly destroying the old one, incessantly creating a new one. This process of Creative Destruction is the essential fact about capitalism. It is what capitalism consists in, and what every capitalist concern has got to live in."

In a later book, Capitalism, Socialism, and Democracy, Schumpeter wrote that the traditional view of competition must be abandoned. "Economists," he said, "are at long last emerging from the stage in which price competition was all they saw. ... However, it is still competition within a rigid pattern of invariant conditions, methods of production and forms of industrial organization ... that ... monopolizes attention. But in capitalist reality as distinguished from its textbook picture, it is not that kind of competition which counts but the competition from the new commodity, the new technology, the new source of supply, the new type of organization — competition which commands a decisive cost or quality advantage and which strikes not at the margins of the profits and the outputs of the existing firms but at their foundations and their very lives."

Entrepreneurs innovate new ways of manipulating nature, and new ways of assembling and coordinating people. It is important to stress that a Schumpeterian entrepreneur is not an inventor, but an innovator. The innovator shows that a product, a process, or a mode of organization can be efficient and profitable, and that elevates the entire economy. But it also destroys those organizations and people who suddenly find their technologies and routines outmoded and unprofitable. There is, Schumpeter was certain, no way of avoiding this: Capitalism cannot progress without creating short-term losers alongside its short- and long-term winners: "Without innovations, no entrepreneurs; without entrepreneurial achievement, no capitalist ... propulsion. The atmosphere of industrial revolutions ... is the only one in which capitalism can survive."

Schumpeter's ideas lay waste to economists' smooth graphs of long-run growth trends and economic evolution. Growth produces progress and wealth, but in unforeseeable ways and in discrete lumps that create many small winners (for example, the people who can now buy their shirts at Wal-Mart for $8.99 as opposed to $12.99 at its less-efficient competitors), a few huge winners (for example, the Walton family of Bentonville, Ark.), and notable substantial losers (the Main Street merchants of the Mississippi Valley, the Great Plains, and the Sun Belt).

Schumpeter, like his contemporary Karl Polanyi, feared for the long-term survival of capitalism. Bureaucrats and ideologues threatened by creative destruction would resist it. The challenge for the government in managing the market thus becomes not just the Adam Smithian task of securing

property rights, enforcing contracts, and providing civil order, but also the tremendously difficult job of managing the creative destruction so that capitalism does not undermine and destroy itself for essentially political reasons. Schumpeter did not think the beast could be managed, because democracy is hostile to great inequalities, and socialism even more so.

Capitalism, however, inevitably generates these mammoth inequalities through creative destruction. The combinations of market economies and political democracies that we see today in the richest countries in the world were, Schumpeter thought, unlikely to be stable. No country that wanted to see rapid economic growth could afford to remain a political democracy for long.

He did not think governments could maintain enough social insurance to counter the destructive part of capitalism without strangling the sources of rapid growth. But why Schumpeter's Capitalism, Socialism, and Democracy places so much blame on "democracy" is unclear to me: Oligarchs fear change at least as much as democratic electorates do.

All told, Schumpeter's economic theorizing was keen, but his political judgment was abysmal. Franklin D. Roosevelt was not a communist aiming to abolish elections and turn himself into a dictator — but Schumpeter thought he was. Efforts to boost demand and employment in 1933 via deficit spending, abandoning the gold standard, and trying to keep interest rates as low as possible were not counterproductive and destructive — but Schumpeter argued that they were. He was, to put it mildly, unsuccessful as finance minister for the brand-new post-World War I Austrian Republic. He erred greatly in his assessment of Germany: A strong Germany was not a stabilizing bulwark against Bolshevism and a source of peace, order, and strength in Europe — at least not in Schumpeter's day.

Moreover, truth be told, even Schumpeter's academic judgment was not that great. His Theory of Economic Development is genius. Much of Capitalism, Socialism, and Democracy is superb. His pre-World War I essay "The Sociology of Imperialism" is also very fine. It argues that imperialism and empire were from an older political tradition given new power and energy by modern capitalism's explosion of wealth, but that in the long run, market capitalism was unsympathetic to imperial domination. Why pay to rule when you can get almost all of what you want through trade? Nevertheless, there are many dry holes and detours in Schumpeter's academic work.

Indeed, Schumpeter's writing late in his career is, in my view at least, not worth reading. The signal-to-noise ratio is simply too low. For instance, the dull, posthumous History of Economic Analysis grades thinkers of the past depending on how closely they were able to approach what Schumpeter regarded as the proper vision and analytical tools of his present.

I think he had decided that he was a Great Man, and that Great Men write Great Books, and so anything that was not a Great Book was beneath him. But the two Great Books written in economics while Schumpeter was working on his were written by people who were not trying to write Great Books but who had something important to say. Keynes's The General Theory of Employment, Interest and Money says that demand management through fiscal and monetary policy can make Say's Law true in practice even though it is false in theory. Samuelson's Foundations of Economic Analysis says that mathematics is a superbly useful language for economics.

Thus Schumpeter's influence was limited. He died half a decade later than Keynes, but Keynes's reputation put Schumpeter's in near-total eclipse for more than a generation. McCraw greatly prefers Schumpeter's Business Cycles to Keynes's General Theory, which he scorns as having no mention of any individual businesses in its 400-odd pages.

McCraw is thus too devoted to his subject to see one essential strand of Schumpeter's story: the academic undervaluation that resulted from his political delusions. This is not the place to write about the rhetorical excellence and the analytic diamonds in Keynes's General Theory (that place is Paul Krugman's introduction to Keynes's book). But I can say that General Theory develops, justifies, rationalizes, extends, applies, and shows economists how to use what had by then become the author's standard themes and approaches in a way that Business Cycles, the History of Economic Analysis, and even Capitalism, Socialism, and Democracy do not. Back in 1970, the economist Harry G. Johnson pointed out that all successful founders of schools not only are geniuses with profound insights but also provide a road map that tells their followers and successors what to do to make a successful academic career within the school. Schumpeter did not do that second part.

Perhaps this next century will give Schumpeter's work its proper place as the power of innovation to transform, create, enrich, and destroy makes itself manifest globally. And while we marvel at how much he got right, we can hope Schumpeter was wrong in his political analysis. One great test of our era will be whether creative destruction can flourish alongside public order and political liberty. If not, we're in big trouble. But if so — and I'm an optimist on the point — the results could be a marvel.

J. Bradford DeLong is a professor of economics at the University of California at Berkeley.



Journal of Regulatory Economics; 21:1 35–56, 2002
© 2002 Kluwer Academic Publishers. Manufactured in The Netherlands.

The Deregulatory Tar Baby: The Precarious Balance Between Regulation and Deregulation, 1970–2000 and Henceforward*

ALFRED E. KAHN
Cornell University
Robert Julius Thorne Professor of Political Economy, Emeritus, Special Consultant, National Economic Research Associates, Inc. (NERA), 308 North Cayuga Street, Ithaca, NY 14850
E-mail: alfred.kahn@nera.com

Abstract
The recent history of the airline, telecommunications and electric utility industries years amply demonstrates the benefits of deregulation. It also demonstrates both the necessity of continued government involvement and its pitfalls. Prominent among the former are the efficient provision and pricing of infrastructure services, enforcement of antitrust-like policies and regulation of bottleneck facilities. The latter come down to an inherent tendency of regulators to engage in continued micromanagement—to continue to assume responsibility not merely for opening the door to competition but for ensuring that competitors go through it, and to prescribe the results they think an efficient competitive process would produce.

1. Introduction

When I became chairman of the New York Public Service Commission, in the summer of 1974, I had not the slightest notion that I was assuming the leadership of an enterprise that either was or should have been in process of going out of business. True, I soon discovered that my distinguished predecessor, Joe Swidler, had agreed with Rochester Telephone to undertake an experiment of opening the provision of terminal equipment—phone instruments, answering machines, PBXs—to competition, subject only to certification of compatibility of competitive equipment with its own network: prevailing policies still permitted the franchised local monopolies to require (and charge prices far above cost for) ''protective connection devices'' between ''alien attachments'' and the interface with the telephone network.[1] The purpose of the experiment was to see whether, as some exuberant Bell representatives claimed, removal of that protection might end up electrocuting

* I thank Dennis Weisman and Tim Tardiff for their extremely careful reading of this manuscript and suggestions for its improvement.

1 See the discussion and criticism of the Hushaphone and Carterphone decisions in Faulhaber (1987, 27–30) and in Kahn (1971, vol. 2, 140–145).

telephone linemen. The notion, however, that other parts of the business could be opened to competition or deregulated still seemed remote. This was true even of long distance service, despite the FCC's decision in 1959 to permit use of the spectrum above 890 Mc for private microwave systems, its interminable proceedings thereafter concerned with defining the acceptable response by AT's to that competition, and its 4–3 decision in 1969 to approve MCI's provision of private carriage alone.[2] The only major active deregulatory issues at the national level related to the field price of natural gas, stock exchange brokerage commissions and whether cable systems should be relieved of the prohibition of their importing over-the-air distant signals, imposed in the interest of protecting local broadcasters (Kahn 1971, vol. 2, 193–209 and Noll et al. 1973, chapter 6).

My preponderant interest, instead, was in applying to electric, gas, telephone and water system charges the elementary economic principles I had previously enunciated in my first volume.[3] Meanwhile, the Kennedy hearings on the airline industry, directed by a Harvard law professor, Steven Breyer, were only just getting underway. Although I testified at them, I was of course totally unaware of how, within fewer than three years, they would change the course of my life.[4]

2. The Continuing Tension Over the Proper Role of Government in the Deregulated Airline Industry

It seems fitting to use the airline industry as my first illustration of ''the precarious balance between regulation and deregulation.'' I have elsewhere described how what we intended to be a gradual process of deregulating the airlines soon took on a life of its own, like the proverbial snowball rolling down a hill—the mirror image of the tendency of regulation, once undertaken, to become increasingly pervasive and thoroughgoing.[5] Unsurprisingly,

---

2 See Kahn (1971, vol. 2, 129–136). Also Faulhaber, chapters 2 and 3. MCI's defiance of the Commission's prohibition of its offering public switched service was not sustained in the courts until 1977 (MCI v. FCC, CCA, DC Circuit, July 28, 1977).

3 See the ''Introduction: A Postscript, Seventeen Years After,'' to the 1988 reprinting, especially its survey and interim assessment of ''The Deregulation Revolution,'' and ''The Anomalies and Distortions of Partial Deregulation.''

4 This is an apt occasion to acknowledge my debt to Walter Adams for calling my—and the world's—attention 20 years earlier to the similar cartelization of the trucking industry, largely in his work for the Senate Small Business Committee. See the discussion and references *ibid*., vol. 2, esp. 14–18 and 186–193.

5 See the reference to Kahn 1971 in note 3, above, a discussion that documents the ''inherent tendency to what might be termed service inflation, in which an equilibrium of cost and price is achieved not by reducing price to marginal cost but raising marginal cost to price'' (207).

> The typical answer of foreign governments to asserted excessively low load factors, namely the imposition of direct limitations on the amount of capacity offered, market by market, provides yet another illustration of the inexorable tendency for regulation of a competitive industry to spread. Control price, and the result will be an artificial stimulus to entry. Control entry as well, and the result will be an artificial stimulus to compete by offering larger commissions to travel agents, advertising, scheduling, free meals, and bigger seats. The response of the complete regulator, then, is to limit

therefore, when and as deregulated airline markets began to demonstrate the consequences of severe imperfections—

1. $13 billion or so of losses industry-wide during the 1990–1993 period—greater, the industry proclaimed, than the totality of its profits since the Wright Brothers' historic achievement in Kitty Hawk, North Carolina—demonstrating all of the symptoms of destructive competition, consequent on the splurge of capital expenditures in the late 1980s and the subsequent deceleration in the growth of demand;
2. the general deterioration in the quality of the air travel experience and especially extreme congestion at particular times and airports in very recent years; and, finally,
3. the extreme profusion and differentiation of fares, reflecting in important measure differences in marginal costs but also genuine price discrimination, reflecting some degree of monopoly power, particularly at increasingly prevalent hubs dominated by single carriers—it predictably and understandably generated increasingly intense pressures for governmental intervention. I confine my attention to the last two; the first has for obvious reasons come to seem much less urgent than it was a decade ago.

In principle, those proposals have fallen into two very different categories—some (coming particularly from parties who were never enthusiastic about deregulation) clearly constitute proposed reversals of the historic decision in 1978—putatively partial but, all experience tells us, certain to be driven in the direction of totality—others, at least in intent, involving no more than an application of antitrust principles to preserve the competition that is the essential substitute for regulation.

But of course, as every student of antitrust knows, that important distinction, simple and unexceptionable in principle, nevertheless in large measure merely shifts the nominal terms rather than the substance of the debate: to what extent do proposed interventions, under the rubric of preserving competition—the very rationalization of comprehensive regulation in the first place—actually involve repressing it in the interest of protecting competitors?.

I propose to offer two examples, knowing that for this audience little elaboration will be necessary.

2.1. Quality of Service Generally and Congestion in Particular

About the deterioration in the quality of airline service generally—the longer lines, more crowded planes—thanks largely to the 20% point increase in average load factors and narrower seating—I can be very brief. Under tight price regulation, travelers were confronted with only one choice—generally good (indeed, inefficiently good[6]) service but

---

advertising, control scheduling and travel agents' commissions, specify the size of the sandwiches and seats and the charge for inflight movies. Each time the dyke springs a leak, plug it with one of your fingers; just as a dynamic industry will perpetually find ways of opening new holes in the dyke, so an ingenious regulator will never run out of regulatory fingers (1978b).

6 See footnote 5, above.

at cartel prices; the purpose of deregulation was, by unleashing price competition, to offer economy-minded travelers an alternative low price and necessarily correspondingly lower quality service; and in that it has dazzlingly succeeded.

In contrast, the intolerable congestion at major airports, either at particular times or chronically, is probably the greatest single black mark on the industry's performance. The temptation is to characterize it, correspondingly, as the single greatest failure of deregulation.

Obviously, it is not the latter: on the contrary, it is the strongest possible vindication of the expectations of the deregulators—expectations of an accentuation of price competition and an elastic response of the traveling public to the sharp consequent reductions in average fares. The failure—the only failure—was that of governmental agencies to respond to the predictably increased demand for infrastructure—specifically, air traffic control and airports—and to price it correctly—a blame that I can definitively deflect from myself (and from deregulation itself), since I explicitly emphasized to the FAA the importance of their fulfilling that function at the very time we were undertaking to deregulate (Kahn 1978a).

I do not have to make the case here that the efficient remedies are and always have been reform of our organizations for providing those infrastructure services—removing them from the regular agencies of government and the federal budget, whether by corporatization or privatization—and the introduction of marginal cost pricing by a combination of congestion charges and auctioning of slots.[7] The only amplifications worth making in the context of this talk are to rebut the naïve assumption—of which surely none of you will be guilty—that congestion charges will produce an additional burden on airline customers; to document the danger that this abominable performance may lead instead to ''solution'' by reregulation, rather than by efficient pricing; and to consider the ways in which antitrust protections may be required.

Consumer organizations rarely put efficient pricing high on their list of proposed remedies, evidently partly for fear that the result will be higher prices for scarce services. But average fares are, unsurprisingly, already higher on flights to and from slot-constrained airports and on peak than off-peak, reflecting the comparative availability of discount tickets. Efficient congestion-based charges—or large proceeds from slot auctions—at times of excessive congestion will extract those rents from the carriers, rather than raise fares. Small wonder the favored carriers oppose them, preferring instead to assume responsibility for tailoring operations to capacity cooperatively, under grant of antitrust immunity![8] The other side of the coin would have to be arrangements assuring that those rents would, in turn, be used for some combination of increasing capacity and subsidizing—in the form of negative charges, if necessary for revenue neutrality—usage at off-peak times or uncongested auxiliary airports. Efficient charges of this kind would further realize the success of deregulation itself (and, incidentally, also of the bumping rules enacted by the CAB) in offering travelers a low-fare/low-quality (or lower-convenience) option denied them under regulation.

---

7 See, for example, Levine (1969).

8 For a thorough criticism of this proposed ''solution,'' see Robyn (2001).

As to the second, it is difficult to do justice, summarily, to the flood of proposals for legislative enactment of ''travelers' bills of rights'' engendered by the clearly unsatisfactory air service to which they have been exposed in recent years, particularly when, at peak travel times, adverse weather conditions bear principal responsibility for the most dramatic deteriorations. As the one who, inspired by the late Julian Simon, persuaded the CAB to adopt its present bumping rules under our authority to prohibit unfair or deceptive methods of competition, I would be among the last to oppose legislation banning ''deceptive scheduling of flights,'' requiring the publication of on-time performance or requiring (or, as I would prefer, requiring some administrative agency to require) compensation of aggrieved passengers in the event of specific failures. Yet the tendency of consumerists to prescribe such rules; to assign the principal blame to the scheduling practices of the carriers—as though they have any choice individually but to schedule their flights, to the maximum extent possible, when and where customers want to travel—rather than to inefficient pricing of infrastructure services; to prescribe preferential allocations of scarce facilities to favored carriers—low-fare or regional; indeed the hostility some of their proposals betray to efficient pricing ''as a way for the FAA to raise additional revenues rather than reducing congestion delays,'' to quote one of them—all reflect a bias in favor of regulatory rather than efficient competitive solutions.

There remains to be considered the possibility that the use of congestion pricing or slot auctioning may be distorted by market power. This is clearest in the case of auctioning: the value of a slot is likely to be systematically higher for the bidding carrier for whom it preserves a dominant position and the accompanying monopoly power than for the carrier to whom it offers only an opportunity to compete with an entrenched rival. The response, in principle, would be that it is precisely the purpose of the antitrust laws, and of section 7 of the Clayton Act in particular, to prevent such agglomerations of assets. Enforcement of that protection through the regular antitrust channels—involving ultimately adjudication by the courts—is both highly uncertain (for good economic reasons, no doubt) and cumbersome. For that reason, the antitrust agencies themselves use essentially administrative means not only, necessarily, to decide what mergers they will attack but, with increasing frequency, to attach structural conditions—typically stipulated divestitures—to an agreement not to prosecute. So the most efficient way of counteracting any such distortion of the bidding process would be to set some limits—say, 20 or 25%—to the share of the total slots for which any carrier could successfully bid.

Without some such protection, congestion pricing would seem to be subject to the same distortion as slot auctioning: as the price at times and places of excess demand rises to whatever level necessary to reduce congestion to optimal levels, presumably those carriers for whom use of the facilities protects a monopoly or dominant position would be the ones who would not drop out. The only answer I can think of would be to impose a similar limitation on the permissible share of any single carrier. To the foreseeable reaction that such a preventive would be inherently regulatory, the answer can only be that it is no more so than section 7 of the Clayton Act and the spirit of the antitrust laws generally, the importance of which is of course enormously increased by the abandonment of direct regulation itself.

These suggestions raise the familiar dilemmas of all such limitist solutions to antitrust issues—in the present instance, the fact that hubs have characteristics of natural—as well

as unnatural—monopoly[9] and that, as Dan Kasper has urgently called to my attention, such limits could well weaken the important competition on long routes among network carriers over their respective hubs.[10] The dominant carrier may well be able to outbid its rivals for the scarce slots or better pay the congestion-based fees because of the superior economies of scope that usage of airports at those times permitted it to exploit.

2.2. Extreme Price Differentiation

> It costs $352.50 for a round-trip coach ticket between Boston and Washington, a distance of 406 air miles. But if you want to go to London, 3,267 miles away, it costs as little as $298—provided you meet the restrictions.[11]

In my response to this complaint, I compared the $352.50 full Boston/Washington fare with the weighted average round trip fare for that route of $226.22, and the $298 Boston/London fare with the range of $401, $506 and $1,587 Boston/London, respectively, for the excursion fares, low and high seasons, and the regular tourist fare.[12]

These differentials are, of course, in important measure not discriminatory at all, but reflect differences in marginal costs; they are in large measure also definitely discriminatory but welfare-enhancing, in an industry characterized by extreme economies of scale and, even more, scope—benefiting the ''victims'' as well as the beneficiaries; but they create a need for the best possible guarantees that the demand-inelastic customers are in fact not being exploited—that is, are not being forced to pay more than the costs of serving them on a stand-alone basis (Baumol et al. 1988, 508).

There are only two generic ways of attempting—and of course the word ''attempting'' requires multiple emphasis—to provide such protection, regulatory and competitive. The first could involve ceilings of estimated stand-alone costs or fully allocated costs or setting some arbitrary maximum to the spread between average and full fares—all of which have been advocated at one time or another. The other would require doing everything possible to strengthen the discipline of competition: perfect contestability is the perfect preventive of rates to any group of customers exceeding stand-alone costs (Baumol et al. 1988, 352–354; 508–509).

My rejection of the regulatory solution is not merely aesthetic or an attempt to exorcise the hobgoblin of inconsistency; it is informed rather by experience and a recognition of the logical and practical impossibility of partial re-regulation. As for the former, the contestant parties in proceedings attempting to set ceiling rail rates for captive shippers under the

---

9 Kahn (1993, 381–387). See also Levine (1987).

10 For example, if such restriction applied at ORD [O'Hare], both AA and UA would be severely handicapped in competing against carriers such as NW and DL whose competing hubs are relatively unconstrained. In addition, it is clear that hubbing carriers at efficient hubs typically use well over 25% of the operations, even at unconstrained hub airports. Thus, the proposed limits would effectively eliminate hubbing if applied to hub airports. Another reason to push hard for airport capacity expansion. (Personal communication.)

11 Kuttner (1988). Undoubtedly more extreme examples could be cited today.

12 Kahn (1988).

provisions of the 1980 Staggers Act have accepted the proposition that stand-alone cost ceilings are theoretically correct but almost hopelessly difficult to administer, with the result that the captive shippers clause itself has provided little protection and is only rarely invoked these days.[13] This experience has suggested to me and, I believe, to others, that something like 180% of actual long-run incremental costs—measured in an approximate way by the railroads' so-called Form A costs—would make much more sense, pragmatically. An additional reason for rejecting any genuine attempt to measure those stand-alone costs has been the bitter experience with the FCC's prescribed standard for pricing unbundled network elements provided by incumbent to competitive local exchange companies—the total long-run incremental costs of a hypothetical competitor writing on an almost completely blank slate and employing the most efficient available technology—rather than the actual LRICs of the incumbent companies—to which I turn in part B, below.

My second reason for rejecting the regulatory protection is the tendency to which I have already referred and that I have already documented in other contexts, for ''regulation, once undertaken, to become increasingly pervasive and thoroughgoing.'' (See note 5, above.) Despite the unfortunate tendency of the discomforts and inconveniences of the low-fare/low-quality service the emergence of which, as I have argued, is one of the proudest achievements of deregulation, to spill over and sully the service of the travelers who pay full fares, the airlines have over the last several years been attempting to differentiate the quality of the two—with automatic upgrades to first class, separate lines at the check-in counter and extra frequent flyer awards for the full-fare passengers. Predictably, any attempt by regulators to put ceilings on those higher fares will have to be accompanied by corresponding, increasingly detailed floors under quality—the percentage of times in which the payment of full tourist fares is indeed accompanied by upgrades, the respective lengths of lines at the check-in counters, the quality of their food service: there is no halfway house.

This leads, as the only logical complement of deregulation, to greatly increased emphasis by both the antitrust agencies and the Department of Transportation—which has the same kind of authority to prosecute unfair or exclusionary methods of competition as the Federal Trade Commission has for industry generally—to protect and improve the conditions of competitive entry.

This means to me, in turn, that the great majority of economists who are of the opinion that airline deregulation has been a success and is worth preserving rethink the highly skeptical view of many if not most of them about the reality of predation.[14] All it takes, in

---

13 The major exception, I am informed, is for shipments of coal moving on essentially stand-alone railroads. See my citation of this experience in my recognition of the theoretical correctness of the stand-alone cost ceiling and my rejection of the FCC's similar prescribed blank-slate pricing of unbundled network elements in my *Whom the Gods Would Destroy* (2001, 61–62), and in part 2, below.

14 See my fuller and latest discussion of this issue (*ibid*., 36–39) and especially lengthy accompanying footnotes (66–70). The day before I first dictated this discussion, there appeared the following in the *New York* Times (2001):

> Early last month, Canada's Competition Bureau, a government antitrust watchdog, asked a federal court to order Air Canada to stop pricing fares below cost in Eastern Canada. In February, for

my opinion, is that DOT and antitrust agencies recognize the soundness of marginal opportunity (rather than only marginal production) costs as the theoretically correct floor of non-predatory prices[15] and solve the administrative (and essentially regulatory) pitfalls of a government agency attempting to apply such a standard[16] by leaving that responsibility to the responding carriers themselves, by requiring them, if their response is indeed followed by withdrawal of the challenger, to maintain their responsive levels of capacity and fares for some period of time—say two years. That would tend to ensure that an incumbent carrier would not lightly undertake a response that was in its own opinion profit-sacrificing (and therefore putatively predatory) in the expectation of being able to withdraw it if it succeeds in driving out the challenger; and, at the same time, give travelers the continuing benefit of the newly introduced competition for some substantial period of time, rather than permit its quick withdrawal.[17]

---

example, Air Canada, which had been charging about $400 for one-way fares between Halifax and Montreal or Ottawa, suddenly dropped fares to around $65, matching those of the six-month-old CanJet.

15 See references to the unanimous agreement with this proposition in the Report of the Transportation Research Board/National Research Council (1999), in my *Whom the Gods Would Destroy*, note 79 (69–70).

16 All the members of the afore-mentioned TRB/NRC committee explicitly recognized these difficulties, despite their roughly 50/50 division on the merits of the Department of Transportation's initiative to proceed against ''unfairly exclusionary'' pricing.

The recent decision of the U.S. District Court granting American Airlines summary judgment in the predation suit brought against it by the U.S. Department of Justice, finding that the plaintiff had failed to demonstrate that the airline's entry-responsive fares fell below average variable *production costs*, rejected the Department's proposed alternative standard—namely whether the carrier had by its particular response sacrificed profits that it might have earned by an alternative response—on the understandable ground that any attempt to determine whether ''the response at issue involved the carrier's forgo[ing] more revenue than it otherwise would have made . . . would be a litigation nightmare . . . and would invite excessive speculation about reasonable alternative strategies and the like . . .'' U.S. District Court for the District of Kansas, *Memorandum and Order*, 2000, *U.S.A. v. AMR Corporation et al.*, Internet version (2001, 49–51).

17 The District Court in the American Airlines case also committed the logical error—emphasized by its own use of italics—of defining successful predation as the restoration of pre-competitive entry levels only if and to the extent that those levels were themselves super-competitive:

> In sum, the government's evidence establishes only that, after decreasing during a period of low fare carrier competition, fares of four of the core routes. . .returned to approximately the same levels as before. *But there is no evidence that the prior fares were in fact supra-competitive*. (*ibid*., 57).

See also the conclusion on p. 62, for the same standard. If, however, one conceives of the function of competitive entry as one of subjecting those previous fares to a test of competition, clearly a highly discriminatory, pinpointed response that succeeds, by driving out the challengers, in restoring those previous levels will have frustrated that competitive test. The District Court's standard would force on the government the highly regulatory responsibility of determining whether those previously prevailing levels were indeed excessive. On this and other related tests, see Kahn (2001, note 79).

3. The Apotheosis of the Tar Baby: Telecommunications ''Deregulation''

Despite the difficult antitrust issues that they continue to raise, deregulation of the transportation industries—as well as crude oil, natural gas, stock exchange brokerage fees, cable television and wireless telephony—has been comparatively simple. It was comparatively easy for us to achieve the measure of success that I proclaimed some time after assuming chairmanship of the CAB-namely, ''the extent to which there would be no such agency when I was through'' or ''getting those eggs so scrambled that no one would ever be able to unscramble them.'' The most complete exemplification of the tar baby effect has been and continues to be telecommunications—25 years after the courts sustained MCI's defiance of the FCC's prohibition of it offering universal switched long-distance service and five years after Congress proclaimed the goal of deregulation. The reason, of course—both here and in electric power—has been the persistence of elements of natural monopoly that provided the justification—and/or rationalization (Gray 1942, 280–303)—for their treatment as public utilities in the first place.

Since I have—in some instances in collaboration with Timothy J. Tardiff and Dennis L. Weisman,[18] including substantial testimony on behalf of the local Bell telephone companies—sharply criticized at length the FCC's execution of its authority under the 1996 Act, it would come close to malpractice for me to use this occasion to repeat those arguments at any length. I will content myself therefore only with the following summary observations, first about its pricing prescription and second about its identification of network elements meeting the statutory standards for mandatory sharing.

As for the first, I find simply incomprehensible the Commission's adoption of the ridiculous blank slate version of TELRIC for the mandatory pricing of network elements—all the more so in view of the fact that I (sometimes in collaboration with Timothy J. Tardiff) and Jerry Hausman, separately, had emphasized directly to them the conceptual fallacy that it embodied. The literature had—to my satisfaction!—established the sufficiency of a requirement that they make network elements available at the actual long run incremental costs of the incumbents, with or without a markup approved by their state regulators sufficient to recover their historically determined revenue entitlement, subject to what I have termed the requirement of competitive parity and Baumol and Willig efficient component pricing (Baumol and Sidak 1994, 171–202, Kahn and Taylor 1994, 225–40).

Taken together with the Commission's totally expansive criteria for identification of the network elements subject to mandatory unbundling—reversed by a seven-to-one vote of the United States' Supreme Court—it is difficult to conceive of a standard more inherently contradictory of the Act's manifest and understandable desire to encourage competitive challenges by entrants constructing their own facilities than their ability to obtain such facilities instead from the incumbent at prices at the lowest level that they could conceivably achieve themselves.[19]

---

18 Kahn (2001, chapter 2); Kahn (1998); Kahn et al. (1999, 319–365).

19 Unsurprisingly, therefore, the only major facilities-based competition has taken place where prices were held by regulators outrageously above the incremental costs of incumbents and challengers alike in order

With Henry Higgins, I can only wonder or marvel ''what can have possessed'' the Commission's Eliza to have presumed in this way to stipulate what the *results* of the hypothetical competitive process would be: ''I cannot understand the wretch at all!'' The only answer I have been able to divine I have put in the title of my little book, *Whom the Gods Would Destroy* (2001).

My only other observation I claim entitlement to make as one of the few surviving students of Schumpeter. Whatever the case for the FCC's identification of ILEC network elements subject to mandatory sharing and its prescribed method of determining the price, it clearly becomes progressively attenuated the more such advantages as the ILECs enjoy from providing them are not simply the heritage of their previous franchised monopolies but flow from the undertaking of new, costly and risky investments, in an environment of technological and commercial uncertainty. Such a policy confronts investors with the asymmetrical prospect of being required to share the results of investments that turn out successfully, at regulated (let alone absurdly regulated) rates, while perforce absorbing the entire costs of the unsuccessful ones. The case turns unequivocally negative when, as the FCC has itself found with respect to the offer of broadband services generally and high speed Internet access specifically, the ''preconditions for monopoly appear absent''; the market is likely ''to accommodate different technologies such as DSL, cable modems, utility fiber to the home, satellite and terrestrial radio''[20]; and the market share of the phone companies is roughly half that of the cable companies, who are subject to no such sharing obligations.

In these circumstances, the FCC's insisting additionally on attaching to its authorization of the ILECs to do anything it is a position to deny—to merge, to offer interLATA services in region, to offer advanced, broadband services—no doubt to go to the bathroom if it could—its favorite condition, ''only if you accept the handicap of offering broadband services through a fully separated subsidiary''—seems, quite simply, anticompetitive.

Although as little as a year or so ago, I could not

> bring myself to deny the regulatory agencies a central role in the transition of public utilities to competition, in view of the special circumstances of those industries: the necessity for settling out and tracking the collection of strandable costs; the more pervasive possibilities in those industries of tying competitive to monopoly services, directly or subtly, and of cross-subsidization, strictly defined; the consequent need for

---

to cross-subsidize other services. The major historical instances of regulatory-mandated egregious overcharging were in (1) the retail charges for long-distance calling and then to independent long-distance companies for access to the local networks—leading to the emergence of independent long-distance access providers in every major metropolitan area in the United States—and (2) the retail charges for high-volume services to business users in those areas.

20 Report, *Inquiry Concerning the Deployment of Advanced Telecommunications Capability*, 14 FCC Rcd 2398, 2423–24, par. 48 (1999); *see also* Memorandum Opinion and Order, *Applications for Consent to the Transfer of Control of Licenses and Section 214 Authorizations from MedioOne Group, Inc. to AT&T Corp.*, 15 FCC Rcd 9816, 9866, par. 116 (2000); *see also* Seventh Annual Report, *In the Matter of Annual Assessment of the Status of Competition in the Market for the Delivery of Video Programming*, CS Docket No. 00–132, FCC 01–1, par. 52 (rel. Jan. 8, 2001).

> accounting separations and the monitoring of transactions between still-regulated utilities and unregulated affiliates; and, finally, the pervasiveness of essential facilities controlled by incumbents—not to mention competitive advantages deriving solely from their historical franchised monopolies, requiring an administrative agency to define them and to prescribe the terms and conditions of sharing (Kahn 2001),

I'm tempted to suggest a moral from the FCC's behavior: that the task of presiding over and establishing the necessary conditions for deregulating an industry should not be entrusted to the agency that regulated them.

4. The California Electric Deregulation Fiasco[21]

I had been promising myself during the last few months to determine a reasonable apportionment of responsibility for what has gone wrong in California better than the explanations one can read in any newspaper any day—to decide what conclusions can reasonably be drawn from it about the wisdom of that venture, both as a general proposition and in its major details, and to use this paper as the occasion for offering my insights to an eagerly expectant world. I have long since abandoned that ambition: there are too many others who understand far better than I the peculiar constraints imposed by the technology of electric power generation and transmission, and have articulated a far clearer vision of the *regulatory* policies and organizational arrangements essential to satisfy those constraints if competition is to be feasible.[22] Sheer cowardice, along with a recognition of my limitations, forces me to confine myself here to a series of questions and suggestions.

4.1. Should We Have Deregulated?

Although the question may be pointless at this stage—too many eggs have become scrambled; a generation sector that it still appears could be effectively competitive (even

---

21 My first and only exposure to the California experience resulted from my appointment by the California Power Exchange to the chairmanship of a ''Blue Ribbon Panel''—whose other members were Peter C. Cramton, Robert H. Porter and Richard D. Tabors—to assess proposals that the Exchange switch from uniform pricing to pay-as-bid. Our report, on January 23, 2001, has since been published in slightly altered form in the *Electricity Journal* (Kahn et al. 2001).

22 See Hogan 2001, in this issue, citing also the prescient discussion of Joskow and Schmalensee 1983; also Joskow 2000.

> The only involvements I have had with the radical restructuring of the electric utility industry have been testimony I have given on (1) the entitlement of the utility companies to recover the costs that, it appeared—indeed, it was fondly expected and hoped by some parties—would be left stranded by the introduction of competition and (2) the formulation of codes of fair conduct governing the anticipated competition at retail between the still-regulated local utility distribution companies or their affiliates and independent marketers. Even if the first of these appeared to involve many billions of dollars and the second was the subject of intense controversy, neither seemed to me of great importance compared with whether or how an electric industry, part deregulated and part, necessarily, still regulated, could work.

though it may not be in the present situation of extreme shortage) has now been separated out; any suggested reforms aimed at restoring regulation may well do more harm than good—it is still one that all of us must attempt to answer for ourselves. An answer would have to take into account, above all, the extraordinary and in some respects literally unique characteristics of this industry—the high degree of interdependence between investments and operations at the transmission and generating levels, hitherto presumably taken fully into account internally by the several vertically integrated franchised monopolists that dominated the industry[23] and by pools or bilateral arrangements that they found in their several interests to enter; and the especial need for integration created by the essential non-storability of power and the corresponding need for an instantaneous matching of generation and consumption, in which failure at any single point can threaten the stability of the entire interconnected grid. California has made especially clear how extraordinarily the effects of failures to satisfy these constraints are compounded by the highly inelastic demand on the part of most consumers, particularly because of the absence of any means of confronting them with the wide fluctuations in marginal supply costs on a real-time basis,[24] and the long-run inelasticity of supply—using long-run in both its popular, temporal, and technical economic sense.[25]

As a result, the industry is extraordinarily susceptible to severe consequences of forecasting errors. For at least the last 10 years, I have been pointing out that the deregulation movement in electric power has been essentially opportunistic: practically no one was calling for it during the 1950s and 1960s, when retail prices declined some 30% in real terms, or in the 1970s and early 1980s, when it appeared—and to this day in states where it still appears—unregulated prices would be higher than regulated; but deregulation became almost irresistible politically by the middle 1990s in those parts of the country where, it seemed certain, competitive prices would be far below regulated ones.[26]

While I, along with many others, recognized that the apparent opportunity for huge short-term benefits from deregulation was created by enormous forecasting errors in the 1970s—in assuming an indefinite continuation of demand growth at accustomed 6 or 7% annual rates; in estimating the costs of nuclear plants and in foreseeing what would in the 1980s and 1990s turn out to be the lowest-cost technology; in regulators' estimates of avoidable costs on the basis of which they were instructed by PURPA to set the prices the utility distribution companies were required to pay for independently generated power, as a result of all of which the industry, particularly on the East and West Coasts, found itself with what appeared to be a heavy load of high-cost excess capacity and a consequent wide spread between book costs and the estimated costs of new combined cycle plants powered

---

23 See, e.g., Kaserman and Mayo (1991, 483–502).

24 There is a story now to be told about the apparently unexpectedly large response of demand and its role in fending off the worst expected shortages in California in the summer of 2001.

25 See, among others, Rosen et al. (2000); and the brief, challenging ''Heresy? The Case Against Deregulation of Electricity Generation,'' by the pseudonymous ''Price C. Watts'' (2001, 19–24). See, in contrast, the confident exposition by Hogan (2001) of the policies of a still-regulated, naturally monopolistic transmission system sufficient to ensure efficient, socially beneficial competition in generation and distribution. Also Rowe et al. (2001).

26 See also Joskow (2000, 133–138, 142–143).

by unexpectedly cheap natural gas—the irony is that we failed to recognize that the generally accepted expectation of huge benefits for consumers from deregulation might similarly prove to have been based on the opposite forecasting errors. This time, it has been the growth in demand that has been underestimated; the persistence of excess capacity overestimated—especially when generators were vowing not to make their previous mistakes again!; the cost of natural gas and NO_X emissions underestimated; and the long-run inelasticity of supply either ignored—because of its seeming irrelevance—or its importance underestimated; and the importance of real-time marginal cost pricing and the short-run demand elasticity of which it is a prerequisite inadequately appreciated, despite our belief that we had learned that lesson in the inflationary 1970s.

In these respects, I have the clear impression, California was not unique: its experience was simply more extreme than elsewhere. In these circumstances, however, it is not surprising that observers in California seem to have been first to identify apparent exercises of market power because of the newly-discovered susceptibility of its spot markets to unilateral withholding by multi-plant producers of some comparatively small proportion of their capacity and the consequent apparent insufficiency of competitive disciplines at times of peak demand.[27]

4.2. Specific Errors or Uncertainties, in Hindsight

4.2.1. The Retail Price Freeze

Postmortems on California, many of them justifying Harold Demsetz's acute observation many years ago that ''believing is seeing'' (Demsetz 1974)—brought to my

---

27 Borenstein and Bushnell (1999); Borenstein et al. (2000); Joskow and Kahn (2001). Our Committee (see note 21, above, recognized the possibility that the Power Exchange's uniform pricing system—in which all bidders on the supply side received the marginal, market-clearing price—might have borne some responsibility for this phenomenon and the plausible consequent case for paying each supplier only what it bid on its offerings:

> This kind of perceived behavior has lent plausibility to the proposal to substitute pay-as-bid for uniform market-clearing prices. It is only the prospect, under the present system, of receiving on all their sales the benefit of the increase in market price caused by withholding some portion of their capacity that large generators can expect to profit from that practice—so goes the reasoning: under pay-as-bid, in contrast, such generators would have to bid the estimated monopolistically elevated price on all their proffered sales in order to reap those gains, at the immensely increased risk that some or all of those higher bids will prove to have been excessive and therefore be rejected, with a consequent loss of the entire difference between their actual marginal costs and the ultimate market price. The proposed change in the pricing method would, by this reasoning, therefore dramatically alter the balance of risks and potential gains.
>
> Just as the naïve expectation that a shift to pay-as-bid will produce a dramatic reduction in the average prices consumers pay ignores the certainty that generators will radically alter their bidding practices to frustrate achievement of that result, however, so here, the expectation that it would discourage monopolistic withholding fails to take into account the ways in which bidders will respond by changing their bidding behavior correspondingly. If and to the extent that monopolistic withholding has occurred in the past, bidders would henceforward, under pay-as-bid, attempt to predict the consequent behavior of the market prices in their several bids and, to the extent they succeed, the anticipated benefits of the change for consumers will prove to have been illusory.

attention by John Kwoka (Kwoka 2001)—have readily identified California's particular mistake as deregulating only partially. Clearly, the mandated reduction and freeze of retail prices, accompanied by deregulation of wholesale prices, has proved catastrophic, initially to the distribution companies but then, inevitably, to consumers and taxpayers called upon to bail them out. Yet the freeze must have seemed at the time a logical part of a pragmatic resolution of the essentially political problem of providing the companies reasonable assurance of recovery of their stranded costs (to which I, as a former practitioner of prudent investment regulation, felt constrained to acknowledge their entitlement), which was apparently their almost exclusive concern, in exchange for giving ratepayers comparable transitional protection. Indeed for the first couple of years it worked unexpectedly well: wholesale markets were weak and the companies were able to use the expanded margins between acquisition cost and their frozen rates to recover their strandable costs ahead of schedule. How, nevertheless, their managements could have blithely, in effect, bet their companies on the expectation of continued soft wholesale markets, without hedging, is itself difficult to comprehend—at least in hindsight. Moreover, as several have observed, the unexpectedly high prices at which they succeeded in selling off their generating plants might or should have suggested that those expectations were probably wrong.

It may well be that setting aside this particular aspect of the deregulatory design in California—and the deplorable populism and misguided consumerism of promising the public the benefits of unleashed competition while shielding them from its risks[28]—is tantamount to asking, ''Other than that, how did you enjoy the play, Mrs. Lincoln?'' It seems to me necessary, however, to ask whether other aspects of the design may also have been at fault.

4.2.2. Vertical Divestiture

The apparent success of the mandatory divestiture of AT's putatively naturally monopolistic local service from its potentially competitive long distance service[29] has clearly encouraged legislators and regulators to prescribe similar protections in other industries in process of deregulation: observe my criticism, above, of the FCC's infatuation with fully separated subsidiaries and the reluctance of the Department of Justice to give up the interLATA prohibition as the RBOCs seek its lifting (Kahn 1998, 66–69).

In the circumstances, it has seemed natural for several of the state commissions

---

28 Severin Borenstein informed me that the statutorily-mandated 10% reduction in residential rates, which went into effect on January 1, 1998, was not reversed until three years later, and while increases averaging 18% went into effect in May, about one half of all households were to see no increase.

29 But see Crandall (2001). See also the citation of evidence from Canada and Japan that competition in long-distance service had progressed more rapidly in those two countries than in the United States, even though neither of them had required vertical separation of long-distance and local service, in Grieve and Levin (1997); and, similarly, by Spiller and Cardilli that facilities-based *local* competition had progressed at a healthy pace in the smaller countries they examined (Australia, Chile, Guatemala and New Zealand), even though none of these countries has the extensive unbundling requirements for an indefinite duration that prevail in the United States or has prevented incumbents from vertically integrating (1997, 127–138).

similarly to require severing of the ties between electric generating companies and both their transmission and local distribution networks—in order to ensure equal, non-discriminatory access of competitive generators[30] and marketers to those essential facilities. Both California and New York have taken the divestiture route, whether mandatorily or by the offer of sufficient inducements—in part, additionally, in order to secure a market estimate of stranded costs.

I have wondered whether this effectively mandatory divestiture may have contributed to California's problems by denying it one of the main benefits of vertical integration itself—the incentives it provides for undertaking investments at one horizontal level necessary for the success of operations at another. Continued ownership of generation capable of satisfying a larger proportion of their needs would of course have protected the major distribution companies that had accepted the obligation to freeze their retail rates. But whether it would have produced more generation capacity sooner is questionable: there seems to be more than sufficient interest of independent generators in building new plants, as evidenced by the huge number of pending applications for licenses: the assumption that competitive entry into generation is or should be a sufficient protector of consumers—in the long run!—seems to have been borne out.

Whether the same complacency is justified at the transmission level is a question crying out for an answer: when I read in the *Wall Street Journal*—in the news section, mind you, not its editorial pages—that there is inadequate incentive in California to construct new transmission facilities (perhaps the reality would be better conveyed by substituting for ''construct'' ''run the gauntlet of obtaining the necessary permissions to construct''), I cannot but raise the question of whether the almost automatic recourse of some regulators to MFJ-like remedies—apparently Pennsylvania and Illinois have been outstanding, successful exceptions[31]—was wise, especially in an industry in which the vertical interdependencies are so extraordinarily close. At the very least, the California deregulators—in contrast with those in New York, Illinois and Pennsylvania—cannot escape blame for the compulsive tidiness and high propensity to micro-meddle that led them to require utility distribution companies to purchase all their power in the spot market—surely an excessively regulatory accompaniment to what purported to be a process of deregulation. As is widely recognized, distribution companies that adopted the commitment to freeze their rates should, in any sensibly devised system, have had available to them the alternative quasi-integration option of entering into long-term supply contracts, which would similarly have insulated them from what proved after the fact to have been a gigantic risk.

And of course the enforced artificial concentration of purchases and sales on the spot or quasi-spot markets was directly responsible for the immediate explosive effect of the suddenly emerging shortages at times of peak demand on not just the profits of the companies whose retail prices were still frozen—profits almost instantaneously more than wiped out—but also on the rates of the customers unprotected by such freezes, and

30 See Joskow (2000), 128–130.

31 See the references to the healthy progress of Illinois and Pennsylvania—neither of which has required divestiture—in developing new generation capacity in Rowe et al. (2001, 12, 15–16).

inevitably, within a very short time, on the ''protected'' customers as well. It also created the opportunity to which I have already referred for generators with individual generating units large enough to make the difference between adequate and inadequate supplies—in markets characterized by highly inelastic supply and short-term demand—dramatically to lever the market price upward by withholding that output.

4.2.3. Misguided Populism

One widely cited explanation for the prohibition of utility distribution companies acquiring significant proportions of their supplies with long-term contracts was the fear I have seen expressed that they would ''muck it up'' [sic]. That would seem to be a reasonable concern so long as it was contemplated that those companies would continue to operate under the old cost-plus regulatory system. But surely the price freeze to which they were subjected during the transition and the effective competition to which it was anticipated they would be subject thereafter were the proper preventives, remedies or punishment for irrational contracting on their part. Moreover, this concern surely came with ill grace from regulators and strong advocates of independent, non-utility generation who had themselves ''mucked it up'' by forcing on the utility companies purchase prices in their contracts with qualifying facilities under PURPA wildly above avoided costs—from the perspective of the last two decades of the century but no longer!

It appears also the utility distribution companies may have had some legitimate basis in opportunistic behavior of their regulators in the past for a reluctance to avail themselves of the opportunity to hedge their obligation to serve with long-term purchase contracts. As our Committee described that experience:

> The IOUs were able to purchase via forward contracts through the CalPX—but only through the PX—after it first made block forward contracts available in July of 1999. Although there are regulatory limits on the amount of forward positions that each IOU may take, they have not reached those limits, in availing themselves of this opportunity. They may have been discouraged from doing so by their past experience with forward purchases in California, in both gas and electric, under which regulators forced them to absorb any losses stemming from the contract prices exceeding wholesale market prices while not being permitted to reap the benefits, when the contractual prices were lower.[32]

Another explanation of the prohibition of long-term contracts proffered to our Committee was the fear expressed by consumer organizations that such contracts would enable large, business customers to ''skim the cream'' of the benefits available from competition, at the expense of small users. Of course, even if this concern were legitimate, it would justify prohibiting the distribution companies entering into long-term sales, not purchase contracts. The apparent purpose of the latter prohibition was, instead, to preserve equal competitive opportunities for independent generators by preventing the utility companies that continued to own generating plants from favoring those affiliates in obtaining their supplies. Since, however, the independent generators—including purchasers of divested

32 Kahn et al. (2001, original report, footnote 15).

utility plants—were subject to no such requirement to sell exclusively through the organized exchanges, the effort in this meddlesome way to prevent ''cream skimming'' was probably futile, as well as—by denying distribution companies the ability to hedge their risks—injurious.

One final manifestation of misguided, albeit understandable, populism to which I have already referred—the insistence on conferring on consumers the anticipated benefits of deregulation while also protecting them from possible injury from exposure to the vicissitudes of competitive markets—was by retaining the obligation of the utility distribution companies to serve as the suppliers of last resort at regulated prices. The aforementioned Mr. Watts points out that distribution companies cannot enter into long-term purchase contracts so long as they remain subject to the supplier of last resort obligation, because it involves playing a game of heads-we-lose, tails-somebody else-wins: if the terms of the contract prove to be unfavorable—that is, if open market prices fall below the contractual level—they lose all their retail customers and are stuck with the difference between the contractual price and open market value of the power; if and when instead the contract proves favorable, retail customers either remain with them or return to them, in which event they merely break even, at best. If the facts are as he states them, one protective provision eliminated the possibility of another.

4.3. Rate Caps?

The reported increase in California's total electricity bill from $7.4 billion in 1999 to more than $27 billion in 2000 and, according to recent projections (subject to all my preceding admonitions about forecasts in this industry!), $50 billion or more this year, have understandably generated demands for the imposition of ceilings on wholesale rates—specifically, since the traffic is clearly interstate, for the Federal Energy Regulatory Commission to exercise its authority to establish ''just and reasonable'' rates.[33] This appeal—my advocacy of which some commentators have characterized as apostasy—elicited the immediate response, from the President on downward, demonstrating a firm command of the first week of Economics 101:[34] that it would not only do nothing to solve the fundamental imbalance of supply and demand but, particularly by discouraging the expansion of capacity, would exacerbate the problem. I can respond most efficiently in the form of a few pertinent facts and related propositions from the second week.

Both the supply and the demand for electricity, at times when the latter presses hard on capacity, are extraordinarily inelastic. Within the limits of installed capacity, supply is highly responsive to the wide fluctuations of demand from one moment, day or season to the next, as generating stations are automatically brought on line in rank order of their operating costs. When that limit is reached, however, it takes years to add to capacity. True, the elasticity response of demand in California has been prevented by the

33 See the letter initiated by Frank Wolak and signed by nine other economists, including me, calling upon FERC to exercise that authority (May 25, 2001).

34 See the scarcely less simplistic letter to the President of May 29, 2001 by Annelise Anderson, Martin Anderson, Michael J. Boskin, Steve Entin, Milton Friedman and James Gwartney and the much more sensible ''Dim Bulbs'' (Hazlett 2001).

ridiculously extended freeze in charges to residential customers, followed by grudgingly inadequate increases. Since, however, the extreme shortages, producing 10- and 20-fold increases in wholesale prices, have taken the form of extreme spikes at particular times and places,[35] and the overwhelming majority of customers do not have meters permitting them to be charged correspondingly fluctuating prices, the result has been very costly blackouts. It is a truism that blackouts have occurred because most retail rates have not been free to increase to whatever extent necessary to prevent them: what is frightening to contemplate is the extent of that ''extent,'' in the absence of real-time metering.

As against the minimal contribution of such exploding prices to an improved balance of supply and demand at such times[36] must surely be weighed not merely the income distributional consequences but the adverse macroeconomic consequences of generators extracting tens, indeed scores, of additional billions of dollars a year from consumers, businesses and government—much like the contribution to stagflation of the three-fold increases in the prices of crude oil nationally in 1973 and again in 1979–1980.

The spectacular historical instances of price controls doing more harm than good by interfering with the expansion of supply cited routinely by opponents of the proposed FERC initiative—notably on crude oil and natural gas in the 1960s and 1970s—have been ones in which regulation held prices below short- and long-term marginal supply costs (in the case of crude oil, the price of imports). The current electricity price cap proposals with which I am familiar are all intended to incorporate flexibility sufficient to accommodate incremental supplies at high short-run marginal costs, which are readily ascertainable and easily adduced by any generator in support of an exemption; and to ensure the presence of optimal levels of reserve capacity as well.[37]

There seems good reason to believe that the explosion of wholesale prices was not a phenomenon of pure competition alone, but reflected the not-necessarily-collusive or antitrust-law-violative withholdings of capacity at peak times, in order to lever up the market-clearing prices, in a process I have already described. In such circumstances, the rote repeaters of the litany that price controls always reduce supply might be interested in discovering, ceilings may actually result in expansions of offerings: with price caps there is no benefit in withholding supply, only sacrificed profits.

Interference with the fundamentally required correctives—of expanded capacity, on the

---

35 In response to a remonstrance by an independent generator, I am attempting to assess the possibility that in a deregulated system, such price spikes become the necessary mechanism for recovering the capital costs of their peaking units.

36 See note 35, immediately preceding.

37 In a preceding draft, I attempted to rebut the widespread assertions that ceilings inevitably conflict with long-run expansions of supply by citing the experience of the electric industry itself during the entire three-quarter-century, 1920–1995, when, if anything, cost-based regulation, as traditionally practiced, encouraged the very excess capacity that seemed during the past decade to promise such enormous benefits to consumers if rates were deregulated. Upon further reflection I see that this experience could be regarded as irrelevant in a world that had abandoned rate base/rate of return regulation of franchised monopolies, with its implicit guaranteed recovery of costs not imprudently incurred. My colleagues at NERA are helping me evaluate the efficiency of alternative arrangements for ensuring the recovery of the capital costs of capacity required to achieve optimal loss of load probability under the new restructuring arrangements. See also Hobbs et al. (2001).

one side, and intensified conservation, on the other—would of course be severely counterproductive. But where those supply-and-demand responses inevitably take time—to cite the most relevant example, at least a couple of years before additional generating capacity is likely to come on the market—any discouragements can readily be prevented by making the price caps (a) designedly temporary, automatically sunsetting within, say, two to three years, and/or (b) inapplicable to new capacity coming on line. Clearly, however, it is essential that they incorporate some such assurance.

5. Conclusion

The experience of the airlines, telecommunications and electric utility industries over the last 30 years amply documents the superiority of competition over comprehensive regulation. It also demonstrates that successful deregulation rarely consists in total laissez-faire. On the other hand, it amply documents the Tar-Baby effect: as I characterized it some 23 years ago,

> Each time the [regulatory] dike springs a leak, plug it with one of your fingers; just as a dynamic industry will perpetually find ways of opening new holes in the dike, so an ingenious regulator will never run out of regulatory fingers.[38]

In the case of the airlines, the greatest failures and sources of consumer complaint have been the increasing congestion and delays at particular times and places and the extreme differentiation of fares—in particular, the very sharp increase in full fares. The proper remedy for the former is neither governmental nor collaborative airline rationing of scarce air and airport space or promulgation of ''consumers' bills of rights'' but restructuring of the agencies responsible for providing the requisite infrastructure and pricing it correctly; and for the latter, not prescription of maximum fare differentials but intensification of efforts to provide fair competitive opportunities for challengers.

Something approximating total deregulation is clearly infeasible in telecommunications and electric power, essentially for the reasons that led them to be identified as regulated natural monopolies in the first place. At the same time, telecommunications provides the most extreme exemplification of the Tar-Baby effect, because of the demonstrated tendency of regulators to assume responsibility not merely for establishing the necessary conditions for efficient competition but for guaranteeing the desired results. The expansiveness of the FCC's definition of the elements of incumbent telephone company networks that it requires them to make available to entrants and its presumption in prescribing charges equivalent to the results that in its (mistaken) judgment would flow from efficient competition not only flatly discourage facilities-based competition but, as applied to new and extraordinarily expensive facilities, requiring costly and risky investments in latest technologies, conflict directly with the Schumpeterian requirements for dynamic competition.

---

38 Kahn (1978b, pp. 17–18).

Even more than in telecommunications, it is no accident that the proper characterization of the required regulatory reform in the electric power industry is characterized as ''restructuring'' rather than mere deregulation. Fully as much as the FCC's practices, however, the California experience illustrates the evils of deregulatory and regulatory opportunism, of excessively detailed design and prescription and misguided populism—promising consumers the anticipated benefits of competition while sheltering them from its risks. The central task is to design the inescapable residual regulation—centrally, of transmission—in such a way as, by the use of price, to ensure the same kind of efficient synchronization of investments and operations in generation, on the one side, and transmission, on the other, as were previously achieved within vertically integrated regulated suppliers; and, whether by explicit orders or organization of associated markets for capacity, ensuring reserve generating and transmission capacity sufficient for optimal system-wide reliability.

The only other prescription is that regulators be required, in keeping with current fashionable trends in national education policy, regularly to pass tests of their comprehension of Uncle Remus' ''The Wonderful Tar Baby.''[39]

References

Baumol, W. J., J. C. Panzar, and R. D. Willig. 1988. *Contestable Markets and the Theory of Industry Structure*. San Diego: Harcourt Brace and Jovanovich.

Baumol, W. J., and J. G. Sidak. 1994. ''The Pricing of Inputs Sold to Competitors.'' *Yale Journal on Regulation* 11(1): 171–202.

Borenstein, S., and J. Bushnell. 1999. ''An Empirical Analysis of Market Power in a Deregulated California Electricity Market.'' *Journal of Industrial Economics* 47(3): 285–323.

Borenstein, S., J. Bushnell, and F. Wolak. 2000. ''Diagnosing Market Power in California's Deregulated Wholesale Electricity Market.'' POWER Working Paper PWP-064, University of California Energy Institute.

Crandall, R. 2001. ''The Failure of Structural Remedies in Sherman Act Monopolization Cases.'' AEI-Brookings Joint Center for Regulatory Studies Working Paper 01–05.

Demsetz, H. 1974. ''Two Systems of Belief About Monopoly.'' In *Industrial Concentration: The New Learning*, edited by H. J. Goldschmid, H. M. Mann, and J. F. Weston. Boston: Little, Brown and Company, 164–184.

Faulhaber, G. R. 1987. *Telecommunications in Turmoil*, rev. edition. Cambridge: Ballinger.

Gray, H. M. 1942. ''The Passing of the Public Utility Concept.'' Reprinted from *The Journal of Land and Public Utility Economics*, XVI, February 1940. American Economic Association, *Readings in the Social Control of Industry*. Philadelphia: Blakiston: 280–303.

Grieve, W., and S. L. Levin. 1997. ''Telecom Competition in Canada and the U.S.: The Tortoise and the Hare.'' *Selected Papers from the 25th Annual Telecommunications Policy Research Conference*, Alexandria, VA, September 27–29.

Harris, J. C. 1881. ''The Wonderful Tar Baby.'' *Uncle Remus, His Songs & His Sayings: The Folk-Lore of the Old Plantation*. New York: Appleton and Company.

Hazlett, T. W. 2001. ''Dim Bulbs.'' AEI-Brookings Joint Center, *Policy Matters* 01–03, January.

Hobbs, B. F., J. Iñón, and S. E. Stoft. 2001. ''Installed Capacity Requirements and Price Caps: Oil on the Water, or Fuel on the Fire?'' *The Electricity Journal* (July): 23–34.

Hogan, W. W. 2002. ''Electricity Market Restructuring: Reforms of Reforms,'' Presented at CRRI's 20th Annual Eastern Conference, May 25, 2001, *Journal of Regulatory Economics* 21(1).

39 McKie (1970, p. 9) first used the term ''tar baby effect'' from the original story (Harris 1881).

Joskow, P. L. 2000. ''Deregulation and Regulatory Reform in the U.S. Electric Power Sector.'' In *Deregulation of Network Industries: The Next Steps*, edited by S. Pelzman and C. Winston. Washington, DC: AEI-Brookings Joint Center for Regulatory Studies, 113–188.

Joskow, P. L., and E. Kahn. 2001. ''A Quantitative Analysis of Pricing Behavior in California's Wholesale Electricity Market during Summer 2000.'' National Bureau of Economic Research Working Paper #8157.

Kahn, A. E. 1970 and 1971. *The Economics of Regulation* (2 volumes). New York: John Wiley & Sons. Reprinted by MIT Press, 1988.

Kahn, A. E. 1978a. ''Route Awards and Airline Scheduling Practices.'' Remarks before the Federal Aviation Agency (March 22).

Kahn, A. E. 1978b. Presentation before a Symposium on the Changing Environment of International Air Commerce, Georgetown University, Washington, D.C., May 4.

Kahn, A. E. 1988. ''Airline Deregulation: 'I Would Do It Again'.'' *The Washington Post* (September 25): C6.

Kahn, A. E. 1993. ''Competitive Consequences of Hub Dominance: A Case Study.'' *Review of Industrial Organization* 8: 381–405.

Kahn, A. E. 1998. *Letting Go: Deregulating the Process of Deregulation*. Michigan State University Institute of Public Utilities.

Kahn, A. E. 2001. *Whom the Gods Would Destroy, or How Not to Deregulate*. Washington, D.C.: AEI-Brookings Joint Center for Regulatory Studies.

Kahn, A. E., and W. B. Shew. 1987. ''Current Issues in Telecommunications Regulation Pricing.'' *Yale Journal on Regulation* 4(2): 191–256.

Kahn, A. E., and W. E. Taylor. 1994. ''The Pricing of Inputs Sold to Competitors: A Comment.'' *Yale Journal on Regulation* 11(1): 225–240.

Kahn, A. E., T. J. Tardiff, and D. L. Weisman. 1999. ''The Telecommunications Act at Three Years: An Economic Evaluation of Its Implementation by the Federal Communications Commission.'' *Information Economics and Policy* (December): 319–365.

Kahn, A. E., P. C. Cramton, R. H. Porter, and R. D. Tabors. 2001. ''Pricing in the California Power Exchange Electricity Market: Should California Switch from Uniform Pricing to Pay-as-Bid Pricing?'' A study commissioned by the California Power Exchange, January 23; published in amended form as ''Uniform Pricing or Pay-as-Bid Pricing: A Dilemma for California and Beyond'' in *The Electricity Journal* 14 (July): 70–79.

Kaserman, D., and J. T. Mayo. 1991. ''The Measurement of Vertical Economies and the Efficiency Structure of the Electric Utility Industry.'' *Journal of Industrial Economics* 39: 483–502.

Kuttner, R. 1988. ''High Costs of Flying.'' *The Boston Globe* (September 16).

Kwoka, J. 2001. ''California: The Perfect Electrical Storm.'' AAI Column in *FTC: Watch* (February 14).

Levine, M. E. 1969. ''Landing Fees and the Airport Congestion Problem.'' *Journal of Law and Economics* XII (April): 79–108.

Levine, M. E. 1987. ''Airline Competition in Deregulated Markets: Theory, Firm Strategy and Public Policy.'' *Yale Journal on Regulation* 4(2): 392-494.

MCI v. FCC, CCA, DC Circuit, July 28, 1977, *Utility Law Reports*, Federal, Paragraphs 12,000–12,024.

New York Times. (2001) ''Canada Plays Tough on Competitors.'' (April 22, V:3).

McKie, J. W. 1970. ''Regulation and the Free Market: the Problem of Boundaries.'' *The Bell Journal of Economics and Management Science* 1(1): 6–26.

Noll, R. G., M. J. Peck, and J. J. McGowan. 1973. *Economic Aspects of Television Regulation*. Washington: Brookings Institution.

Robyn, D. 2001. ''Congress' Plan to Reduce Flight Delays is Not Airworthy.'' AEI-Brookings Joint Center for Regulatory Studies, Policy Matters 01–15 (May).

Rosen, R. A., F. Sverrisson, and J. Stutz. 2000. *Can Electric Utility Restructuring Meet the Challenges It Has Created?* Boston: Tellus Institute.

Rowe, J. W., P. Thornton, and J. B. Szczypinski. 2001. ''Competition Without Chaos.'' AEI-Brookings Joint Center for Regulatory Studies, Working Paper 01–07, June.

Spiller, P. T., and C. G. Cardilli. 1997. ''The Frontier of Telecommunications Deregulation: Small Countries Leading the Pack.'' *The Journal of Economic Perspectives* 11: 127–138.

Transportation Research Board/National Research Council. 1999. ''Entry and Competition in the U.S. Airline Industry, Issues and Opportunities.'' Special Report 255. Washington, D.C.

U.S. District Court for the District of Kansas. 2001. *Memorandum and Order, 2000, U.S.A. v. AMR Corporation et al.* Internet version (April): 49–51.

Watts, P. C. (pseudonym of Michael Rothkopf). 2001. ''Heresy? The Case Against Deregulation of Electricity Generation.'' *The Electricity Journal* (May): 19–24.

IMPROVING THE ECONOMIC FOUNDATIONS OF COMPETITION POLICY

Timothy J. Muris[*]

I am delighted to speak at a symposium dedicated to my teacher, mentor, and colleague, Jim Liebeler. Every day I spend at the Federal Trade Commission is another step on a career path Jim opened for me thirty years ago. Antitrust anchored Jim's professional life, and he would have been proud to be celebrated by this gathering. Thank you for allowing me to honor him.

Today, I will address a subject that often occupied Jim Liebeler's formidable intellect: How can we improve the economic foundations of competition policy? Finding good answers to this question is vital to the success of antitrust enforcement. More than any other body of U.S. law, economics plays a central role in guiding courts and enforcement agencies about the proper design and application of legal rules.[1] If the economic foundations of antitrust analysis are infirm, competition law topples.

Policy discourse no longer focuses on *whether* economics should guide antitrust policy; that debate was settled long ago. The pressing question today is *how*. Which theories from the vast, diverse body of industrial organization economics should courts and enforcement agencies use to address antitrust problems? What hypotheses best explain business behavior in an increasingly complex and fast-changing business environment? How are economic ideas to be translated into operational rules?

In discussing how we can improve the economic foundations of antitrust, I will first identify several normative propositions about integrating economic ideas into antitrust policy. Part I, thus, focuses on the importance of regularly reassessing the economic assumptions of current policy, of distilling economic insights into workable rules and analytical techniques, and of doing empirical research to test the economic effects of judicial decisions and public enforcement activities.

Part II of the paper discusses the application to antitrust analysis of the New Institutional Economics ("NIE"). For antitrust, NIE teaches that the nature of industry organization (e.g., the type and extent of vertical rela-

---

[*] Chairman, Federal Trade Commission. Presented at George Mason University School of Law Winter Antitrust Symposium, January 15, 2003. This speech reflects the views of Chairman Muris and not necessarily those of the Commission or any other Commissioner.

[1] *See* William E. Kovacic, *The Influence of Economics on Antitrust Law*, 30 ECON. INQUIRY 294, 295-96 (1992) (describing features of the U.S. competition policy system that gives economists a major role in shaping antitrust rules).

tionships and the level of horizontal concentration) and of competition in a specific industry are not a black box to be analyzed only through the lens of industry structure and market power-based theories. A broader set of tools and presumptions is required to enforce the antitrust laws in the public interest. One of my messages today is that antitrust analysis, if done correctly, uses the NIE approach—that is, a careful, fact-based economic analysis grounded in a thorough understanding of the relevant institutions. Especially through its emphasis on transaction costs, NIE is a most promising strand of economic research, both for its theoretical elegance and for its ability to explain real-world phenomena.

Part III identifies issues for future research that could benefit antitrust policy, enforcement, and litigation. This section also highlights what the FTC, with one of the world's largest teams of industrial organization economists, is doing to advance empirical knowledge.

To provide some context for my remarks, let me acknowledge an intellectual debt. The origins of this talk reach back to the early 1970s when I enrolled at UCLA. There, I not only studied under Jim Liebeler but also began a lasting acquaintance with Armen Alchian, Harold Demsetz, and Ben Klein—all luminaries in UCLA's remarkable constellation of industrial organization economists.[2] As you will see, the echoes of their ideas have carried from Westwood to Washington.

I. INTEGRATING ECONOMICS INTO ANTITRUST POLICY: THREE NORMATIVE PROPOSITIONS

At UCLA, and during my professional life, I have learned three basic propositions about integrating economics into antitrust law that have deeply influenced my views about formulating competition policy:

* *Reassessment*. Antitrust agencies should engage in continuing efforts to assess the validity of existing hypotheses about the economic impact of business conduct in light of new developments in economic theory and the business environment.

* *Administrability*. The suitability of an economic hypothesis for shaping antitrust doctrine should be measured by whether the hy-

---

[2] Since leaving UCLA, I have enjoyed many opportunities to learn from and work with these economists. For example, when I was part of the University of Miami's Law and Economics Center, I had long discussions with Armen Alchian and worked extensively with one of Armen's pupils, Louis DeAlessi, who also taught at Miami.

pothesis lends itself to the development of standards that courts and enforcement agencies can administer effectively.

* *Empirical Testing*. The soundness of doctrine and enforcement policy over time depends heavily on the strength of empirical research that evaluates the economic effects of judicial rulings and enforcement decisions.

A. *Continuing Reassessment and Adjustment*

My professional career began in the mid-1970s at the FTC as an assistant to Jim Liebeler, who headed the Office of Policy Planning and Evaluation. The FTC was then expanding an already ambitious effort to reshape the American economy. Before I arrived, the Commission had begun cases to restructure the breakfast cereal, petroleum, and photocopier sectors.[3] To many observers, this was merely a good start.[4]

In August 1976, the Commission began a formal investigation of the automobile industry.[5] The decision to begin the highly publicized investigation followed an extensive preliminary inquiry by the agency's staff. In a critique covering nearly 100 single-spaced pages, the staff argued that there was widespread evidence justifying a fundamental restructuring of the U.S.

---

[3] Kellogg Co., [1970-1973 Transfer Binder] Trade Reg. Rep. (CCH) No. 8883 ¶ 19,898 (Apr. 26, 1972) (issuing complaint alleging maintenance of highly concentrated, noncompetitive market structure and shared monopolization in ready-to-eat breakfast cereal industry; seeking divestiture and other structural relief), *complaint dismissed*, [1979-1983 Transfer Binder] Trade Reg. Rep. (CCH) ¶ 21,864 (Sept. 10, 1981), *dismissal affirmed*, 99 F.T.C. 8 (1982); Exxon Corp., [1973-1976 Transfer Binder] Trade Reg. Rep. (CCH) No. 8934 ¶ 20,388 (July 17, 1973) (issuing complaint alleging agreement to monopolize and maintenance of highly concentrated market structure in petroleum refining; seeking divestiture and other structural relief), *complaint dismissed*, 98 F.T.C. 453 (1981); Xerox Corp., 86 F.T.C. 364 (1975) (consent decree imposing mandatory patent licensing and restrictions on conduct; settling complaint alleging monopolization, attempted monopolization, and maintenance of a noncompetitive market structure in dry paper copier sector).

[4] In the Fall of 1974, FTC Chairman Lewis Engman appeared before the Joint Economic Committee of the U.S. Senate to discuss the Commission's antitrust program. The FTC already had initiated the cereal and petroleum shared monopolization cases and the photocopier monopolization case against Xerox. For some committee members, these measures only scratched the surface of the industry concentration problem. Senator William Proxmire told Engman that "the FTC, like a number of other regulatory agencies, seems to concern itself with minor infractions of the law, and to spend much of its time on cases of small consequences." *Market Power, The Federal Trade Commission, and Inflation: Hearing Before the Joint Economic Comm. of Congress*, 93d Cong., 2d Sess., 58-59 (1974).

[5] *See FTC Focuses on Detroit*, BUS. WK., Aug. 16, 1976, at 62 (reporting FTC's opening of automobile industry investigation).

industry. The staff endorsed the horizontal and vertical dismemberment of the industry leader (General Motors) and indicated that the second and third members of the American "Big Three" (Chrysler and Ford) might be worthy candidates for divestiture as well.

A crucial analytical basis for the staff's critique was the simple market concentration doctrine—the belief that concentration and economic performance were closely and inversely correlated. Had the year been 1966, a Commission decision to embrace this belief would have been more understandable. In 1966, the view that high levels of concentration inevitably degraded economic performance commanded considerable academic support.[6] Many commentators saw the American automobile industry, dominated by General Motors for decades, as the paradigmatic example. Ten years later, however, there was serious reason for the FTC to doubt the validity of the simple market concentration hypothesis or to presume the invincibility of U.S. producers.

By 1976 the academic consensus condemning market concentration was crumbling. As discussed in more detail below, changes in economic theory and, more importantly, empirical research had undermined the simple concentration hypothesis. At least three specific developments concerning the auto industry also undermined the FTC's staff approach. The first was John McGee's 1973 article, *Economies of Size in Auto Body Manufacture*,[7] which persuasively argued that much of GM's success derived from its ability to spread the enormous costs of setting up dies to stamp out auto body parts across a much larger volume than its competitors.[8]

---

6 One of the most influential scholarly works of this period was Carl Kaysen's and Donald Turner's ANTITRUST POLICY: AN ECONOMIC AND LEGAL ANALYSIS, which appeared in 1959. CARL KAYSEN & DONALD TURNER, ANTITRUST POLICY: AN ECONOMIC AND LEGAL ANALYSIS 110 (1959). Kaysen and Turner wrote that "[t]he principal defect of present antitrust law is its inability to cope with market power created by jointly acting oligopolists." *Id*. They urged Congress to adopt new legislation compelling the deconcentration of various sectors of the economy. *Id.* at 110-19, 261-66. In 1969 a blue ribbon presidential task force, headed by Dean Phil Neal of the University of Chicago, recommended deconcentration variants of the Kaysen and Turner proposals. *See* WHITE HOUSE TASK FORCE REPORT ON ANTITRUST POLICY, *reprinted in* 2 ANTITRUST L. & ECON. REV. 11, 14-15, 65-76 (1968-69). Task force members who endorsed the deconcentration measure included such prominent academics as Dean Neal, William Baxter, William K. Jones, Paul MacAvoy, James McKie, Lee Preston, and James Rahl. A number of Task Force members who endorsed the deconcentration proposals later withdrew their support from such policies; *see, e.g.*, *Panel Discussion*, 54 ANTITRUST L.J. 31, 31-33 (1985) (discussing changes in the views of William Baxter).

7 John S. McGee, *Economies of Size in Auto Body Manufacture*, 16 J. L. & ECON. 239 (1973).

8 I knew of McGee's article in part because it was based upon the rate/volume effect that Armen Alchian first had identified. *See* Armen Alchian, *Reliability of Progress Curves in Airframe Production*, 31 ECONOMETRICA 679 (1963). McGee's view of the source of the superiority of General Motors was itself beginning to lose its relevance as he wrote. By 1983, when the Commission investigated the General Motors-Toyota joint venture, it had become clear that it was the Japanese, not Detroit, who were the masters at manufacturing automobiles. The success of Japanese producers in pioneering low

A second real world phenomenon that undermined the rationale for the auto industry inquiry was that the FTC proponents of restructuring U.S. producers had dismissed foreign suppliers as likely to have little competitive impact. Having grown up in California, where foreign cars were increasingly prominent, I found this conclusion bizarre. At that time, I owned a Toyota Corolla; like many other baby boomers, I did not own an American-made car until I discovered the SUV in the 1990s, when I also purchased a Saturn. A third phenomenon that undercut the case for conducting an investigation was uncertainty over gasoline prices. The crude oil price shock of 1973-74 increased gasoline prices dramatically and gave an enormous boost to sales of fuel efficient vehicles. This spurred imports, particularly from Japan.

The auto industry investigation collapsed from its own weight and marketplace realities in May 1981.[9] Although one can take some satisfaction that the investigation ended, it is sobering that the agency did not perceive fundamental flaws in the inquiry when it began in 1976. Contemporary economic learning raised grave doubts about the simple market concentration doctrine. The rapidly changing market environment also should have induced caution in dismissing the entry and expansion by foreign suppliers.

This cautionary tale from my youth contains important lessons. Both economic theory and industry circumstances are ever changing. The prevailing consensus must be tested in the face of new theory and evidence. Good antitrust policymaking requires getting the model right and doing the hard slogging necessary to apply economics to what I have called the "stubborn" facts.[10]

---

cost production methods is documented in JAMES P. WOMACK ET AL., THE MACHINE THAT CHANGED THE WORLD: THE STORY OF LEAN PRODUCTION (1990).

9 *See* A.O. Sulzberger, Jr., *F.T.C. Ends Car Maker Inquiries*, N.Y. TIMES, May 14, 1981, at D1 (reporting Commission decision to end investigation; quoting Richard Rosen, an FTC staff attorney with the Bureau of Competition, as stating: "We have a hard time finding any monopoly profits being earned. Monopoly losses maybe, but not profits.").

10 Timothy J. Muris, *Antitrust Enforcement at the Federal Trade Commission: In a Word—Continuity*, Address before the American Bar Association Antitrust Section Annual Meeting, Chicago, Illinois (Aug. 7, 2001), *available at* http://www.ftc.gov/speeches/muris/murisaba.htm. John Adams discussed the importance of facts in a manner directly relevant to us today. In 1770, Adams defended the British officer and soldiers accused of murder in the Boston Massacre. He said: "Facts are stubborn things; and whatever may be our wishes, our inclinations, or the dictates of our passions, they cannot alter the state of facts and evidence." DANIEL B. BAKER, POLITICAL QUOTATIONS: A COLLECTION OF NOTABLE SAYINGS ON POLITICS FROM ANTIQUITY THROUGH 1989 52 (1990).

B. *Administrability: Distilling Economic Concepts into Workable Rules and Analytical Techniques*

Our most influential law and economics scholars have realized a fundamental principle concerning the link between economic analysis and competition policy. The insights of economics have their greatest impact on antitrust law and policy when they are embodied in *workable* rules and analytical techniques for evaluating business conduct. For economics to have a more important role in antitrust, economists need to pay much more attention to how the sausage is made rather than to the theory of the perfect sausage.

The importance of administrability is evident in those who have played a central role in shaping antitrust doctrine and policy in my professional lifetime. Many of the strongest contributions have come from scholars who realized the importance of translating economic concepts into practical rules and analytical techniques that courts and enforcement agencies could apply successfully. In this cohort I include my UCLA mentors[11] and such figures as Phillip Areeda, William Baxter, Betty Bock, Robert Bork, Frank Easterbrook, Ernest Gellhorn, Richard Posner, and Donald Turner.[12]

The evolution of the U.S. merger guidelines provides an example. Donald Turner's 1968 Guidelines[13] took a formative first step toward rationalizing merger policy that faced a danger of becoming completely detached from any sound conception of economics. Though modest in retrospect, Turner's self-limiting guidelines were revolutionary when adopted, in part because they refused to push enforcement policy to the court-established limits.[14] The most significant breakthrough came in 1982, when

---

11 Armen Alchian was unequaled in teaching economics to lawyers. He often presented economics Socratically—a technique familiar to lawyers. For years Armen was one of the most popular instructors in Henry Manne's programs for teaching economics to lawyers. In short courses, he taught literally hundreds of federal judges and law professors.

12 *See* William E. Kovacic & Carl Shapiro, *Antitrust Policy: A Century of Economic and Legal Thinking*, J. ECON. PERSPECTIVES, Jan. 1, 2000 at 43, 53 (describing role in modern era of economically astute attorneys and legally sophisticated economists who have taken economic concepts and "translated them into operational principles that judges readily could apply"); *see also* Andrew I. Gavil, Sylvania *and the Process of Change in the Supreme Court*, 17 ANTITRUST 8, 11-12 (2002) (based on examination of papers of Supreme Court Justice Lewis Powell, discussing prominent role played by views of William Baxter, Robert Bork, Richard Posner, and Donald Turner in shaping Justice Powell's analysis in *Continental T.V., Inc. v. GTE Sylvania Inc.,* 433 U.S. 36 (1977)).

13 U.S. Department of Justice, Merger Guidelines (1968), *reprinted in* 4 Trade Reg. Rep. (CCH) ¶ 13,101.

14 *See* Oliver E. Williamson, *The Merger Guidelines of the U.S. Department of Justice—In Perspective* (June 4, 2002) (paper prepared for the 20th Anniversary of the Department of Justice Merger Guidelines) (analyzing the 1968 Justice Department Merger Guidelines as precursor to the 1982 Merger Guidelines and underscoring Donald Turner's role in formulating the 1968 Merger Guidelines), *avail-*

Bill Baxter issued new DOJ Merger Guidelines.[15] Baxter's guidelines presented an economically sound and administrable approach to market definition and competitive effects analysis for merger control.[16] Although the hypothetical monopolist had antecedents in economics,[17] it took a lawyer like Bill Baxter, with a solid understanding of economics, to craft a sound and administrable approach to market definition.

Since the 1970s and the structure-conduct-performance ("SCP") debate, the number of industrial organization ("IO") economists and their research have soared. During the 1980s and into the 1990s, industrial organization attracted many of the best young economists. Although IO was once a largely empirical discipline, in recent decades empirical research has lost much of its market share. The lure of IO for most young economists was to apply modern mathematical economics to the relatively undeveloped turf of industrial organization.

There have been important advances in this mathematical literature that have been distilled into useful operational principles. For example, modern oligopoly theory built on the work of George Stigler[18] to provide a more rigorous approach to the analysis of tacit coordination that provides useful guidance for policy and the law.[19] The enhancement by DOJ and

---

able at http://www.usdoj.gov/atr/hmerger/11257.pdf.

15 U.S. Department of Justice, Merger Guidelines (1982), *reprinted in* 4 Trade Reg. Rep. (CCH) ¶ 13,102.

16 *See* William Blumenthal, *Clear Agency Guidelines: Lessons from 1982*, 68 ANTITRUST L.J. 5, 17-20 (2000) (emphasizing how 1982 Guidelines made skillful tradeoffs between simplicity and flexibility and presented useful operational framework); Thomas B. Leary, *The Essential Stability of Merger Policy in the United States*, 70 ANTITRUST L.J. 105, 114-21 (2002) (discussing contributions of 1982 Guidelines and subsequent refinements).

17 *See* Gregory J. Werden, *The 1982 Merger Guidelines and the Ascent of the Hypothetical Monopolist Paradigm* (June 4, 2002) (paper prepared for the 20th Anniversary of the 1982 Department of Justice Horizontal Merger Guidelines), *available at* http://www.usdoj.gov/atr/hmerger/11256.pdf.

18 *See* George J. Stigler, *A Theory of Oligopoly*, 72 J. POL. ECON. (1964).

19 *See* Carl Shapiro, *Theories of Oligopoly Behavior*, *in* 1 HANDBOOK OF INDUSTRIAL ORGANIZATION 329, 356-57 (Richard Schmalensee & Robert D. Willig eds., 1989) (calling Stigler's *Theory of Oligopoly* a "classic paper" and observing: "Stigler's view of oligopoly as a problem of policing a tacitly collusive industry configuration is now the norm."); Jonathan B. Baker, *Two Sherman Act Section 1 Dilemmas: Parallel Pricing, the Oligopoly Problem, and Contemporary Economic Theory*, 38 ANTITRUST BULL. 143, 150 (1993) ("Stigler profoundly changed the way economists understand coordination among oligopolists; and his analysis has also influenced antitrust law."). Stigler's insights about the conditions for effective coordination resonate in modern cases; *see, e.g.,* Airtours v. Commission, Case T-342/99, Judgment of the Court of First Instance (Fifth Chamber, Extended Composition) of 6 June 2002, *available at* http://europa.eu.int/smartapi/cgi/sga_doc?smartapi!celexplus!prod!CELEXnumdoc&lg=en&numdoc=61999A0342 (discussing requirements that must be satisfied to prove collective dominance under the European Union's merger regulation).

other competition authorities of leniency programs employs the basic intuition of the prisoner's dilemma to induce individual cartel participants to reveal their unlawful collaboration.[20] Despite these accomplishments, there have been relatively few successful efforts to translate the mathematically elaborate, game theoretic models into administrable antitrust rules or analytical techniques to support enforcement.[21]

C. *Centrality of Empirical Research*

Economics tells us that monopoly can be "bad," but that is the "easy" part. How do we know when we have a monopoly? How do we know which conduct by a monopolist is "bad"? Even when we know it is "bad," what can we do about it? The efficient administration of statutes against monopolies, or trusts, requires presumptions, preferably ones with sound empirical support. The contribution of economics in this regard is improving. Especially over the past few decades, economics had a critical role in correctly characterizing the state of competition in the U.S. economy and, therefore, in guiding the presumptions used in antitrust policy and litigation.

1. Statistical Analysis

During the first few decades after World War II, economists spent considerable effort seeking to determine whether the U.S. economy was rife with market power that could be cured through antitrust enforcement or whether remediable market power problems were relatively rare. The fundamental presumptions about the competitive health of the U.S. economy were resolved in one of the finest hours of industrial organization economics—the debate between Joe Bain and his "disciples" and those who came to be called "Chicago school" economists and lawyers.[22]

---

20 The evolution of the Justice Department's leniency program is traced in Donald C. Klawiter, *Corporate Leniency in the Age of International Cartels: The American Experience*, 14 ANTITRUST 13 (2000). The incentive effects of leniency programs are analyzed in Bruce H. Kobayashi, *Antitrust, Agency, and Amnesty: An Economic Analysis of the Criminal Enforcement of the Antitrust Laws Against Corporations*, 69 GEO. WASH. L. REV. 715 (2001).

21 This was a major theme expressed by the participants in the FTC's Empirical Industrial Organization Roundtable in September 2001. *See, e.g.*, Federal Trade Commission, *Empirical Industrial Organization Roundtable* 93-100 (Sept. 11, 2001) (comments of Michael Whinston), *available at* http://www.ftc.gov/be/empiricalioroundtabletranscript.pdf.

22 The large body of statistical work performed by industrial organization economists associated with the University of Chicago was especially important. *See, e.g.,* YALE BROZEN, CONCENTRATION, MERGERS, AND PUBLIC POLICY (1982) (questioning "link" between concentration and prices and pro-

The cutting edge of this debate was empirical—theory alone could not resolve the issues. Many today probably do not know that George Stigler called for economy-wide industrial deconcentration in the early 1950s.[23] Stigler's recommendation was based on existing empirical economic research on economies of scale at the plant level in manufacturing that appeared to indicate that American industry was concentrated far beyond "efficiency requirements." Stigler changed his position when he learned of various analytical flaws in the research and of empirical work inconsistent with deconcentration.[24]

Major support for deconcentration also came from statistical analyses of the relationship between market structure and measures of "performance." These studies produced the SCP approach to industrial organization economics popularized in Mike Scherer's text, which first appeared in 1970.[25] This was the first major debate in IO economics in which statistical analysis was central. For many antitrust lawyers and industrial organization economists, the debate turned at the 1973 Airlie House conference memorialized in *Industrial Concentration: The New Learning*,[26] among the most influential volumes ever written for antitrust policy. That book showed that the SCP paradigm had theoretical flaws and lacked empirical support. This new learning fundamentally changed the antitrust community's view about the American economy's competitiveness.

---

viding evidence of alternative, efficiency-related reasons for observed prices); *see also* Harold Demsetz, *Two Systems of Belief about Monopoly*, *in* INDUSTRIAL CONCENTRATION: THE NEW LEARNING 164 (Harvey J. Goldschmid et al. eds., 1974) (criticizing preoccupation of antitrust policy with market concentration and private exclusionary conduct).

23 In a much-read magazine article in 1952, Stigler said "[w]hen a small number of firms control most or all of the output of an industry, they can individually and collectively profit more by cooperation than by competition . . . . These few companies, therefore, will usually cooperate." George J. Stigler, *The Case Against Big Business*, FORTUNE, May 1952, at 123. Decades later, Stigler said in his memoirs that "[u]ntil the 1950s I accepted the prevailing view of my profession that monopoly was widespread . . . . I was an aggressive critic of big business." GEORGE J. STIGLER, MEMOIRS OF AN UNREGULATED ECONOMIST 97 (1988). Stigler noted that in 1950 he "believed monopoly posed a major problem in public policy . . . and that it should be dealt with boldly by breaking up dominant firms and severely punishing businesses that engaged in collusion." *Id.* at 99. In the early 1950s Stigler advocated breaking U.S. Steel, which then accounted for 30% of steel production, into several smaller firms. Explaining this position, Stigler said: "Economists (including me) generally believed that this level of industry concentration [a four-firm steel industry concentration ratio of 60%] allowed a substantial amount of noncompetitive behavior, but the belief rested more upon consensus than upon evidence." *Id.* at 99-100.

24 Stigler himself created the "survivorship" analysis for identifying efficient firm size. *See* George J. Stigler, *The Economies of Scale*, 1 J. L. & ECON. 54 (1958) (introducing "survivor principle").

25 F.M. SCHERER, INDUSTRIAL MARKET STRUCTURE AND ECONOMIC PERFORMANCE (1970).

26 INDUSTRIAL CONCENTRATION: THE NEW LEARNING (Harvey J. Goldschmid et al. eds., 1974).

The SCP paradigm was overturned because its empirical support evaporated. Re-estimation of structure-performance equations accounting for efficiency explanations and data problems made the results "go away."[27] Further, various case studies, particularly involving antitrust cases and investigations, indicated that although some industries appeared to have market structures favorable for the existence and exercise of substantial market power, the industries were, nonetheless, quite competitive. This research made clear that sound theory *plus* the details of markets and institutional factors are necessary to understand competition.[28]

2. Case Studies

As suggested above, one foundation for competition policy has been statistical analyses across industries that reduce market and institutional factors into a relative few variables. Though useful for some purposes, broad statistical studies may provide only limited help in understanding competition in a *specific* industry. Broad statistical analysis typically cannot provide the perspective a detailed examination of important institutional factors offers.

What the FTC routinely does in antitrust enforcement and litigation is to analyze specific industry details and institutional arrangements. The agency's methodology is analogous to case studies and, in its finest form, pays proper attention to institutions that influence competition. Beyond the context of individual enforcement matters, careful case studies have enriched our understanding of such issues as market power and efficiencies, contributing to improvements in antitrust policy.[29]

---

27 In this regard, perhaps the most significant contribution was that of Harold Demsetz. *See* Harold Demsetz, *Two Systems of Belief About Monopoly*, *supra* note 22 at 164-84; *see also* Timothy J. Muris, *Economics and Antitrust*, 5 GEO. MASON L. REV. 303, 303-06 (1997) (describing role of Airlie House Conference in focusing attention on Demsetz and other researchers whose empirical work undermined assumptions of SCP model).

28 Hundreds of Hart-Scott-Rodino Act pre-merger investigations confirm this observation, explaining why structure remains, at best, a crude screen, not a dispositive tool.

29 Noteworthy examples of this type work can be found in FEDERAL TRADE COMMISSION, IMPACT EVALUATIONS OF FEDERAL TRADE COMMISSION VERTICAL RESTRAINTS CASES (Ronald N. Lafferty et al. eds., 1984).

II. EXPLORING THE BLACK BOX: NEW INSTITUTIONAL ECONOMICS AND TRANSACTION COSTS

Economics is neither monolithic nor static. In its modern manifestation, researchers have devised theories to condemn or praise virtually any business practice. The challenge for courts and enforcement agencies is to identify methodologies for the most accurate diagnosis of the competitive consequences of business behavior.

One of the most promising developments for antitrust in modern economic analysis is the New Institutional Economics ("NIE").[30] In general terms, this body of work seeks to extend and enrich understanding of the microanalytic details of business behavior and the industry settings that shape firm conduct.[31] The most impressive recent competition policy work I have seen reflects the NIE's teachings about the appropriate approach to antitrust analysis. Much of the FTC's best work follows the tenets of the NIE and reflects careful, fact-based analyses that properly account for institutions and *all* the relevant theories, not just market structure and market power theories.

In horizontal cases, the largest element of the FTC's enforcement program, the important institutions usually are not the government or property rights regimes.[32] Rather, the key institutions are the determinants of the specifics of competition in each industry. For example, how do transactions

---

30 Douglass North defines "institutions" as "the rules of the game in a society or . . . the humanly devised constraints that shape human interaction." He adds that "[i]n the jargon of the economist, institutions define and limit the set of choices of individuals." DOUGLASS C. NORTH, INSTITUTIONS, INSTITUTIONAL CHANGE, AND ECONOMIC PERFORMANCE 3-4 (1990). Modern economics uses the concept of institutions to examine a wide range of phenomena that shape the behavior of individuals and organizations. *See* MASAHIKO AOKI, TOWARD A COMPARATIVE INSTITUTIONAL ANALYSIS 1-29 (2001) (surveying literature on definition of "institutions"); Christopher Clague, *The New Institutional Economics and Economic Development*, *in* INSTITUTIONS AND ECONOMIC DEVELOPMENT 13, 18 (Christopher Clague ed., 1997) (Institutions "can be organizations or sets of rules within organizations. They can be markets or particular rules about the way a market operates. They can refer to the set of property rights and rules governing exchanges in a society. They may include cultural norms of behavior. The rules can be either formally written down and enforced by government officials or unwritten and informally sanctioned.").

31 *See* Oliver E. Williamson, *The New Institutional Economics: Taking Stock, Looking Ahead*, 38 J. ECON. LIT. 595 (2000) (surveying the modern NIE literature). The predecessor to the "New" Institutional Economics was the institutionalist school of the early 20th century. The literature of the original institutionalist school often included fact-intensive, historical accounts of specific industries, but without applying the microanalytic tools and theory employed by NIE scholars. *See* Kenneth J. Arrow, *Reflections on the Essays*, *in* ARROW AND THE FOUNDATIONS OF THE THEORY OF ECONOMIC POLICY 727, 734 (George Feiwel ed., 1987).

32 In some of the FTC's non-merger matters, other institutions (including patent law and the operation of the Hatch-Waxman statute) that are external to the specific market have been important.

occur, and what are their determinants? Can the process be properly approximated as an auction, and, if so, what kind? Are transactions negotiated? Are suppliers "qualified," and what does this qualification involve? What is the nature of supplier/buyer relationships? How important are long-term relationships? What information do the transacting parties possess that is relevant to the transaction's outcome?

A. *Coase and the Black Box*

To understand the NIE's contributions, it is useful to consider some of the weaknesses in economics this body of work has addressed. In his 1991 Nobel Prize Lecture, Ronald Coase discussed the explanatory power of modern neoclassical economic theory.[33] Coase described this theory as "a state of the world that lives in the minds of economists, but not on earth."[34] Coase went on to consider how traditional price theory fails to account for "non-market" parameters that significantly influence market outcomes and specific exchange relationships. Describing the mainstream theory as "blackboard economics," Coase explained:

> The firm and the market appear by name but they lack any substance. The firm in mainstream economic theory has often been described as a 'black box,' [a]nd so it is. This is very extraordinary given that most resources in a modern economic system are employed within firms, with how these resources are used dependent on administrative decisions and not directly on the operation of a market. Consequently the efficiency of the economic system depends to a very considerable extent on how these organizations conduct their affairs, particularly, of course, the modern corporation. Even more surprising, given economists' interest in the pricing system, is the neglect of the market or more specifically the institutional arrangements which govern the process of exchange. As these institutional arrangements determine to a large extent what is produced, what we have is a very incomplete theory.[35]

This is sharp criticism from one of the giants of 20th century economics. What we learned from the SCP debate and, of course, from the work of Coase himself and the research that he inspired in others is that institutions and facts matter. Unfortunately, the typical IO theory article contains little description or analysis of institutions and factual details. Of course, I do not claim that theoretical, mathematically-oriented economics lacks any value. Improving theory strengthens any discipline, and important theoretical de-

---

33 By "neoclassical" economic theory, I mean theory that is based primarily on downward sloping demand curves and highly stylized models of competition not firmly grounded in any specific institutional setting.

34 Ronald H. Coase, *The Institutional Structure of Production* (Alfred Nobel Memorial Prize Lecture in Economic Sciences, Dec. 9, 1991), *reprinted in* Ronald H. Coase, ESSAYS ON ECONOMICS AND ECONOMISTS 3, 5 (1994) [hereinafter *Nobel Lecture*].

35 *Id.* at 5-6.

velopments in economics often come through mathematical modeling. Nonetheless, I am struck that, despite the enormous recent volume of theoretical, highly mathematical IO literature, its effect on antitrust policy and law has been quite small.

The fundamental reason for this modest influence is that, although empirical work overturned the SCP paradigm, too much of modern IO theory adopts the SCP approach by making market structure the only important market feature in the model. Put differently, although the SCP debate reveals that there is no systematic relationship between market structure and the competitiveness of the market, much of modern IO theory allows *only* market structure and *assumed* market power to be important determinants of the competition.

The mathematical IO literature illuminates how substantial market power might be exercised, *assuming* it exists. Undoubtedly, this question is important. But it identifies and considers few bases for business decision-making other than market power, thereby greatly overemphasizing the importance of such power. As Coase said in 1972, "One important result of this preoccupation with the monopoly problem is that if an economist finds something—a business practice of one sort or another—that he does not understand, he looks for a monopoly explanation. And as in this field we are very ignorant, the number of understandable practices tends to be rather large, and the reliance on a monopoly explanation, frequent."[36]

Thus, literature of antitrust economics abounds with theoretical models that go far beyond horizontal mergers and cartelizing conduct and "simple" theories of vertical foreclosure. Because they start with the assumption of market power and then ask how that market power might be exercised, you can find theoretical support for, among other things, predatory pricing at prices above costs, tying as a monopolizing device, and *even* the Robinson Patman Act.[37]

A visitor from Mars reading this literature would infer that the U.S. economy is rife with monopoly power. However, unlike 1972, there is con-

---

36 Ronald H. Coase, *Industrial Organization: A Proposal for Research*, *in* POLICY ISSUES AND RESEARCH OPPORTUNITIES IN INDUSTRIAL ORGANIZATION 59, 67 (Victor R. Fuchs ed., 1972).

37 *See, e.g.*, Aaron S. Edlin, *Stopping Above-Cost Predatory Pricing*, 111 YALE L.J. 941 (2002) (discussing how a monopolist with a cost advantage over its potential rivals might deter entry despite its high pre-entry price); Dennis W. Carlton & Michael Waldman, *The Strategic Use of Tying to Preserve and Create Market Power in Evolving Industries*, 33 RAND J. ECON. 194 (2002) (discussing how firms can use the tying of complementary products to create or protect monopoly power); Michael L. Katz, *The Welfare Effects of Third-Degree Price Discrimination in Intermediate Goods Markets*, 77 AM. ECON. REV. 154 (1987) (finding possible benefits of forbidding third-degree price discrimination when bargaining power of chain stores comes from their ability to credibly threaten to integrate backward into the supply of intermediate goods); Michael D. Whinston, *Tying, Foreclosure, and Exclusion*, 80 AM. ECON. REV. 837 (1990) (discussing possible exclusionary effects of certain tying arrangements).

sensus today—especially among empiricists—that significant market power "problems" are special cases, not the norm. Compared to 1972, major contributions to empirical research have improved our understanding of competition. Antitrust litigation (for example, the monopolization cases of the 1970s) and the wealth of data collected from hundreds of Hart-Scott-Rodino pre-merger filings and investigations created or inspired much of the relevant empirical work.

Of course, market structure and market power are important. The federal merger guidelines and the FTC's enforcement decisions use these variables and sometimes accord them great weight. Nonetheless, the trend of analysis for the past 20 years under the federal merger guidelines and in other areas of antitrust has been to use market structure and market power as two elements of a broader, fact-based analysis of potential competitive effects. My point is that having a theoretical paradigm that largely can explain business conduct *only* by market structure and assumed market power is flawed. Fortunately, there are important alternative theories and analyses that are richer in their examination of business conduct and its effects.

B. *An Example Outside the Black Box: Auctions*

Auction theory and empirical research based on it probably have made the greatest contribution to merger enforcement. The work on auctions provides explicit empirical analyses relevant to assessing a merger's potential effects *if* an auction structure adequately approximates the market setting.[38] The adequacy of the approximation requires close study of the specifics of the nature of competition—that is, proper attention to market institutions. The key to the success of auction theory is that, when appropriate, it actually fits how parties carry out transactions rather than an ad hoc, simplistic model based on market structure and assumed market power. As with the NIE, institutional knowledge is critical.

C. *An Example Inside the Black Box: One-Shot Bertrand/Unilateral Effects Analyses*

The approach to unilateral effects analysis that many economists follow ignores the basic tenets of NIE. Econometric analyses of retail scanner data and highly simple simulation models are used, based on what econo-

---

38 *See* Serdar Dalkir et al., *Mergers in Symmetric and Asymmetric Noncooperative Auction Markets: The Effects on Prices and Efficiency*, 18 INT'L J. INDUS. ORG. 383 (2000); Steven Tschantz et al., *Mergers in Sealed versus Oral Auctions*, 7 INT'L J. ECON. BUS. 201 (2000). For many years the federal antitrust agencies have applied auction theories in antitrust investigations.

mists call a "one-shot Bertrand" model. This approach has at least two major problems that proper attention to NIE would have identified. First, the data are at the *retail* level, while the mergers are between *manufacturers*.[39] The applicability of even properly estimated retail level demand systems to mergers of manufacturers is only now beginning to be addressed, even though the models in question have been used for a decade. Second, the Bertrand model is imposed with virtually no analysis of its actual ability to explain competition in the market. Theory cannot perfectly replicate reality, but applying a highly simplistic theory without any empirical basis that the theory adequately approximates reality is not sound economics.

I am not implying that careful empirical work that involves econometric modeling lacks value in merger investigations—quite the opposite. Certainly, the work of our Bureau of Economics, along with that of outside economists, in connection with last year's Cruise Ship investigation reveals that empirical work can be crucial in Commission decisions. For empirical work and modeling to be useful, however, they must be based firmly in the institutions and the less quantitative evidence about an industry.

D. *Deciphering the Black Box: The Role of NIE and Transaction Costs*

Much of the NIE literature has significant potential to improve antitrust analysis and policy. In particular, one branch of the NIE considers the implications of "transactions costs" on economic behavior. This analysis specifically accounts for the institutional arrangements in which the actors participate.[40] As first developed by Coase, this "transactions cost economics" ("TCE") originally focused on demystifying the "black box" firm and on clarifying important determinants of vertical relationships. TCE builds on the insights of Coase's classic 1937 article on the firm[41] and shifts the analysis toward exchange relationships.[42] Coase taught us that the transac-

---

39 Various data and econometric issues arise in the estimation of demand systems using scanner data. These issues are the subject of a recent Bureau of Economics working paper. *See* Daniel Hosken et al., *Demand System Estimation and its Application to Horizontal Merger Analysis* (April 2002) (FTC Bureau of Economics Working Paper #246), *available at* http://www.ftc.gov/be/workpapers/wp246.pdf.

40 My UCLA instructors contributed extensively to this literature. Armen Alchian was an important developer of the transaction costs approach, as were Harold Demsetz and Ben Klein. Although Demsetz may be best known for his criticism of the market concentration doctrine, he also wrote important articles involving transaction costs analysis. Several of his most influential works, including *Toward a Theory of Property Rights*, AM. ECON. REV., May 1967, at 347 and *When Does the Rule of Liability Matter?*, 1 J. LEGAL STUD. 13 (1972), elaborated on Coase's initial work.

41 Ronald H. Coase, *The Nature of the Firm*, 4 ECONOMICA 386 (1937).

42 Perhaps the seminal example of this effort is OLIVER E. WILLIAMSON, MARKETS AND HIERARCHIES: ANALYSIS AND ANTITRUST IMPLICATIONS (1975).

tions costs arising from problems of contracting and coordination will determine a firm's organization, particularly its degree of vertical integration. The profit-maximizing firm will weigh the costs, risks, and efficiencies associated with performing functions via market-based transactions against the like parameters associated with performing the functions by internal integration, and it will choose the cost-minimizing alternative. Modern TCE builds on this foundation and attempts to explain and better understand the consequences of market-based exchanges by placing them within their specific institutional context. In this sense, modern TCE enhances and complements traditional neoclassical economics.[43]

Two early illustrations of the insights from TCE come to mind. First, antitrust rules based on initial market-power screens typically relied on the textbook definition of market or monopoly power; that is, evidence that prices exceed marginal cost. Such a test is not particularly useful. In the textbook perfect competition model, firms face flat demand curves—they are price-takers. In the real world, when virtually any firm raises its price, it retains some sales. Consider hot dog vendors on the Mall in Washington, D.C. and on street corners in other major cities. Any vendor that raises its price would be unlikely to lose all of its sales. In other words, the demand curve is sloped downward, not flat. Few persons, however, would claim that the hot dog vendor has the degree of market power that courts or enforcement agencies should regard as important for antitrust. The demand curve is sloped because of positive transactions costs. When purchasing a hot dog, most consumers will simply not find it cost-effective to survey available prices first.[44]

Second, scholarly research on the value for advertising in reducing transactions costs has transformed the way antitrust policy makers think about the competitive implications of advertising.[45] George Stigler was an early contributor to this research. In typically blunt fashion, he said advertising is "an immensely powerful instrument for the elimination of igno-

---

43 Early assessments of the empirical support for TCE are favorable. *See* Howard A. Shelanski & Peter G. Klein, *Empirical Research in Transaction Cost Economics: A Review and Assessment*, 11 J.L. ECON. & ORG. 335, 336 (1995) (finding that "a remarkable amount of the empirical work we have examined is consistent with TCE predictions—much more so, perhaps, than is the case with most of industrial organization"); *see also* Bruce Lyons, *Empirical Relevance of Efficient Contract Theory: Inter-Firm Contracts*, 12 OXFORD REV. ECON. POL'Y 27 (1996); Keith Crocker & Scott Masten, *Regulation and Administered Contracts Revisited: Lessons from Transaction Cost Economics for Public Utility Regulation*, 9 J. REG. ECON. 5 (1996).

44 Timothy J. Muris, *The Federal Trade Commission and the Rule of Reason: In Defense of Massachusetts Board*, 66 ANTITRUST L. J. 773, 779-780 (1998). For an excellent discussion of this concept, see Benjamin Klein, *Market Power in Antitrust: Economic Analysis After* Kodak, 3 SUP. CT. ECON. REV. 43, 47-63 (1993).

45 J. HOWARD BEALES & TIMOTHY J. MURIS, STATE AND FEDERAL REGULATION OF NATIONAL ADVERTISING (1993).

rance."[46] In other words, advertising can reduce search costs significantly and thereby foster competition. Ads inform consumers about the availability of new products, new features, or new information about existing products. Such information is vital to competition.[47]

TCE can also shed light on the validity of the anticompetitive stories that market power theories emphasize. TCE enriches existing theory by considering such factors as the efficiencies of contracting given the institutional setting within which trading partners interact and the uncertainties trading partners have about each other. TCE also stresses the importance of an *ex ante* perspective in evaluating the efficiency of conduct and judging whether legal intervention is useful.

Ben Klein's case studies in this area have been especially illuminating. Focusing on transactions costs, the elimination of opportunistic behavior, and the impossibility of drafting fully-contingent legally-enforceable contracts, Klein argues that distributional restraints are best understood not as means to harm consumers[48] or as mechanisms to avoid consumer free riding,[49] but instead to compensate dealers optimally for an increased supply of product promotion services and to prevent price competition that would eliminate the desired marketing plan. In the typical case, manufacturers induce the desired services through sharing of profits, active monitoring, and the threat of termination, which would cause the dealers to lose the compensation stream.[50]

By emphasizing the specific institutional settings in which contracting parties operate, the TCE framework provides an enriched means for antitrust enforcers to evaluate specific behavior. Unless we are evaluating con-

---

46 George J. Stigler, *The Economics of Information*, 69 J. POL. ECON. 213, 220 (1961).

47 A large empirical literature demonstrates the pro-competitive benefits of advertising. I review this literature in Timothy J. Muris, California Dental Association v. FTC: *The Revenge of Footnote 17*, 8 SUP. CT. ECON. REV. 265, 293-304 (2000).

48 The hypothesis that vertical contractual restraints commonly injured consumers motivated federal antitrust enforcement policy in the 1960s and 1970s. *See* Donald F. Turner, *Some Reflections on Antitrust, in* N.Y. STATE BAR ASS'N ANTITRUST LAW SYMPOSIUM 1, 1-2 (1966) (comments by Donald Turner, Assistant Attorney General for Antitrust; remarking that he approached vertical "territorial and customer restrictions not hospitably in the common law tradition, but inhospitably in the tradition of antitrust law"); Wesley J. Liebeler, *Bureau of Competition: Antitrust Enforcement Activities*, *in* THE FEDERAL TRADE COMMISSION SINCE 1970: ECONOMIC REGULATION & BUREAUCRATIC BEHAVIOR 65, 75-84 (Kenneth W. Clarkson & Timothy J. Muris eds., 1981) (discussing FTC vertical restraints enforcement in 1970s).

49 *Compare* Lester G. Telser, *Why Should Manufacturers Want Fair Trade?*, 3 J.L. & ECON. 86 (1960) (arguing that in many instances manufacturers imposed vertical restraints to cure free rider problems).

50 The theory and evidence are summarized in Benjamin Klein & Kevin M. Murphy, *Vertical Restraints as Contract Enforcement Mechanisms*, 31 J.L. & ECON. 265 (1988); Benjamin Klein, *The Economics of Franchise Contracts*, 2 J. CORP. FINANCE 9 (1995).

duct for which adverse competitive effects are almost certain,[51] anticompetitive effects in antitrust cases must be proven with hard facts.[52] The theory supporting enforcement must either be consistent with the hard facts or must explain them away convincingly. Inconvenient facts cannot be ignored; we must deal with them. Traditional market power theories, based solely on generalized, non-case-specific modeling, are rarely helpful. TCE and other applications of the new institutional economics, by contrast, can illuminate the meaning of facts—particularly in the context of complex contractual relations—that otherwise cannot be explained, or worse, are explained incorrectly.

Complex contractual settings are pervasive in our economy.[53] This is particularly apparent to those of us who have taught and written about contract law as well as antitrust. For other academics and myself who cover both fields, the process of reading hundreds of contract cases, old and new, has underscored the value of the NIE and transactions costs perspectives in understanding the interaction of sellers and buyers. The contracts case law reveals the extraordinary degree of experimentation and adaptation by which firms test various approaches to solve traditional and novel problems of cooperation and exchange. One sees the extent that even seemingly "simple" contracts, when examined in detail, frequently provide only a general framework in which the parties develop elaborate, unwritten "relational" understandings that supply the operational rules for executing the commitments described in the formal text.[54] The cases also reveal how various contract provisions can at first glance appear particularly harsh or excessively restrictive but actually constitute legitimate measures to discourage opportunism or other efforts to exploit imperfections in the specification of contract terms or the implementation of legal remedies for reneging.[55]

---

51 Examples include naked price fixing, bid rigging, or customer allocation agreements.

52 *See supra* note 10 and accompanying text.

53 Timothy J. Muris, *The FTC and the Law of Monopolization*, 67 ANTITRUST L.J. 693 (2000) [hereinafter *Law of Monopolization*].

54 "Relational contracts" usually refer to situations in which parties, often in a setting involving long-term interaction, "are incapable of reducing important terms of the arrangement to well-defined obligations." Charles J. Goetz & Robert E. Scott, *Principles of Relational Contracts*, 67 VA. L. REV. 1089, 1091 (1981). The parties rely heavily upon unwritten elaboration and adjustment of their responsibilities due to difficulty with writing contract terms that "identify uncertain future conditions or because of inability to characterize complex adaptations adequately even when the contingencies themselves can be identified in advance." *Id*.

55 Opportunistic behavior refers, in part, to "hold up" problems. On the economic issues and the role of contract law in policing holdups, see Timothy J. Muris, *Opportunistic Behavior and the Law of Contracts*, 65 MINN. L. REV. 521 (1981). For more recent discussions of opportunism in contractual settings with strong relational features, see John P. Esser, *Institutionalizing Industry: The Changing Forms of Contract*, 21 L. & SOC. INQUIRY 593 (1996); Claire Moore Dickerson, *Cycles and Pendulums:*

Many of the most difficult competition issues confronting the FTC and other enforcement bodies involve intricate and fast-changing contractual arrangements. High technology provides a noteworthy example. Bengt Holmstrom and John Roberts documented that complex contractual relationships are common in high tech industries, such as computers and biotechnology, as well as in more traditional ones, such as automobiles and steel.[56] In biotechnology, for example, Holmstrom and Roberts explain that the activities of the different industry members are highly interrelated, with most firms engaged in many partnerships. They note that, in 1996, one firm reported ten marketing partnerships, twenty licensing arrangements, and more than fifteen formal research collaborations.[57] In such a world, change is inevitable. Parties may become disillusioned with each other, and lawsuits may result. Although anticompetitive consequences are possible when relationships change, they are by no means likely, let alone inevitable.

To date, TCE's most important contribution to antitrust is its ability to provide much better explanations for most vertical relationships than market power theories. This contribution has come not so much from theory (although many theoretical papers apply TCE to vertical relationships) as from careful empirical work. Some papers use econometric analysis to estimate the importance of broad categories of transaction costs to the nature of vertical relationships across companies or industries. Following the scholarship of Coase, however, the core of the empirical implementation of transaction cost economics is the intensive case study. For example, Coase's *The Theory of the Firm* was based on in-depth case studies of several companies, comparing and contrasting organizational structure against firm and market parameters.[58] Quantitative data and econometric analysis belong in the case study, but the core of the analysis is identifying the key features of the relevant institutions and competitive processes and the nature and importance of the relevant transaction costs. The analysis then proceeds to analyze how institutions and transaction costs affect relationships and competition.

Good Faith, Norms, and the Commons, 54 WASH. & LEE L. REV. 399 (1997); Benjamin Klein, *Contracts and Incentives: The Role of Contract Terms in Assuring Performance*, *in* CONTRACT ECONOMICS 149-72 (Lars Werin & Hans Wijkander eds., 1992).

56 Bengt Holmstrom & John Roberts, *The Boundaries of the Firm Revisited*, J. ECON. PERSP., Fall 1998, at 73.

57 *Id*. at 85-86.

58 Coase, *Nobel Lecture*, *supra* note 34, at 6-8.

E. *Three NIE Applications*

Let me turn to some specific examples of the application of TCE.

1. *Kodak* and Hold-ups

Applying the insights of TCE, a number of scholars have critically re-examined *Eastman Kodak Co. v. Image Technical Services, Inc.*[59] and similar cases with a view toward a more thorough and rigorous understanding of the underlying conduct. Focusing on transactions costs, protection against opportunistic behavior, and risks embedded in contractual relations, this literature allows much better understanding of the disputes at issue than do court decisions and other analyses of these issues.[60]

The *Kodak* facts are generally well-known. To review briefly, in the early 1980s, independent service organizations ("ISOs") began servicing Kodak's photocopier and micrographics equipment. The ISOs competed with Kodak, often at substantially lower prices. In the mid-1980s, Kodak limited the availability of its replacement parts for Kodak equipment to ISOs, forcing many out of business and prompting the lawsuit.

There is potential for harm to the purchasers of high-volume equipment in this setting. Such purchasers generally make product and relationship-specific investments, including training employees on use of the equipment. Once these investments are made, it is costly to switch brands. It may also be, as the Court claimed, that the equipment's value decreases rapidly in the second-hand market. These low salvage values and high product-specific investments imply that purchasers are "locked in" after their initial equipment purchase, perhaps allowing the seller to take advantage of them by increasing the price it charges for service above the level buyers anticipated when they purchased the original equipment.

Whether Kodak could have engaged in such a hold-up depends upon several facts. Most importantly, for buyers to be harmed the behavior must have been unanticipated.[61] Moreover, buyers can attempt to negotiate spe-

---

59 504 U.S. 451 (1992).

60 The hold-up issues in Kodak are analyzed in Benjamin Klein, *Market Power*, *supra* note 44; Benjamin Klein, *Market Power in Franchise Cases in the Wake of* Kodak*: Applying Post-Contract Hold-Up Analysis to Vertical Relationships*, 67 ANTITRUST L. J. 283 (1999); Carl Shapiro, *Aftermarkets and Consumer Welfare: Making Sense of* Kodak, 63 ANTITRUST L. J. 483 (1995); Muris, *Law of Monopolization*, *supra* note 53, at 704-07.

61 Moreover, buyers could not be harmed unless they made specific investments allowing the seller to engage in a "hold-up." Further, protection against opportunism, either through contract law or some other mechanism, must have been inadequate. Finally, it is important to note that the hold-up problem is distinct from pre-contractual monopoly.

cific contract terms to protect against opportunistic behavior, rely on contract law's prohibition against such behavior, or take other steps to protect themselves. In *Kodak*, for example, the buyer could have purchased at a price far enough below what otherwise would be the market value to reflect potential switching costs.[62]

Because buyers usually can protect themselves, we cannot assume that a change in practice, such as occurred in *Kodak*, is unfair, let alone anticompetitive. Although the buyers are locked in, the change may have been anticipated. If so, how can it be said to be unfair? Of course, an unanticipated hold-up may still have occurred in *Kodak*; we do not know. Ben Klein argues that the arrangement Kodak adopted by tying the sale of some of its equipment to the sale of replacement parts and services was a device for price discrimination.[63] Regardless of the purpose, analyzing hold-up problems under TCE analysis reveals the nature of the problem better than other forms of analysis.[64]

2. Carbonated Soft Drinks

Another example of the use of TCE is the study of the carbonated soft drink industry I did with David Scheffman and Pablo Spiller,[65] which was conducted to understand the nature of vertical relationships in that industry and the reasons why those relationships were changing. This research was

---

62 Price adjustments may not always perfectly deter opportunism. Consider the problem of employees working at less than full capacity, often called "shirking." Even if an employer could hire more employees at lower wages to solve the problem of employee shirking, a greater quantity of lower-quality labor at a low price may not perfectly substitute for a smaller quantity of more expensive, higher-quality labor. In the *Kodak* example, the lower equipment price would, as Klein notes, distort the relative prices of equipment and aftermarket services, leading customers inefficiently to economize on service. *See* Klein, *Market Power, supra* note 44, at 51. Additional contract terms to avoid the potential hold-up include a long-term service agreement or a "most favored purchase" clause on equipment sales that would prevent discriminatory pricing against old purchasers. Finally, a common way to avoid a hold-up problem is to contract with parties who possess sufficiently strong reputations for fair dealing that they have more to lose than gain by a hold-up policy. *See id.* at 50, 51; Muris, *Opportunistic Behavior*, *supra* note 55, at 527. Reliance on reputation may make the arrangement look "one-sided," misleading observers to conclude that the contracts are "unfair."

63 *See* Klein, *Market Power*, *supra* note 44, at 62.

64 *See also* Muris, *Law of Monopolization*, *supra* note 53, at 707 (observing that an analysis focused on hold-ups and opportunism "does not mean that *Kodak* was decided incorrectly. Indeed, it supports the *Kodak* majority's view that the defendants were wrong in asserting that no potential problems existed. Whether Kodak could defend itself by claiming that its practices could not have been anticompetitive was not before the Court, nor did Kodak argue that, although hold-ups were possible, it did not engage in one.").

65 TIMOTHY J. MURIS ET AL., STRATEGY, STRUCTURE, AND ANTITRUST IN THE CARBONATED SOFT-DRINK INDUSTRY (1993).

stimulated by PepsiCo's and Coca-Cola's acquisitions of what had been for decades independent bottlers. Our work, both theoretical and empirical, demonstrated that changes in market conditions made vertical integration superior to independent bottlers to reduce the transaction costs of making, distributing, and selling the product. We also explained that these transactions costs rose dramatically in recent years. In the early years of the industry, independent bottling was a sensible response to the difficulty of managing hundreds of manufacturing and distribution operations for what was then a simple product with a relatively simple marketing strategy.

3. Antitrust Rules and Remedies

TCE also has valuable insights for framing antitrust rules and remedies. Antitrust policy normally seeks to deter misconduct rather than subject businesses to regulatory controls. Antitrust rules thus embody both potential benefits (the promotion of competition) and potential costs (the reduction of competition and innovation, and the augmentation of private sector and judicial costs). Errors in designing and applying enforcement rules—especially errors that chill pro-competitive conduct—can have severe adverse consequences. Thus, antitrust rules must account for these benefits and costs.

Given the transactions costs, it is clear that antitrust rules are not well-suited to "fix" all market imperfections. Rules that are "flexible" enough to accommodate every theory of "market power" place inordinate burdens on the courts to sort through huge amounts of economic facts, comprehend complex and often quite abstruse theory, and ultimately distinguish lawful competition from unlawful acts. The risk of error is substantial.

By contrast, relatively clear and simple antitrust rules may allow some amount of anticompetitive conduct to escape prosecution. The efficient rule will minimize the sum of failures to capture anticompetitive conduct and interventions that challenge or chill pro-competitive conduct, along with the costs of understanding and following the rules and of litigation. Paul Joskow puts it well when he says "the test of a good legal rule is not primarily whether it leads to the correct decision in a particular case, but rather whether it does a good job deterring anticompetitive behavior throughout the economy given all of the relevant costs, benefits, and uncertainties associated with diagnosis and remedies."[66] Thus, optimal rules will create appropriate incentives for efficient behavior *ex ante*.

---

66 Paul L. Joskow, *Transaction Cost Economics, Antitrust Rules, and Remedies*, 18 J. LAW, ECON. & ORG. 95, 99-100 (2002).

TCE may also yield important insights in fashioning antitrust remedies. Joskow argues, for example, that antitrust enforcers should consider divestiture remedies with caution, and only after careful analysis reflecting transaction cost issues. It is obvious that for merger divestitures to restore competition to pre-merger levels, the divested assets must remain viable not only during the dispute, but also after the divestiture takes place. The problem is how to assess the long term viability of assets to-be-divested *ex ante*. Joskow argues that the organizational design and governance arrangements for the assets require careful attention. He cites an FTC study of thirty-five divestiture orders, taking place from 1990-1994, that concluded that twenty-five percent of the divestitures failed to create a sustained and viable competitive force within the target market.[67]

Joskow finds this result unsurprising. TCE, he suggests, tells us:

> Firms subject to 'voluntary' divestitures to mitigate market power should be expected to behave strategically; ongoing businesses that have been divested are likely to fare better post-divestiture than are assets that require the creation of a complete new business organization to be used effectively; buyers negotiating divestiture agreements in which they depend on the seller and have not protected themselves against ex post holdups are likely to face the consequences of these holdups; contractual arrangements for input supplies between competing firms can soften competition between them; it's not the size of the acquirer but its ability to utilize the assets effectively that matters.[68]

Consequently, enforcement agencies must become more sophisticated in their analysis of alternative remedies. That is exactly what we are trying to do at the FTC.[69]

---

67 STAFF OF THE BUREAU OF COMPETITION, FEDERAL TRADE COMMISSION, A STUDY OF THE COMMISSION'S DIVESTITURE PROCESS (1999), *available at* http://www.ftc.gov/os/1999/9908/divestiture.pdf.

68 Joskow, *supra* note 66, at 114.

69 *See* Federal Trade Commission, *Frequently Asked Questions About Merger Control Consent Order Provisions, at* http://www.ftc.gov/bc/mergerfaq.htm; *see also* Robert Pitofsky, Chairman, Federal Trade Commission, *The Nature and Limits of Restructuring in Merger Review*, Prepared Remarks Before the Law Seminars International Cutting Edge Antitrust Conference, New York, New York (Feb. 17, 2000) (discussing importance of improving Commission's approach to devising remedies), *available at* http://www.ftc.gov/speeches/pitofsky/restruct.htm.

III. A RESEARCH AGENDA TO IMPROVE ANTITRUST'S ECONOMIC FOUNDATIONS

A. *The FTC's Research Agenda*

As Chairman, I have emphasized the need for the antitrust agencies to devote adequate resources to competition policy "research and development."[70] Over the past nineteen months, the FTC has embarked on several initiatives to improve and enhance its use of empirical economic analysis. The starting point for these initiatives was an Empirical Industrial Organization Roundtable, one of the first major events I convened on my return to the FTC. This event, organized by the Bureau of Economics ("BE") and Dennis Carlton, brought together leading industrial organization scholars to discuss a wide array of topics that they felt needed more attention from researchers. Among the ideas that received the greatest support was more research evaluating merger outcomes—both in terms of competitive effects and efficiencies.[71] We are actively engaged in a number of these research projects.

Our most significant effort is research on consummated hospital mergers.[72] We have formed a Merger Litigation Task Force, whose mission includes investigating hospital mergers recently consummated. (Antitrust agencies unsuccessfully challenged some of these mergers.) These studies are using proprietary data on payments to hospitals by third-party payers to assess whether these mergers raised prices, taking into account any impact on quality.

---

70 Timothy J. Muris, Chairman, Federal Trade Commission, *Looking Forward: The Federal Trade Commission and the Future Development of U.S. Competition Policy*, Prepared Remarks before the Milton Handler Annual Antitrust Review, New York, New York (Dec. 10, 2002), *available at* http://www.ftc.gov/speeches/muris/handler.htm.

71 *See* Federal Trade Commission, *Empirical Industrial Organization Roundtable* 115-41 (Sept. 11, 2001), *available at* http://www.ftc.gov/be/empiricalioroundtabletranscript.pdf; *see also* William E. Kovacic, *Evaluating Antitrust Experiments: Using Ex Post Assessments of Government Enforcement Decisions to Inform Competition Policy*, 9 GEO. MASON L. REV. 843, 855-56 (2001) (collecting authorities who have recommended that government antitrust agencies devote more resources to analyzing effects of completed cases). BE staff has long been involved in this type of research; *see, e.g.*, LAURENCE SCHUMANN ET AL., CASE STUDIES OF THE PRICE EFFECTS OF HORIZONTAL MERGERS (Bureau of Economics Report, April 1992).

72 The FTC's efforts to study the effects of consummated hospital mergers are addressed in Timothy J. Muris, Chairman, Federal Trade Commission, *Everything Old Is New Again: Health Care and Competition in the 2*1st *Century*, Prepared Remarks Before the 7th Annual Conference in Health Care Forum, Chicago, Illinois (Nov. 7, 2002), *available at* http://www.ftc.gov/speeches/murishealthcarespeech0211.pdf.

Other merger retrospectives involve the petroleum industry. A major revision of the 1982 and 1989 FTC staff reports on oil mergers is underway.[73] The agency also is in the early stages of empirical research to assess the effects of various oil mergers of the past decade. On a related topic, the FTC's staff also will report on what we learned in conferences we held in August 2001 and May 2002 on the reasons for the volatility of refined petroleum products.

Another substantial topic of discussion at the Empirical Roundtable was the use of scanner data estimation and simulation models to investigate branded products mergers. Last year, BE economists produced a Working Paper on the use of scanner data estimation in antitrust.[74] Relying on econometric and economic analysis, including analyses from the merging parties, the paper identifies a number of significant issues in the estimation of demand systems and the interpretation of retail demand elasticities using scanner data. Some tentative conclusions were described by an FTC economist and an Antitrust Division economist in a luncheon program hosted by the American Bar Association's Section of Antitrust Law. Finally, an FTC economist is working with an academic economist on a retrospective study of several branded products mergers.

Yet a different BE paper explores the general role of econometrics and empirical analyses in antitrust, and it suggests best practices for developing econometric studies that will be useful for FTC decisions.[75] The paper provides examples of many commonly used analyses in investigations, including the use of scanner data for demand estimation, the use of merger simulation models, and the use of manufacturer level (especially transaction specific) data in consumer and industrial product mergers. In addition, we have published best practices for parties and outside counsel for interacting with the FTC regarding data and empirical analyses.[76]

Another significant research project is to develop an analytical approach and empirical analyses to help assess the impact of mergers on the potential for coordinated interaction. Several types of analyses were used in

---

73 FEDERAL TRADE COMMISSION, MERGERS IN THE PETROLEUM INDUSTRY (Sept. 1982); FEDERAL TRADE COMMISSION, BUREAU OF ECONOMICS, MERGERS IN THE U.S. PETROLEUM INDUSTRY, 1971-1984: AN UPDATED COMPARATIVE ANALYSIS (1989).

74 DANIEL HOSKEN ET AL., DEMAND SYSTEM ESTIMATION AND ITS APPLICATION TO HORIZONTAL MERGER ANALYSIS (FTC Bureau of Economics Working Paper #246, 2002), *available at* http://www.ftc.gov/be/workpapers/wp246.pdf. The final version will be published in an American Bar Assocation Section of Antitrust Law monograph on the use of econometrics in antitrust.

75 David Scheffman & Mary Coleman, *FTC Perspectives on the Use of Econometric Analyses in Antitrust Cases* (2002), *available at* http://www.ftc.gov/be/ftcperspectivesoneconometrics.pdf.

76 Bureau of Economics, Federal Trade Commission, *Best Practices for Data, and Economics and Financial Analyses in Antitrust Investigations* (2002), *available at* http://www.ftc.gov/be/ftcbebp.pdf.

the Commission's investigation of the cruise company mergers.[77] At today's conference, BE Deputy Director Mary Coleman is discussing her research with Bureau Director David Scheffman regarding an analytical approach to coordinated interaction.

We also conducted a two-day Roundtable focusing on various issues about the rate of success of mergers, including key measures for success or failure.[78] This Roundtable blended empirical research on merger outcomes with knowledge from academics, consultants, and business leaders experienced in mergers and acquisitions. The transcript of this Roundtable, soon to be posted on the FTC's website, will be valuable for economists and others interested in studying merger efficiencies.

The FTC has also been extremely active in studying competition in pharmaceutical markets. Prompted by its own experience in antitrust cases and a request from Congress, the FTC began a lengthy industry-wide study of procedures to introduce generic drugs *prior* to the expiration of patents protecting the brand name drug. In July 2002, the Commission released *Generic Drug Entry Prior to Patent Expiration: An FTC Study*.[79] Based on responses from branded and generic pharmaceutical companies, this report clarified several important issues regarding the interplay between patents on pioneer pharmaceuticals and the ability of generic drugs to compete. The report greatly increased knowledge about the extent and potential effects of patent litigation and settlements, and its finding and recommendations were valuable to government decision makers considering modifications to the Hatch-Waxman statute.[80]

---

77 *See* Statement of the Federal Trade Commission Concerning Royal Caribbean Cruises, Ltd./P&O Princess Cruises plc and Carnival Corp./P&O Cruises plc, FTC File No. 021-0041 (Oct. 4, 2002), *available at* http://www.ftc.gov/os/2002/10/cruisestatement.htm; *see also* Joseph J. Simons, Director, Bureau of Competition, Federal Trade Commission, *Merger Enforcement at the FTC*, Prepared Remarks Before the Tenth Annual Golden State Antitrust and Unfair Competition Law Institute, Santa Monica, California (Oct. 24, 2002) (discussing cruise inquiry), *available at* http://www.ftc.gov/speeches/other/021024mergerenforcement.htm; Bureau of Economics, Federal Trade Commission, *Cruise Investigation: Empirical Economic & Financial Analyses* (Nov. 2002), *available at* http://www.ftc.gov/be/hilites/ftcbeababrownbag.pdf.

78 The Roundtable, titled *Understanding Mergers: Strategy and Planning, Implementation and Outcomes*, was held in December, 2002.

79 The study is available at http://www.ftc.gov/os/2002/07/genericdrugstudy.pdf.

80 President Bush cited our study as the basis for the FDA's recent proposal to limit branded drug companies to a single 30-month stay under the Hatch-Waxman Act. *President Takes Action to Lower Prescription Drug Prices by Improving Access to Generic Drugs*, (Oct. 21, 2002), *available at* http://www.whitehouse.gov/news/releases/2002/10/20021021-4.html.

B. *Important Areas for More Empirical Research*

The FTC is making a strong commitment to the fact-intensive research that improves the economic foundations of antitrust enforcement. We find that each new case and research project raises issues that require thoughtful analysis. Looking ahead, several areas are leading candidates for further study by both the FTC and the academy.

1. Research on Competitive Effects

The most important area for further research is empirical study to improve our understanding of the nature of competition and of how and why horizontal mergers affect competition. I discussed earlier the benefits of work on auction mechanisms and the deficiencies in the current state of Bertrand simulation models. The analysis of FTC economists in the Cruise Ships investigation was one useful contribution. The FTC economists highlighted the importance of understanding the relevant institutions and of detailed analysis of actual transactions, changes in capacity, and other facts. Mary Coleman's and David Scheffman's paper at this conference provides a further contribution. Closely related to the development of sound empirical analyses to assess the competitive effects of mergers is the need to understand manufacturer pricing for consumer products. We will be working with outside experts on this issue.

2. Retrospectives on Merger Enforcement Decisions

To assess the efficacy of merger enforcement, we need to analyze the effects of past enforcement actions, including non-enforcement decisions. Specifically, we need to understand the industry and firm specific conditions relevant to the potential for anticompetitive effects. We also need to know much more about the nature and likelihood of significant procompetitive effects of mergers. Understanding the efficiencies that can arise from mergers and how they are achieved would provide us with greater ability to evaluate prospective mergers.

3. High Margin Industries

Mergers in high margin (high fixed costs) industries frequently present challenging issues. High margins can be simplistically equated with substantial existing market power (unilateral or coordinated). Given the nature

of costs in such industries, however, high margins are often necessary for firms to be profitable. Our understanding of the nature of competition in such industries is limited—particularly regarding how mergers might affect the nature of this competition.

4. Price Discrimination

Some lawyers and economists use evidence of "price discrimination" to infer market power and market definition, raising several issues. Most real world markets, even those for relatively "homogenous" products and a market structure inconsistent with significant market power, exhibit significant price variation. These price differences do not prove that the firms have market power. Moreover, price discrimination can be pro-competitive. A significant deficiency of the economics literature is the fragmentary explanation of why significant price variation is common and understanding the implications of this fact.

A related issue occurs when the agency learns of customers concerned about targeted price increases. These concerns are difficult to assess, especially without detailed industry data. Greater focus on techniques to evaluate and analyze transaction data will yield insights into the likelihood of potential anticompetitive pricing. Thus, more research is needed concerning how to identify price discrimination that raises competitive concerns and the role that price discrimination should have in merger analysis. We are beginning studies in this area and encourage others to do so as well.

5. Other Issues

This list is only the beginning of the numerous topics on which antitrust would benefit from a better empirical foundation. For example, the many theories of possible anticompetitive behavior in the theoretical literature—including raising rivals' costs, naked exclusion, and the impact of network effects—all lack significant empirical analysis.[81]

---

[81] To say that the relevant literature is limited is not to say it does not exist. For an interesting example of research that brings an NIE perspective to bear on analyzing an apparent episode of raising rivals' costs, see Elizabeth Granitz & Benjamin Klein, *Raising Rivals' Costs: The Standard Oil Case*, 39 J.L. & ECON. 1 (1996).

CONCLUSION

In passing the federal antitrust laws, Congress adopted an evolutionary scheme in which courts would alter doctrine by "recognizing and adapting to changed circumstances and the lessons of accumulated experience."[82] The rationality of our antitrust system requires continuing efforts to make this process of adaptation well-informed by refinements in economic theory and empirical research. Too often in our antitrust history, sound understanding of business behavior has lagged behind, not accompanied by the formulation and application of legal rules.[83]

As the dynamism and complexity of the economy grow, competition policy institutions face ever greater pressure to improve their ability to analyze business conduct accurately and swiftly. Meeting this challenge requires greater use of analytical approaches that enrich our understanding of the institutions that govern behavior within firms and industries. By combining microeconomic theory with close attention to industry-specific facts, New Institutional Economics provides a valuable framework for achieving the necessary understanding. In the words of one of the Federal Trade Commission's most illustrious economists, our concern must be "with how specific companies and markets operate, with what specific received data on competition mean, and with the economic development of the law."[84]

With its unique combination of economic research and legal policy functions, the FTC occupies a special position to improve antitrust's economic foundations. With its own resources, the agency is expanding efforts to examine the effects of its enforcement decisions and otherwise to strengthen the empirical basis for future policy making. Our own initiatives alone are important but not sufficient. For decades, the Commission has drawn heavily on the contributions of academics and other researchers. We seek to extend this intellectual partnership through greater interaction with scholars to identify priorities for industrial organization research.

I finish by expressing my hope that IO returns to be a much more empirically-oriented discipline, guided of course, by sound theory. We all have much to do to ensure that antitrust repeats the successes and avoids

---

82 State Oil Co. v. Khan, 522 U.S. 3, 20 (1997).

83 *See* Dennis W. Carlton, *A General Analysis of Exclusionary Conduct and Refusal to Deal—Why* Aspen *and* Kodak *are Misguided*, 68 ANTITRUST L.J. 659, 680 (2001) ("[A]s the literature in economics shows, economists often take decades to understand certain business practices.").

84 Betty Bock, *Ethical Considerations for an Antitrust Economist: An Economist's View*, 48 ANTITRUST L.J. 1875 (1979). Betty Bock was a leading figure in the use of economic analysis to inform antitrust law in the post-World War II era. Among other distinctions, she served in the FTC's Bureau of Economics in the 1950s.

the mistakes of its past. Jim Liebeler would have been happy with the progress to date, but impatient to continue with the job ahead. So am I.

Revised November 7, 2000

ANTITRUST IN THE NEW ECONOMY

Richard A. Posner[1]

Concern has been expressed recently that U.S. antitrust law may not be well suited to regulating the "new economy." Doctrines developed to deal with competition and monopoly in smokestack industries is not well adapted, it is argued, to dealing with the dynamic economy of the twenty-first century. What I shall argue is that there is indeed a problem with the application of antitrust law to the new economy, but that it is not a doctrinal problem; antitrust doctrine is supple enough, and its commitment to economic rationality strong enough, to take in stride the competitive issues presented by the new economy. The real problem lies on the institutional side: the enforcement agencies and the courts do not have adequate technical resources, and do not move fast enough, to cope effectively with a very complex business sector that changes very rapidly. This problem will be extremely difficult to solve; indeed, I cannot even glimpse a complete solution, although I will suggest some partial solutions.

I.

I shall use the term the "new economy" to denote three distinct though related industries. The first is the manufacture of computer software. The second consists of the Internet-based businesses (Internet access providers, Internet service providers, Internet content providers), such as AOL and Amazon. And the third consists of communications services and equipment designed to support the

---

[1] Judge, U.S. Court of Appeals for the Seventh Circuit; Senior Lecturer, University of Chicago Law School. Previous versions of this paper were given as talks in New York City on September 14, 2000, at a conference on antitrust sponsored by the American Bar Association and the American Law Institute, and the next day at a meeting of the Council of Foreign Relations; at DePaul University in Chicago on September 21, 2000, as a lecture in the Distinguished Lecture Series in Finance; as a presentation at the Law and Economics Workshop of the University of Chicago Law School on October 2, 2000; and as a lecture at the University of Hamburg on November 6, 2000. I thank Dennis Carlton, Frank Easterbrook, Eric Posner, William Landes, and participants at the various sessions at which the paper was given for their very helpful comments.

first two markets. There are other candidates for inclusion in the new economy, but these three will do for my purposes.

These industries differ markedly from most of the industries in which modern antitrust doctrine emerged, and particularly from industries that manufacture traditional physical goods such as steel, automobiles, pipe, wire, aluminum, railroad cars, roadbuilding materials, and cigarettes. The traditional industries are characterized by multiplant and multifirm production (indicating that economies of scale are limited at both the plant level and the firm level, or in other words that average total costs are rising at relatively modest output levels), stable markets, heavy capital investment, modest rates of innovation, and slow and infrequent entry and exit. The new-economy industries that I'll be discussing tend to lack these features. They are characterized instead by falling average costs (on a product, not firm, basis) over a broad range of output, modest capital requirements relative to what is available for new enterprises from the modern capital market, very high rates of innovation, quick and frequent entry and exit, and economies of scale in consumption (also known as "network externalities"), the realization of which may require either monopoly or interfirm cooperation in standards setting. And while vertical integration is a common feature of the old economy, it tends to be even more common in the new one, precipitating an unusually large number of firms into customer or supplier relations with other firms that are also its competitors.

A key to understanding these industries is that, with the partial exception of communications equipment, their principal output is intellectual property, namely computer code, rather than physical goods. (Although the intellectual property may be shipped on a disk, this is incidental, and anyway software is increasingly shipped to the purchaser over the Internet.) This is obviously the case for computer software but is also true to a large extent of the Internet-based businesses, even those that sell physical products such as groceries and books. Their ability to take and fill orders and carry out the other operations (such as marketing, billing, handling returns, responding to customer questions and complaints, and allaying the customer's privacy and security concerns) required to give their customers whatever goods or services the business provides is a function to a large extent both of the sophistication of the business's computer software and of its trademarks and copyrights.

Intellectual property is characterized by heavy fixed costs relative to marginal costs. It is often very expensive to create, but once it

is created the cost of making additional copies is low, dramatically so in the case of software, where it is only a slight overstatement to speak of marginal cost as zero.[2] Without legal protection, the creator of intellectual property may be unable to recoup his investment, because competitors can free ride on it; and so legal protection can expand output rather than, as in the usual case of monopoly, reduce it. At the same time, legal protection of intellectual property may operate to deflect consumers to more costly substitutes. The owner of the patent or copyright will charge a positive price for copying, even though the marginal cost may be zero; the positive price will deflect some consumers to substitutes. To prevent these defections would require perfect price discrimination, which is infeasible because (administrative costs to one side) it would require the seller to have complete information about the elasticity of the demand for his product by all his customers and potential customers.

The patent and copyright laws try to strike the output-maximizing balance by giving the creator of intellectual property some but not complete protection from competition. A copyright confers protection for a longer period than a patent does because, traditionally, less was protected—just the form in which the composer, painter, writer, etc. had chosen to express his ideas, and just copying, not independent discovery. The extension of copyright to software has been controversial. Many observers believe that it confers excessive protection, in the sense either of restricting rather than expanding the amount of intellectual property or of attracting excessive resources into the production of that property relative to investment in markets that do not yield a monopoly return (or both), because of difficulties in inventing around an innovative code. Even worse, it is argued, the methods of distributing software often enable the creator to obtain by contract even more protection than copyright law gives him. Copyright gives the holder a property right in his intellectual property even when it is in the hands of a person with whom he has no contract, such as the purchaser of a copyrighted book from a bookseller. To the extent that the creator of software contracts directly with the ultimate purchaser, he can impose via contract more restrictions than the copyright law would allow him to do in the absence of contract;[3] for example, he can forbid the pur-

---

[2] But still an overstatement, because there are selling and servicing costs associated with each sale or rental of software.

[3] See Mark A. Lemley, "Beyond Preemption: The Law and Policy of Intellectual Property Licensing," 87 *California Law Review* 111, 124–134 (1999).

chaser to make an extra copy for his own use, as copyright law permits. It is true that a contract is unlikely to have the same duration as copyright protection, but length of protection is academic in the case of software, which becomes obsolete long before the copyright on it expires. When discrimination is less than perfect, the effect on output is indeterminate, as I'll explain shortly.

The possibility that the combination of copyright and contract gives software manufacturers too much monopoly power in the economic sense, that is, either causes a lessening rather than an increase in the output of the intellectual property in question or distorts the allocation of resources between industries that produce intellectual property and industries that do not, or has both bad effects, creates a natural concern with any further practice or circumstance that might increase the manufacturer's power over the price of his software. It is at this point that another feature of the new economy that I mentioned, economies of scale in consumption, becomes troublesome. Economies of scale in manufacture are familiar; up to a point, the longer the production run, the lower average cost is. Economies of scale in consumption refer to the situation in which the larger the firm's output is (up to some point), the more valuable that output is to its customers. The traditional example was the telephone. Telephone service is worthless if there is only one subscriber; he has no one to talk to. The more subscribers, the more valuable the service is to each one, or at least to many of them. Interactive services, such as email and online auctions, are similar. Likewise the sharing of computer programs, as where two or more academics collaborate on writing a scholarly article by means of word-processing and spreadsheet programs. Literal networking or sharing to one side, computer programs tend to be more valuable the more people use them because training, support by IT personnel, and standardization of equipment and procedures are facilitated. It is the same reason that the typewriter keyboard is standardized. We might call this "figurative" networking, by analogy to the distinction between literal and figurative queuing (for example, standing in line to buy a ticket versus waiting for one's case to be called for trial).

These economies of consumption presuppose uniformity rather than a common source. The international telephone system is a single network, but its components are owned by a vast number of separate firms and individuals. The components have, however, been standardized to assure interoperability, in the same way that the gauge of the railroad track has been standardized throughout the

nation. A firm that manufactures one of the essential components of a network (a term I am using broadly, to denote any situation in which there are economies of scale in consumption) would prefer to be the exclusive source of that component rather than be required to disclose the information that would enable competitors to duplicate it. If the component is subject to intellectual-property protection through patent, copyright, or contract, or can be held as a trade secret, then the requisite uniformity may be more readily achievable by monopoly provision than by standardization.

Networks are not valuable to the consumer in themselves; they are conduits for the services that the consumer values. This is one point at which vertical integration enters the new economy. An operating system is a platform for software applications, and so the writer of operating-system software may decide to write software applications to ride on it, in much the same way that AT&T used to manufacture the terminal equipment attached to its telephone lines. Modern operating systems are themselves composites of separate programs which may be provided by separate companies or by one company; in the latter case there is an analogy to AT&T s practice of manufacturing the switching equipment for its telephone system as well as the telephone lines themselves. Firms that provide dial-up connections and other facilities for accessing and browsing the World Wide Web can integrate forward into the provision of Web-based services such as shopping and video.

The features of the new economy that I have been describing, all but the last (vertical integration, of which more later), tug it toward monopoly yet also, oddly, toward competition. The paradox is dissolved by remembering that competition to *obtain* a monopoly is an important form of competition. The more protection from competition the firm that succeeds in obtaining a monopoly will enjoy, the more competition there will be to become that monopolist; and provided that the only feasible or permitted means of obtaining the monopoly are socially productive, this competition may be wholly desirable. A firm that will have the protection both of intellectual-property law and of economies of scale in consumption if it is the first to come up with an essential component of a new-economy product or service will have a lucrative monopoly, and this prospect should accelerate the rate of innovation, in just the same way that, other things being equal, the more valuable a horde of buried treasure is, the more rapidly it will be recovered. What is more, the successful monopolist is likely to be a firm that initially charges a very low price for the

new product that it has created. Think back to the telephone. Since every new subscriber increases the value of the service to the existing subscribers, a telephone company has an incentive to provide price inducements to new subscribers, as the money it will lose on them may be more than made up for by the higher price that existing subscribers will pay for access to a larger network. This is especially likely if the network will be a natural monopoly, in the sense that no competitor would find it feasible to duplicate it—then the faster the network reaches maturity the longer the monopolist will be protected from challenges to his monopoly. The prospect of a network monopoly should thus induce not only a high rate of innovation but also a low-price strategy that induces early joining and compensates the early joiners for the fact that eventually the network entrepreneur may be able to charge a monopoly price.

I emphasize "may be able" in the preceding sentence. Traditional networks such as the telephone system and the railroads required enormous capital investments and were therefore difficult to duplicate. If you owned such a network, or an important part of it, you had a pretty secure monopoly. The less capital investment the creation of a substitute network involves, the less secure the network monopolist's monopoly is. Because of the extraordinary pace of innovation not only in computer software but also in communications technology, the extraordinary amount of capital that is available worldwide for investment in new enterprises, and the rapidity with which new networks that are primarily electronic can be put into service, the networks that have emerged in the new economy do not seem particularly secure against competition. We have seen all manner of firms rise and fall in this industry—falling sometimes from what had seemed a secure monopoly position. The gale of creative destruction that Schumpeter described, in which a sequence of temporary monopolies operates to maximize innovation that confers social benefits far in excess of the social costs of the short-lived monopoly prices that the process also gives rise to, may be the reality of the new economy.

The feasibility of challenging an existing network monopolist is critical, however. Even if the only way to become a network monopolist in the new economy is to be the first to come up with a new technology that benefits consumers, the existence of the monopoly may discourage subsequent technological innovation by other firms. If network externalities are large, they may give the monopolist a cost advantage that exceeds the benefit of a superior new technology.

This is the issue of "path dependence": an industry may be stuck with an inferior technology because of the cost advantage of the existing network.[4]

Appearances to the contrary notwithstanding, the antitrust laws are not much concerned with monopoly as such.[5] It is not a violation of those laws to acquire a monopoly by lawful means, and those means include innovations protected from competition by the intellectual-property laws. If copyright protection of computer software is too broad (too much intellectual-property protection can reduce output, and, for that matter, too much innovation can be inefficient too[6]), that is a matter to take up with Congress. Nor is it a violation of antitrust law to charge a monopoly price, or to price discriminate in an effort to maximize monopoly profits. And now that the *Alcoa* doctrine[7] is discredited, it is understood that a monopolist is free to compete, whether against the competitive fringe in its monopoly

---

[4] See, for example, Stanley M. Besen and Joseph Farrell, "Choosing How to Compete: Strategies and Tactics in Standardization," *Journal of Economic Perspectives*, Spring 1994, p. 117; Joseph Farrell and Garth Saloner, "Installed Base and Compatibility: Innovation, Product Preannouncements, and Predation," 76 *American Economic Review* 940 (1986); Michael L. Katz and Carl Shapiro, "Technology Adoption in the Presence of Network Externalities," 94 *Journal of Political Economy* 922 (1986).

[5] This is what makes Eastman Kodak Co. v. Image Technical Services, Inc., 504 U.S. 451 (1992), involving the tying of repair services to the purchase of the product, along with most other tying cases, so questionable, as it was plain that, at worst, Eastman Kodak was merely exploiting its customers, not excluding or discouraging equally or more efficient competitors. Benjamin Klein, "Market Power in Antitrust: Economic Analysis after *Kodak*," 3 *Supreme Court Economic Review* 43 (1994).

[6] A patent confers a lawful monopoly, but the expenditures on obtaining a patent may exceed the social benefits of that lawful monopoly. Imagine a case in which those social benefits, which I'll assume are equal to the royalties that the patentee will obtain, are $10 million, and the optimal path to creating the patentable innovation involves two firms each spending $4 million. The firm that creates it first will obtain the patent. Suppose that after each has spent $3 million, they estimate that it will take another six months to complete their respective patent applications. Since the expenditure of $6 million is a sunk cost, each will spend up to an additional $10 million to complete his application a day before his rival; but the social benefit of the acceleration in the completion day will be negligible. Monopoly in this example draws excessive resources into innovating.

[7] United States v. Aluminum Co. of America, 148 F.2d 416 (2d Cir. 1945) (L. Hand, J.).

market or against potential competitors, as vigorously as a firm in an ordinary competitive market would be, provided it doesn't employ tactics calculated to drive an equally or more efficient firm from the market.[8] The fact that a monopolist buttressed by network externalities may be hard to dislodge even by a firm with a superior technology has no antitrust significance in itself. What is true is that a firm is forbidden to enter into a price-fixing agreement with its competitors, or to acquire a competitor if the acquisition will alter the structure of the market to make it much more conducive to price fixing, but these rules do not require any modification to deal with mergers and price fixing in new-economy industries. Nor is there a flat rule against communication among competitors on matters such as standards setting where consumers may benefit from a degree of interfirm cooperation.

The focus of concern with the application of antitrust law to the new economy is on the methods by which a firm that has a monopoly share of some market in a new-economy industry might seek to ward off new entrants. The lawyers and economists who express this concern are fearful lest a "Chicago school" approach to antitrust deny the possibility that a single firm, without collaborating with competitors or potential competitors (thus inviting application of the rules against price-fixing and large horizontal mergers), can, at least under new-economy conditions though probably more generally as well, prevent efficient challenges to its monopoly. If the Chicago-school approach so understood is law, these critics want it modified to do service in new-economy antitrust cases.

This is a misunderstanding of the Chicago school, at least if I can be considered a member of the school in good standing.[9] The approach is skeptical—but no stronger word would be correct—about the danger to competition that is posed by unilateral firm action,

---

[8] See, for example, United States v. Syufy Enterprises, 903 F.2d 659, 669–669 (9th Cir. 1990); Olympia Equipment Leasing Co. v. Western Union Telegraph Co., 797 F.2d 370, 375–376 (7th Cir. 1986); Foremost Pro Color, Inc. v. Eastman Kodak Co., 703 F.2d 534, 544 (9th Cir. 1983). These cases reflect a series of emphatic post-*Alcoa* statements by the Supreme Court that the antitrust laws protect competition in the sense of efficient business practices rather than in the sense of rivalry per se. See, for example, Broadcast Music, Inc. v. Columbia Broadcasting System, 441 U.S. 1, 19–20 (1979); Reiter v. Sonotone Corp., 442 U.S. 330, 343 (1979).

[9] See Richard A. Posner, *Antitrust Law: An Economic Perspective*, ch. 8 (1976).

unilateral in the special sense that it does not require cooperation with competitors (it usually requires cooperation with customers or suppliers). The approach emphasizes both the difficulty of squashing competition by such means and the danger that heavy-handed antitrust enforcement may suppress a practice that may seem anticompetitive but actually is efficient, or at least neutral, from the broader social standpoint. A classic example is the tying agreement, which used to be thought a means by which a firm having a monopoly of one market (the market for the tying product) could obtain a second monopoly (over the market for the tied product). Overlooked was the fact that the cases involved complementary products (hammers and nails, for example), so that, by definition of complementarity, an attempt to increase the price of one would reduce the demand for the other.[10] Owning both monopolies would produce a net increase in the monopolist's profits only if the second monopoly enabled him to engage in price discrimination more easily, with sales of the tied product being used to monitor the intensity of the consumer's demand for the service (such as hammering nails) that the products jointly produce. Although price discrimination has no general tendency to increase efficiency,[11] banning one form is unlikely to do any good, since there is no general antitrust prohibition against price discrimination.

---

[10] A more fundamental point, however, is that tying *makes* the tied product a complement of the tying one; the buyer cannot get the tying product without the tied product and will therefore pay less for the tying product the higher the price for the tied product. The significance of the fact that the tied product is usually a complement of the tying product independently of the tie is that the quantity demanded of the tied product is a measure of the intensity of the buyer's demand for the tying product.

[11] Perfect price discrimination would bring about the same output as under competition, because no customer willing to pay the seller's marginal cost would be turned away. But perfect price discrimination is infeasible, and imperfect price discrimination can result in a lower or higher output than under competition, or the same output. See Frederick M. Scherer and David Ross, *Market Structure and Industrial Performance* 494–496 (3d ed. 1990); Paul A. Samuelson, *Foundations of Economic Analysis* 42–45 (1947); Joan Robinson, *The Economics of Imperfect Competition* 188–195 (1933). Many economists believe that even crude discrimination is more likely to expand than to reduce output, see, for example id. at 201; Scherer and Ross, *supra*, at 494–496; Peter O. Steiner, Book Review, 44 *University of Chicago Law Review* 873, 882 (1977), but there does not appear to be a firm basis for this belief. See Hal R. Varian, "Price Discrimination," in *Handbook of Industrial Organization*, note 12 above, vol. 2, pp. 597, 629–633.

Or consider exclusive dealing. A manufacturer of some consumer product requires its distributors to agree not to carry any potential competitor's products, thus increasing (he hopes) the cost of distribution to potential competitors. With an important exception noted below for the case where there are economies of scale in distribution, the manufacturer will have to compensate its distributors for agreeing to the restriction, and this will increase its costs, making entry more attractive. If a potential competitor has a promising product, other distributors will be delighted to carry it; if there are no other distributors, new ones will appear. Even if the potential competitor has to do its own distribution, and thus enter on two levels, manufacturing and distribution, the fact that the capital requirements for entry are now greater should not be an obstacle to entry, since there is no shortage of capital for promising new ventures. For these reasons, exclusive dealing is unlikely to be an effective means of forestalling entry. At the same time, it is easy to see how exclusive dealing might promote efficiency by increasing the likelihood that a distributor will use his best efforts to promote the manufacturer's goods. Exclusive dealing may also help a seller of intellectual property to prevent piracy, a serious concern in intellectual-property markets.

But skepticism about unilateral monopolizing actions is not the same as denial. The *Standard Fashion* case[12] has long been used by some members of the Chicago school to illustrate a unilateral action that might well have increased the defendant's monopoly power.[13] The defendant manufactured a line of women's dress patterns (which women used to make their own dresses) that was very popular. Retailers thought it essential to be able to sell the line. The defendant required retailers to agree not to carry competing lines. Competing manufacturers could in principle create their own retail outlets, but who would shop there if the most popular brand could not be found? Competing manufacturers would have to create a line as long and as popular as Standard Fashion's line, and that would be difficult, maybe impossible, to do.

---

[12] Standard Fashion Co. v. Magrane-Houston Co., 258 U.S. 346 (1922).

[13] Aaron Director and Edward H. Levi, "Law and the Future: Trade Regulation," 51 *Northwestern University Law Review* 281, 293 (1956); Richard A. Posner, *Antitrust Law: An Economic Perspective* 202 n. 48 (1976). See generally John E. Lopatka and William H. Page, "Posner's Program for the Antitrust Division: A Twenty-Five Year Perspective," 48 *SMU Law Review* 1713, 1734, 1734 (1995).

What distinguished *Standard Fashion* from a garden-variety exclusive-dealing case was the existence of economies of scale at the distribution level. Consumers didn't want to traipse from store to store. They wanted a full line in each store, so anyone entering the women's clothing business had to provide the full line if it was excluded from stores that carried the dominant firm's line. Restricting its retailers no doubt cost Standard Fashion something. But it is plausible that the cost was less than the increase in its expected monopoly profits from forestalling new entry by compelling any prospective entrant to enter on a full-line basis.

The point is not that the new entrant would have to invest more capital, as in the previous example. The point is that it would have to embark on a riskier undertaking, that of creating not a single successful product but a whole line of such products. It's as if one couldn't make commercial aircraft without making military aircraft as well. In such a case, as in *Standard Fashion* itself, exclusive dealing, while it would increase the defendant's net costs of distribution (if it were indeed adopted for anticompetitive purposes and had no efficiency rationale), might increase the costs (not the capital requirements) of new entrants much more.

The analogy to a new-economy network externality should be plain. The network corresponds to the full-line retail store in *Standard Fashion*. A firm may wish to enter the market by producing one component of the network or one value-added service, but if a competitor by virtue of owning or having an exclusive-dealing contract with the network refuses to cooperate with the firm, the firm won't be able to duplicate the network in order to get distribution of its component or service.

It is true that piecemeal entry is the norm in many industries. A department store carries the products of many producers, many of which do not offer a full line of products. One can imagine a number of designers of women's dress patterns, each specializing in one pattern, and the department store assembling them into a full line to compete with Standard Fashion's full line. The risk of entry to each designer would be minimized. The case was decided in 1922, however, and many towns may not have had department stores. We may be in a similar stage in the development of the new economy, where distribution facilities may be sufficiently limited to create bottlenecks that monopolists can exploit to perpetuate monopoly.

It is important to note that the monopolist would have no incentive to engage in exclusionary conduct unless his monopoly were

fragile, that is, vulnerable to new entry. The more vulnerable it is, however, the less likely it is to endure even if the monopolist does resort to such conduct. This point is missed in Robert Bork's criticism of the decision.[14] He argues that Standard couldn't extract a monopoly price from its dealers twice, once by charging them what the market would bear and then by forcing them to enter into exclusive-dealing contracts. That's true, but what Standard may have been able to do was to increase the duration of its monopoly.

The likelihood that the monopoly profits obtained during the extension period (as we may call the period for which a monopoly is extended by means of exclusionary practices) will exceed the costs of the exclusionary practice to the monopolist is enhanced by the fact that, both in *Standard Fashion* and even more dramatically in parallel new-economy cases, the monopoly is of intellectual property. Intellectual property is characterized as we know by a large difference between total and marginal cost. In the extreme case, which is approximated in some software markets, marginal cost is close to zero, meaning that almost all the revenues earned by a firm that monopolizes the market go directly to the bottom line. This makes it plausible that the profit from extending the monopoly another year or two will exceed the cost of the exclusionary practices required to achieve the extension. To put this differently, there is no reason to think that the cost of an exclusionary practice in such markets will exceed the additional monopoly profits that the practice makes possible, even though the cost may be incurred earlier, which requires that any future profits be discounted to present value before being compared to the cost.

A complicating factor in both *Standard Fashion* and new-economy cases is the utility of exclusive dealing as a means of dealing with piracy. *Standard Fashion* may have been worrying about the piracy of its patterns by competing manufacturers, and attempting to prevent this by denying outlets to the pirates. (This concern was explicit in a later case, *Fashion Originators' Guild of Amer-*

---

[14] Robert H. Bork, *The Antitrust Paradox: A Policy at War with Itself* 305–307 (2d ed. 1993). Bork also argues that Standard Fashion's market share, 40 percent, was too small to enable it to foreclose rivals. But that was its nationwide market share, and the Supreme Court pointed out that in many towns and smaller cities it had a monopoly. It should probably be regarded therefore as a case involving a series of local monopolies rather than a national monopoly.

ica v. FTC.[15]) To anticipate the second part of my discussion, whenever an antitrust court is called on to balance efficiency against monopoly, there is trouble; legal uncertainty, and the likelihood of error, soar.

Exclusive dealing, by the way, is analytically the same as tying. Exclusive dealing ties distribution to manufacture; equivalently, tying is exclusive dealing in the tied product. All you need is sensible law on exclusive dealing to deal sensibly with tying cases. If there are economies of scale in the tied product, a firm that must in order to enter the market for the tying product produce the tied product as well will have higher costs than the monopolist and this will reduce the expected gain to him of entering the market for the tying product. Notice, finally, that exclusive dealing and tying can be accomplished in a variety of different ways, including vertical integration, contract, product design (in the case of tying), and bundling (charging a zero price for the tied product, and so in effect paying the consumer to buy it from the supplier of the tying product). And, speaking of bundling, while documented cases of predatory pricing are rare, it is not always an irrational method of deterring entry. Especially when it takes the form of area price discrimination or "fighting brands," so that the predator does not have to lose more money than the new entrant by lowering prices throughout his market, a monopolist may maximize his profits by "investing" in a reputation for predatory response to threatened entry.

II.

Well, the plaintiff won in *Standard Fashion*, and the decision remains good law and its principle is not limited to exclusive dealing but extends as we have just seen to any business method by which a monopolist can exclude an equally or more efficient competitor; likewise the illegality of predatory pricing is settled antitrust doctrine. I conclude that, as I said at the outset, existing antitrust doctrine is sufficiently supple, and sufficiently informed by economic theory, to cope effectively with the distinctive-seeming antitrust problems that the new economy presents.

What is troublesome is the institutional structure of antitrust enforcement. To begin with, cases in the new economy present unusually difficult questions of fact because of the technical complexity of the products and services produced by new-economy industries.

---

[15] 312 U.S. 457 (1941).

Such questions can be central rather than peripheral to the antitrust issues, as where a plaintiff complains that the defendant has altered a technical protocol or interface to make it more difficult for the plaintiff s product to work with the network, or a defendant contends that disclosure of a protocol would enable its competitors by reverse engineering to deprive it of a valuable trade secret that it cannot feasibly protect by means of copyright or patent law. The second example is an aspect of two broader questions, both also very technical and very difficult—the relative merits of monopoly provision and standardization as methods of optimizing a network, and the adequacy in fact of copyright or patent law relative to trade-secret law to protect investments in software and other new-economy products. Similar questions have arisen in other antitrust network settings, notably in the telephone industry, but they simply are much more difficult for a lay person to understand in the new-economy context.

A further complication is that it is difficult to find truly neutral competent experts to advise the lawyers, judges, and enforcement agencies on technical questions in the new economy. There aren' t that many competent experts, and almost all of them are employed by or have other financial ties to firms involved in or potentially affected by antitrust litigation in this sector. The Antitrust Division does not employ *any* computer scientists or electrical engineers, but is wholly dependent on consultants, as are also, I believe, all the state antitrust offices; and, as I say, it is difficult to find a consultant in the new economy who is both competent and disinterested.

We deal with technical questions in the judiciary not by having judges or jurors who have the requisite technical knowledge or by giving them technical assistants but by having technical experts present evidence which the judge and jury (if it is a jury case) is expected somehow to assimilate. This system does not work as badly as its critics maintain; but the more technical the area of litigation and the fewer experts are disinterested, the worse it is apt to work. Computer science and communications technology are much more difficult areas than the average body of scientific or engineering knowledge that lay judges and jurors are asked to absorb en route to rendering a decision.

One possible measure would be to borrow from arbitration procedure, as follows: In a technically complex case, the judge could direct each party to nominate a technical expert, and the two experts selected in this way would then agree upon a third, a neutral, expert,

whom the judge would appoint as a court-appointed expert[16] in confidence that he would not be misled by a partisan. The problem, as I mentioned earlier, is that there may be very few genuine neutrals in the high-technology sector who have the expertise required for a particular case, and they may be reluctant to become involved in a litigation. Setting that problem aside, a technical committee composed of two party-nominated and one neutral technical expert might assist the judge or special master in the administration of a consent or litigated decree if one were entered in an antitrust case against a new-economy company. This is more needful than in the case of warring economic experts, because the economic principles that bear on antitrust can be explained in intuitive terms; technical principles often cannot be.

The difficult factual questions presented by new-economy antitrust cases are not limited to technical areas, unfortunately. The combination of intellectual property, network externalities, and rapid growth in consumer demand creates difficult questions involving the ascertainment and measurement of monopoly. Suppose a firm has 100 percent of some new-economy product market and charges a price for it that is greatly in excess of marginal cost. Suppose also that the product is one that involves economies of scale in consumption (that is, it is a network market) and the demand for it is growing very rapidly. This is a common new-economy scenario. Does the firm have a monopoly in an economically relevant sense? The ratio of price to marginal cost is meaningless, since often pricing intellectual property at marginal cost is nonremunerative (because marginal cost is lower than average total cost, owing to heavy fixed costs) and leads to bankruptcy. What about the market share? But one characteristic of intellectual property is its durability, and the more durable a product is, the fewer repeat sales the manufacturer will have. (It is different if the firm leases rather than sells its product.) New-economy firms that don't lease get around this problem with frequent upgrades, but the installed base acts as a drag on the price that can be charged for an upgrade, since the prospective customer for the upgrade has the option of sticking with the existing product. If demand is growing rapidly, moreover, the firm in deciding on a price must trade off current against future profits, since a higher price today will slow the growth in demand. And by the assumption that it is a network market, the firm will have a strong in-

[16] As expressly authorized by Fed. R. Evid. 706.

centive to charge a very low price in order to increase usage because that will enable it to charge a much higher price in the future—unless rapidity of innovation makes the future wholly uncertain, in which event it may abandon the hope of a network monopoly and charge a high price now. Still another factor is that a vertically integrated firm competes against itself: a firm that owns a network but also sells value-added services may want to keep the price of network access low to increase sales of those services. This is the other side of the tying coin: holding down the price of the tied product will keep up the demand for the tying product. The result of all these factors may be that a firm has a monopoly market share only because it is not charging a monopoly price.

And while it is possible to argue that monopoly can have other bad effects besides the limitation of output over the competitive level that monopoly pricing brings about, namely a reduction in the rate or distortion in the direction of innovation, economic theory and empirical evidence have yet to generate a consensus on whether monopoly is on balance good or bad for innovation. The possible effect of network externalities in discouraging subsequent innovation (the "path dependence" problem) not only is speculative, but is operative even if the monopolist is passive, yet in that case there would be no arguable antitrust violation.

The peculiarities of new-economy markets that I have been describing are apt to make the trial of a new-economy case a daunting challenge to the factfinding capacity of the judiciary. Nor are the economic issues that I have identified issues that can be left to a neutral expert to resolve, given the many imponderables they involve; they are not problems simply to be solved by the application of the relevant economic principles.

The rapidity of innovation in the new economy has another very important institutional implication. Federal courts are highly efficient by the standards of the American legal system. The federal court queue is short, and strong district judges can move even complex cases along briskly. But this is speaking relatively. Antitrust litigation moves very slowly relative to the new economy. Law time is not real time. The law is committed to principles of due process that limit the scope for summary proceedings, and the fact that litigation is conducted by lawyers before tribunals that are not technically trained or experienced inevitably slows the process.

The mismatch between law time and new-economy real time is troubling in two respects. First, an antitrust case involving a new-

economy firm may drag on for so long relative to the changing conditions of the industry as to become irrelevant, ineffectual. That was a problem even in the old economy. One recalls for example that by the time the monopolization case against Alcoa completed its journey through the courts, Alcoa had lost its monopoly for reasons unrelated to the litigation; as a result, the decree finally entered against Alcoa offered little more than nominal relief (the divestiture of Alcoa's Canadian subsidiary, a substantial producer but one which U.S. tariffs prevented from having a big impact on the U.S. market[17]). This type of problem is likely to be more frequent in the new economy.

Second, even if the case is not obsoleted by passage of time, its pendency may cast a pall over parties to and affected by the litigation, making investment riskier and complicating business planning.

These problems are aggravated by the tendency of antitrust litigation to create multiple lawsuits out of a single dispute; call this the cluster-bomb effect. No sooner does the Antitrust Division bring a case, but the states and now the European Union are likely to join the fray, followed at a distance by the antitrust plaintiffs' class-action bar. The effect is to lengthen out the original lawsuit, complicate settlement, magnify and protract the uncertainty engendered by the litigation, and increase litigation costs.

I wish I had a good solution to these problems, or at least some of them, but I don't. As minor palliatives, I would like to see greater use of neutral experts, as I have explained, and I would also like the Antitrust Division and the Federal Trade Commission to be given the necessary appropriations to enable each of these agencies to hire a competent technical staff.[18] That won't be easy, given the salaries that competent new-economy scientists and engineers command in the private market. But there is a more serious problem, and that is the rapid obsolescence of technical knowledge, which combined with tenure practices in government agencies may prevent the agencies from maintaining a technical staff that is actually abreast of current technology. This is an objection, by the way, to an alternative that I consider distinctly unpromising, which is to have a specialized, technical court composed of judges with technical training. Their training

---

[17] See See Merton J. Peck, *Competition in the Aluminum Industry 1945–1958* (1961), under index references to Limited. (The name of the Canadian subsidiary was "Aluminium Limited.")

[18] I am not proposing a net increase in the appropriations for these agencies, just that they be authorized to allocate a portion of their existing appropriations to hiring technical experts.

is likely to become out of date before they become judges, unless we are to have a system in which the regulation of the new economy is entrusted to 25 year old judges, lawyers, and engineers.

Now turning to more radical measures, I would like to see, first, the states stripped of their authority to bring antitrust suits, federal or state,[19] except under circumstances in which a private firm would be able to sue, as where the state is suing firms that are fixing the prices of goods or services that they sell to the state. (In other words, only their power to bring *parens patriae* suits would be abrogated.) States do not have the resources to do more than free ride on federal antitrust litigation, complicating its resolution; in addition they are too subject to influence by interest groups that may represent a potential antitrust defendant's competitors. This is a particular concern when the defendant is located in one state and one of its competitors in another, and the competitor, who is pressing his state's attorney general to bring suit, is a major political force in that state. A situation in which the benefits of government action are concentrated in one state and the costs in other states is a recipe for irresponsible state action. This is a genuine downside of federalism. The federal government, having a larger and more diverse constituency, is, as James Madison recognized in arguing for the benefits of a large republic, less subject to takeover by a faction. I am not myself inclined to make a fetish of federalism.

When I was a law clerk at the Supreme Court, almost 40 years ago, I was struck by the poor quality of the briefs and arguments of the lawyers in the offices of the state attorneys general. Since becoming a judge almost 20 years ago, I have been struck by the poor quality of the briefs and arguments of most though not all of the lawyers in the offices of the state attorneys general of my circuit. On several occasions our court has found it necessary in published opinions to criticize the performance of these lawyers. In response to one of these opinions a meeting was arranged between the Attorney General of Illinois and myself (I was chief judge then), at which he explained that the very low salary scale of his office makes it difficult to hire competent lawyers.

One of the sources of the quality problem is that in most states the attorney general runs for office independent of the governor and is thus not part of the governor's administration and as a result does

---

[19] That is, Congress should enact a statute that would make antitrust law, like patent, copyright, and bankruptcy law, an exclusively federal body of law enforceable only in federal courts.

not receive much support for the governor in negotiating for appropriations from the legislature. In addition, in a field like antitrust, the offices of the state attorneys general are not able to scale up to anything like the level of specialization and professionalism of the Antitrust Division of the U.S. Department of Justice, with its hundreds of lawyers and economists.[20]

A possible, partial solution to the delay and cluster-bomb problems would be to empower the Justice Department to bring antitrust suits for damages by whomever suffered and to parcel out the proceeds of the suit among the victims. If the Department decided to bring a suit, private suits (and state suits, if they were still permitted) would be preempted. In effect the Department would have a right of first refusal, which it presumably would exercise in major cases in order to consolidate and thus simplify and expedite the litigation. This is not an original suggestion, as it is a power already enjoyed by government agencies such as the Department of Labor and the Equal Employment Opportunities Commission in the field of employment law.[21]

Any suggestion for streamlining antitrust enforcement is likely to be met, from the conservative side of the policy-analysis spectrum, by the argument that it is a mistake to make government more efficient, when efficiency connotes the better matching of means to ends and the ends may be bad ones. I raised this possibility in discussing antitrust enforcement by the states. I must confront it with regard to the federal agencies as well. The essential question is the degree to which these agencies are buffeted by interest-group pressures, as I believe the state antitrust enforcers are. I think that, today, only to a small degree. There is a history of efforts to explain antitrust enforcement as just another example of interest-group politics, an approach that has worked for a number of government agencies and programs. But the efforts have not been successful,[22] and the reasons

---

[20] Cf. Robert H. Lande, "When Should States Challenge Mergers: A Proposed Federal/State Balance," 35 *New York Law School Law Review* 1047, 1064–1066 (1990).

[21] See, for example, 29 U.S.C. § 626(c)(1); EEOC v. G-K-G, Inc., 39 F.3d 740, 744–745 (7th Cir. 1994).

[22] See, for example, William F. Baxter, *The Political Economy of Antitrust* (Robert D. Tollison ed. 1980); George J. Stigler, "The Origin of the Sherman Act," 14 *Journal of Legal Studies* 1 (1985); Malcolm B. Coate, Richard S. Higgins, and Fred S. McChesney, "Bureaucracy and Politics in FTC Merger Challenges," in *The Causes and Consequences of Antitrust: The Public-*

may be that the agencies are dominated by lawyers most of whom go on to service in the private sector and that antitrust law itself is dominated by federal judges exercising a broad discretion because of the open-endedness of the major federal antitrust statutes. To land good berths in the private practice of law the antitrust enforcers must demonstrate their professionalism, which means keeping within the boundaries fixed by the courts. Federal judges with their secure tenure seem largely insulated from the interest-group pressures that play on the other branches of government. Looking over the entire history of U.S. antitrust law, I conclude that the most powerful explanatory variable is simply the state of economic opinion. Antitrust doctrine has changed more or less in tandem with changes in economic theory, albeit with a lag.[23] As a result, I am not fearful of making antitrust enforcement more efficient.

Unfortunately, the measures that I have suggested, even in the unlikely event that they are adopted, would probably not do a great deal to correct what seems to me a serious mismatch between the conditions of the new economy and the institutional structure of antitrust enforcement. That brings me to the final question I address, which is society's proper response to a situation of at least temporarily ineradicable uncertainty concerning the effect of governmental intervention in the economy. We really don't know what the effect of applying antitrust principles to the new economy will be, except when they are applied just to stop horizontal price-fixing or mergers of major competitors in highly concentrated markets. I think a policy of zero enforcement against alleged exclusionary practices in the new economy would be a mistake, because there is a pretty solid theoretical basis for concern both that some new-economy firms would find it in their rational self-interest to employ such practices and that natu-

---

Choice Perspective 213 (Fred S. McChesney and William F. Shughart II eds. 1995). But see Roger L. Faith, Donald R. Leavens, and Robert D. Tollison, "Antitrust Pork Barrel," 25 *Journal of Law and Economics* 329 (1982), finding evidence that the FTC in the 1960s was unduly influenced in its enforcement activity by the parochial interests of members of Congress who had budgetary and oversight powers with respect to the Commission.

[23] As Louis Kaplow has pointed out, it is not true that the federal courts in antitrust cases suddenly discovered economics in the 1970s. Kaplow, "Antitrust, Law and Economics, and the Courts," *Law and Contemporary Problems*, Autumn 1987, pp. 181, 187. See also Herbert Hovenkamp, "The Antitrust Movement and the Rise of Industrial Organization " 68 *Texas Law Review* 105 (1989).

ral market forces would not undo those practices in time to avoid significant social costs. A policy of zero enforcement would also deprive us of important information about competition and monopoly in this vital sector of the national economy. Cleary, though, the byword of a prudent enforcement agency and a sensible court will be: caution.

Sunday

May 20, 2012

Testimony of Sir Timothy Berners-Lee
CSAIL Decentralized Information Group
Massachusetts Institute of Technology

Before the
United States House of Representatives
Committee on Energy and Commerce
Subcommittee on Telecommunications and the Internet

Hearing on the "Digital Future of the United States: Part I -- The Future of the World Wide Web"

This document on the Web [http://dig.csail.mit.edu/2007/03/01-ushouse-future-of-the-web] [PDF]

Chairman Markey, Ranking Member Upton, and Members of the Committee. It is my honor to appear before you today to discuss the future of the World Wide Web. I would like to offer some of my experience of having designed the original foundations of the Web, what I've learned from watching it grow, and some of the exciting and challenging developments I see in the future of the Web. Though I was privileged to lead the effort that gave rise to the Web in the mid-1990s, it has long passed the point of being something designed by a single person or even a single organization. It has become a public resource upon which many individuals, communities, companies and governments depend. And, from its beginning, it is a medium that has been created and sustained by the cooperative efforts of people all over the world.

To introduce myself, I should mention that I studied Physics at Oxford, but on graduating discovered the new world of microprocessors and joined the electronics and computer science industry for several years. In 1980, I worked on a contract at CERN, the European Particle Physics Laboratory, and wrote for my own benefit a simple program for tracking the various parts of the project using linked note cards. In 1984 I returned to CERN for ten years, during which time I found the need for a universal information system, and developed the World Wide Web as a side project in 1990. In 1994, the need for coordination of the Web became paramount, and I left to come to MIT, which became the first of now three international host institutes for the World Wide Consortium (W3C). I have directed W3C since that time. I hold the 3Com Founders chair at MIT where I pursue research on advanced Web technologies with the MIT Decentralized Information Group. The testimony I offer here today is purely my own opinion and does not necessarily reflect the views of the World Wide Web Consortium or any of its Members.

The special care we extend to the World Wide Web comes from a long tradition that democracies have of protecting their vital communications channels. We nurture and protect our information networks because they stand at the core of our economies, our democracies, and our cultural and personal lives. Of course, the imperative to assure the free flow of information has

only grown given the global nature of the Internet and Web. As a Federal judge said in defense of freedom of expression on the Internet:

> The Internet is a far more speech-enhancing medium than print, the village green, or the mails.... The Internet may fairly be regarded as a never-ending worldwide conversation.[1]

Therefore it is incumbent on all of us to understand what our role is in fostering continued growth, innovation, and vitality of the World Wide Web. I am gratified that the United States and many other democracies around the world have taken up this challenge. My hope today is to help you to explore the role this committee and this Congress has in building upon the great advances that are in store for the Web.

I. Foundations of the World Wide Web

The success of the World Wide Web, itself built on the open Internet, has depended on three critical factors: 1) unlimited links from any part of the Web to any other; 2) open technical standards as the basis for continued growth of innovation applications; and 3) separation of network layers, enabling independent innovation for network transport, routing and information applications. Today these characteristics of the Web are easily overlooked as obvious, self-maintaining, or just unimportant. All who use the Web to publish or access information take it for granted that any Web page on the planet will be accessible to anyone who has an Internet connection, regardless whether it is over a dialup modem or a high speed multi-megabit per second digital access line. The last decade has seen so many new ecommerce startups, some of which have formed the foundations of the new economy, that we now expect that the next blockbuster Web site or the new homepage for your kid's local soccer team will just appear on the Web without any difficulty.

Today I will speak primarily about the World Wide Web. I hesitate to point out that the Web is just one of the many applications that run on top of the Internet. As with other Internet applications such as email, instant messaging, and voice over IP, the Web would have been impossible to create without the Internet itself operating as an open platform. [2]

A. Universal linking: Anyone can connect to anyone, any page can link to any page

How did the Web grow from nothing to the scale it is at today? From a technical perspective, the Web is a large collection of Web pages (written in the standard HTML format), linked to other pages (with the linked documents named using the URI standard), and accessed over the Internet (using the HTTP network protocol). In simple terms, the Web has grown because it's easy to write a Web page and easy to link to other pages. The story of the growth of the World Wide Web can be measured by the number of Web pages that are published and the number of links between pages. Starting with one page and one site just about 15 years ago, there are now over 100,000,000 web sites[3] with an estimated over 8 billion publicly accessible pages as of 2005. What makes it easy to create links from one page to another is that there is no limit to the

number of pages or number of links possible on the Web. Adding a Web page requires no coordination with any central authority, and has an extremely low, often zero, additional cost. What's more, the protocol that allows us to follow these links (HTTP) is a non-discriminatory protocol. It allows us to follow any link at all, regardless of content or ownership. So, because its so easy to write a Web page, link to another page, and follow these links around, people have done a lot of this. Adding a page provides content, but adding a link provide the organization, structure and endorsement to information on the Web which turn the content as a whole into something of great value.

A current example of the low barriers to reading, writing and linking on the Web is the world of blogs. Blogs hardly existed five years ago, but have become an enormously popular means of expression for everything from politics to local news, to art and science. The low barriers to publishing pages and abundance of linking ability have come together, most recently with blogs, to create an open platform for expression and exchange of all kinds.[4] The promise of being able to reach anyone over a communications system that will carry virtually anything (any type of information) is somewhat like other infrastructures we depend upon: the mail system, the road system, and the telephone system. It stands in contrast to more closed systems such as the broadcast or cable television networks. Those closed systems perform valuable functions as well, but their impact in society is different and less pervasive.

The universality and flexibility of the Web's linking architecture has a unique capacity to break down boundaries of distance, language, and domains of knowledge. These traditional barriers fall away because the cost and complexity of a link is unaffected by most boundaries that divide other media. It's as easy to link from information about commercial law in the United States to commercial law in China, as it is to make the same link from Massachusetts' Commercial Code to that of Michigan. These links work even though they have to traverse boundaries of distance, network operators, computer operating systems, and a host of other technical details that previously served to divide information. The Web's ability to allow people to forge links is why we refer to it as an abstract information space, rather than simply a network. Other open systems such as the mails, the roads or the telephones come to perform a function in society that transcends their simple technical characteristics. In these systems, phone calls from the wireline networks travel seamlessly to wireless providers. Mails from one country traverse borders with minimal friction, and the cars we by work on any roads we can find. Open infrastructures become general purpose infrastructure on top of which large scale social systems are built. The Web takes this openness one step further and enables a continually evolving set of new services that combine information at a global scale previously not possible. This universality has been the key enabler of innovation on the Web and will continue to be so in the future.

B. Open Foundation for Information-driven Innovation

The Web has not only been a venue for the free exchange of ideas, but also it has been a platform for the creation of a wide and unanticipated variety of new services. Commercial applications including eBay, Google, Yahoo, and Amazon.com are but a few examples of the extraordinary innovation that is possible because of the open, standards-based, royalty-free

technology that makes up the Web. Whether developing an auction site, a search engine, or a new way of selling consumer goods, e-commerce entrepreneurs have been able develop new services with confidence that they will be available for use by anyone with an Internet connection and a Web browser, regardless of operating system, computer hardware, or the ISP chosen by that user.[5] Innovation in the non-commercial and government domains has been equally robust. Early Web sites such as Thomas have led the way in efforts to make the legislative process more open and transparent, and non-commercial sites such as the Wikipedia have pioneered new collaborate styles of information sharing. The flexibility and openness inherent in Web standards also make this medium a powerful foundation on which to build services and applications that are truly accessible for people with disabilities, as well as people who need to transform content for purposes other than that for which it was originally intended.

The lesson from the proliferation of new applications and services on top of the Web infrastructure is that innovation will happen provided it has a platform of open technical standards, a flexible, scalable architecture, and access to these standards on royalty-free ($0 fee patent licenses) terms. At the World Wide Web Consortium, we will only standardize technology if it can be implemented on a royalty-free basis. So, all who contribute to the development of technical standards at the W3C are required to agree to provide royalty-free licenses to any patents they may hold if those patents would block compliance with the standard. [6] Consider as a comparison the very successful Apple iTunes+iPod music distribution environment. This integration of hardware, software, Web service shows an intriguing mix of proprietary technology and open standards. The iTunes environment consists of two parts: sales of music and videos, and distribution of podcasts. The sale of music is managed by a proprietary platform run by Apple with the aim of preventing copyright infringement. However, because Apple uses closed, non-standard technology for its copy protection (known as Digital Rights Management), the growth is seen as limited. In fact, Apple CEO Steve Jobs recently wrote that the market for online music sales is being limited by the lack of open access to DRM technology.[7] By contrast, the podcast component of iTunes is growing quite dramatically, providing a means for many small and large audio and video distributors to share or sell their wares on the Web. Unlike the music and video sales, podcasts are based on open standards, assuring that it's easy to create, edit and distribute the podcast content.

C. Separation of Layers

When, seventeen years ago, I designed the Web, I did not have to ask anyone's permission. The Web, as a new application, rolled out over the existing Internet without any changes to the Internet itself. This is the genius of the design of the Internet, for which I take no credit. Applying the age old wisdom of design with interchangeable parts and separation of concerns, each component of the Internet and the applications that run on top of it are able develop and improve independently. This separation of layers allows simultaneous but autonomous innovation to occur at many levels all at once. One team of engineers can concentrate on developing the best possible wireless data service, while another can learn how to squeeze more and more bits through fiber optic cable. At the same time, application developers such as myself can develop new protocols and services such as voice over IP, instant messaging, and peer-to-

peer networks. Because of the open nature of the Internet's design, all of these continue to work well together even as each one is improving itself.

II. Looking forward

Having described how the Web got to where it is, let us shift to the question of where it might go from here. I hope that I've already persuaded you that the evolution of the Web is not in the hands of any one person, me or anyone else. But I'd like to highlight three areas in which I expect exciting developments in the near future. First, the Web will get better and better at helping us to manage, integrate, and analyze data. Today, the Web is quite effective at helping us to publish and discover documents, but the individual information elements within those documents (whether it be the date of any event, the price of a item on a catalog page, or a mathematical formula) cannot be handled directly as data. Today you can see the data with your browser, but can't get other computer programs to manipulate or analyze it without going through a lot of manual effort yourself. As this problem is solved, we can expect that Web as a whole to look more like a large database or spreadsheet, rather than just a set of linked documents. Second, the Web will be accessible from a growing diversity of networks (wireless, wireline, satellite, etc.) and will be available on a ever increasing number of different types of devices. Finally, in a related trend, Web applications will become a more and more ubiquitous throughout our human environment, with walls, automobile dashboards, refrigerator doors all serving as displays giving us a window onto the Web.

A. Data Integration

Digital information about nearly every aspect of our lives is being created at an astonishing rate. Locked within all of this data is the key to knowledge about how to cure diseases, create business value, and govern our world more effectively. The good news is that a number of technical innovations (RDF which is to data what HTML is to documents, and the Web Ontology Language (OWL) which allows us to express how data sources connect together) along with more openness in information sharing practices are moving the World Wide Web toward what we call the Semantic Web. Progress toward better data integration will happen through use of the key piece of technology that made the World Wide Web so successful: the link. The power of the Web today, including the ability to find the pages we're looking for, derives from the fact that documents are put on the Web in standard form, and then linked together. The Semantic Web will enable better data integration by allowing everyone who puts individual items of data on the Web to link them with other pieces of data using standard formats.

To appreciate the need for better data integration, compare the enormous volume of experimental data produced in commercial and academic drug discovery laboratories around the world, as against the stagnant pace of drug discovery. While market and regulatory factors play a role here, life science researchers are coming to the conclusion that in many cases no single lab, no single library, no single genomic data repository contains the information necessary to discover new drugs. Rather, the information necessary to understand the complex interactions

between diseases, biological processes in the human body, and the vast array of chemical agents is spread out across the world in a myriad of databases, spreadsheets, and documents.

Scientists are not the only ones who need better data integration. Consider the investment and finance sector, a marketplace in which profit is generated, in large part, from having the right information, at the right time, and reaching correct conclusions based on analysis and insight drawn from that information. Successful investment strategies are based on finding patterns and trends in an increasingly diverse set of information sources (news, market data, historical trends, commodity prices, etc.). Leading edge financial information providers are now developing services that allow users to easily integrate the data they have, about their own portfolios or internal market models, with the information delivered by the information service. The unique value creation is in the integration services, not in the raw data itself or even in the software tools, most of which will be built on open source components.

New data integration capabilities, when directed at personal information, pose substantial privacy challenges which are hardly addressed by today's privacy laws. The technology of today's Web already helps reveal far more about individuals, their behavior, their reading interest, political views, personal associations, group affiliations, and even health and financial status. In some cases, this personal information is revealed by clever integration of individual pieces of data on the Web that provide clues to otherwise unavailable information. In other cases, people actually reveal a lot about themselves, but with the intent that it only used in certain contexts by certain people. These shifts in the way we relate to personal information require serious consideration in many aspects of our social and legal lives. While we are only just beginning to see these shifts, now is the time to examine a range of legal and technical options that will preserve our fundamental privacy values for the future without unduly stifling beneficial new information processing and sharing capabilities. Our research group at MIT is investigating new technologies to make the most of the Semantic Web, as well as both technical and public policy models that will help bring increased transparency and accountability to the World Wide Web and other large scale information systems.[8] Our belief is that in order to protect privacy and other public policy values, we need to research and develop new technical mechanisms that provide great transparency into the ways in which information in the system is used, and provide accountability for those uses with respect to what ever are the prevailing rules.

B. Network Diversity and Device Independence

The Web has always been accessible from a variety of devices over a variety of networks. From early on, one could browse the Web from a Macintosh, a Windows PC or a Linux-based computer. However, for a long time the dominant mode of using the Web was from some desktop or laptop computer with a reasonably large display. Increasingly, people will use non-PC devices that have either much smaller or much larger displays, and will reach the Internet through a growing diversity of networks. At one end of this spectrum, the devices will seem more like cell phones. At the other end, they will seem more like large screen TVs. There are, of course, technical challenges associated with squeezing a Web page designed for a 17 inch screen into the two to four inch display available on a mobile phone or PDA. Some of these will be

solved through common standards and some through innovative new interface techniques. All of this means more convenience for users and more opportunity for new Web services that are tailored to people who are somewhere other than their desks.

Growth in access networks and Web-enabled applications presents a number of important opportunities. For example, more robust, redundant network services together with innovative uses of community-based social networks on the Web are coming to play an increasing role in areas such as emergency planning and notification.[9] Reports about ad hoc communication networks supporting disaster relief efforts are just one illustration of the benefit of the openness, flexibility and accessible of the Internet and Web. This one area is a microcosm of many of the issues that we are discussing today, because in order to work well it requires seamless integration of diverse types of data; repurposing that data instantly into valid formats for a myriad of different Web devices; and including appropriate captions, descriptions, and other necessary accessibility information. I would encourage all web sites designers to ensure that their material conforms not only to W3C standards, but also to guidelines for accessibility for people with disabilities, and for mobile access.

C. Ubiquitous Web Applications

In the future, the Web will seem like it's everywhere, not just on our desktop or mobile device. As LCD technology becomes cheaper, walls of rooms, and even walls of buildings, will become display surfaces for information from the Web. Much of the information that we receive today through a specialized application such as a database or a spreadsheet will come directly from the Web. Pervasive and ubiquitous web applications hold much opportunity for innovation and social enrichment. They also pose significant public policy challenges. Nearly all of the information displayed is speech but is being done in public, possibly in a manner accessible to children. Some of this information is bound to be personal, raising privacy questions. Finally, inasmuch as this new ubiquitous face of the Web is public, it will shape the nature of the public spaces we work, shop, do politics, and socialize in.

D. The Web is Not Complete

Progress in the evolution of the Web to date has been quite gratifying to me. But the Web is by no means finished.

The Web, and everything which happens on it, rest on two things: technological protocols, and social conventions. The technological protocols, like HTTP and HTML, determine how computers interact. Social conventions, such as the incentive to make links to valuable resources, or the rules of engagement in a social networking web site, are about how people like to, and are allowed to, interact.

As the Web passes through its first decade of widespread use, we still know surprisingly little about these complex technical and social mechanisms. We have only scratched the surface of what could be realized with deeper scientific investigation into its design, operation and impact

on society. Robust technical design, innovative business decisions, and sound public policy judgment all require that we are aware of the complex interactions between technology and society. We call this awareness Web Science: the science and engineering of this massive system for the common good.[10] In order to galvanize Web Science research and education efforts, MIT and the University of Southampton in the United Kingdom have created the Web Science Research Initiative. In concert with an international Scientific Advisory Council of distinguished computer scientists, social scientists, and legal scholars, WSRI will help create an intellectual foundation, educational atmosphere, and resource base to allow researchers to take the Web seriously as an object of scientific enquiry and engineering innovation.

III. Conclusion

So how do we plan for a better future, better for society?

We ensure that that both technological protocols and social conventions respect basic values. That Web remains a universal platform: independent of any specific hardware device, software platform, language, culture, or disability. That the Web does not become controlled by a single company -- or a single country.

By adherence to these principles we can ensure that Web technology, like the Internet, continues to serve as a foundation for bigger things to come. It is my hope, Chairman Markey, members of the committee, that an understanding of the nature of the Web will guide you in your future work, and that the public at large can count on you to hold these values to the best of your ability. I am grateful for the opportunity to appear before you and am ready to help your efforts in future.

Notes

[1] American Civil Liberties Union v. Reno, 929 F. Supp. 824, 844 (E.D. Pa. 1996) (Dalzell, J.)

[2] Kapor, M. and Weitzner, D. "Social and Industrial Policy for Public Networks: Visions for the Future". Harasim and Walls, eds. Global Networks: Computers and International Communication. Oxford University Press. Oxford. (1994)

[3] Netcraft February 2007 Web Server Survey. http://news.netcraft.com/archives/web_server_survey.html

[4] Weinberger, D., Small Pieces Loosely Joined: A Unified Theory of the Web. Perseus Books. (2002)

[5] Note that due to failure by some browser vendors to comply fully with standards, web site developers sometimes have to go to extra trouble to make it so that their sites actually work properly on all browsers.

[6] Overview and Summary of the W3C Patent Policy, http://www.w3.org/2004/02/05-patentsummary.html. W3C Patent Policy. D. Weitzner, Standards, Patents and the Dynamics of Innovation on the World Wide Web. http://www.w3.org/2004/10/patents-standards-innovation.html

[7] Jobs wrote on the Apple Web site: "Imagine a world where every online store sells DRM-free music encoded in open licensable formats. In such a world, any player can play music purchased from any store, and any store can sell music which is playable on all players. This is clearly the best alternative for consumers, and Apple would embrace it in a heartbeat. If the big four music companies would license Apple their music without the requirement that it be protected with a DRM, we would switch to selling only DRM-free music on our iTunes store. Every iPod ever made will play this DRM-free music," Steve Jobs, Thoughts on Music (February 6, 2007), http://www.apple.com/hotnews/thoughtsonmusic/

[8] Weitzner, Abelson, Berners-Lee, Hanson, Hendler, Kagal, McGuinness, Sussman, Waterman, Transparent Accountable Data Mining: New Strategies for Privacy Protection,; MIT CSAIL Technical Report MIT-CSAIL-TR-2006-007 (27 January 2006).

[9] B. Shneiderman, and J. Preece, PUBLIC HEALTH: 911.gov, Science 315 (5814), 944 (16 February 2007)

[10] "Creating a Science of the Web" Tim Berners-Lee, Wendy Hall, James Hendler, Nigel Shadbolt, Daniel J. Weitzner. Science 313, 11 August 2006. And see the Web Science Research Initiative, http://www.webscience.org/

Avoiding a Tech Train Wreck

By James V. DeLong

From the May/June 2008 Issue

Filed under: Science & Technology, Boardroom

The industry is consumed by disputes similar to those that once roiled American politics. We can avoid past mistakes.

The temperature of business reporting these days is raised by numerous disputes that lie at the intersection of technology and politics. Arguments over Internet network management involving telecommunications companies preoccupy the Federal Communications Commission. The European Union is pressuring several U.S. technology companies, including Microsoft, on antitrust grounds. Firms such as Qualcomm, Broadcom, and Nokia are involved in disputes over the use of patented cell phone technologies. Business writer Nicholas Carr lists entertainment, news, software, and financial companies as all having reason to cast a fearful eye on Google.

These battles may seem novel, a result of the frictions of the technological revolution worked by Moore's Law and its progeny. In fact, while the context is new, the underlying structure of the disputes is similar to the issues that roiled American politics a century and more ago, when the great infrastructures of transportation, utilities, telecommunications, and finance were laid down.

And therein lies a cautionary lesson. The battles of that time resulted in an armistice: the creation of regulatory structures that met immediate crises, but which also failed in many ways. These failures caused significant problems of stasis, rent-seeking, political manipulation, and corruption.

It took the great and only partly successful deregulatory effort that started in the 1970s to unchain the economy from the problems caused by the solutions of an earlier era.

As these old problems appear in new areas, or sometimes in the same old areas, the pressure to revert to the old regulatory solutions grows. Yielding to them would be unfortunate, since the same old solutions would inevitably restart the cycle of regulatory dysfunction.

So, while it is foolish to deny that the problems are real and troublesome, we should think in terms of new solutions. These could be hybrids based on contract, public commitment, and self-regulation.

In some cases the existence and persistence of multiple entities are not a realistic option.These contemporary disputes are united by a common thread. They involve situations in which one company controls a resource upon which other firms must rely to pursue their own businesses. The resource may be a transportation network, such as a railroad, or a telecommunications network, such as access to the Internet, upon which the firm must depend for distribution. Or it may be a platform, such as a computer operating system (Windows or Linux), for which an application writer provides a complement. Or it could be a distribution network, such as eBay, on which numerous small businesses have staked their futures, which has the characteristics of both. Or it could be a Google search engine that becomes the entry portal for millions of Web users.

The essential characteristics of networks and platforms drive these disputes. For example, the network or platform is commonly characterized by high fixed costs and low marginal costs. A railroad or telecom company is a quintessential example; each requires massive up-front investment in rights-of-way, construction, and equipment before a single customer is served. Thereafter, the extra (“marginal”) costs of adding each additional customer are pretty trivial. This structure makes entry by competitors difficult, since they must duplicate the investment before starting business.

The network or platform can also be characterized by a need for interoperability. This creates strong pressures for the industry to coalesce around a single standard. Thus, the proprietor of the resource enjoys considerable market power. A dependent company cannot easily shift its reliance on the network or platform.

Each dependent fears that the network or platform, given its druthers, will raise the price of access to a level that extracts almost all the value from the total enterprise and leaves the dependent with only crumbs. They also know that the network and platform owners have a strong interest in turning the complements into cheap commodities. Doing so increases overall demand for its product and improves its leverage over the dependents.

Think of the position of software developers who write applications for Windows. Microsoft needs them because they add value to its product. But Microsoft also benefits if their returns are low so as to keep the total costs of the system low, thereby increasing its own sales.

In the worst case, the network/platform can wait until dependents commit resources, and then pounce. The classic example here is from history: the farmer who invests in land near a rail line and then sees freight rates zoom. The case of cheap long-haul rates, where there was competition among the railroads, versus high short-haul rates, where there was not, became one of the great raw sores of American politics during the Industrial Revolution.

But it is not enough to look only at the challenges faced by dependent enterprises. Each network/platform company also has a problem. It must spend substantial resources to attain its position. But it is wary of investing to the optimum because it fears that a political response organized by the dependent companies will result in confiscation once its capital is sunk.

Again an example from history is instructive here. In the 1870s, Illinois imposed maximum rates on grain elevators in Chicago. The excuse was that the elevator owners were over-charging and, perhaps, price-fixing, but the record is unclear. It may well have been simply that the customers preferred to pay less, had political clout with the legislature, and could take advantage of the fact that the elevator owners had committed their capital. One cannot pick up a grain elevator and move it to Indiana, so as long as the law allowed the owners to recover at least their operating costs, they were stuck.

The fears of both sides in this controversy are well founded. And, for reasons explored below, neither a “let the free market do it” nor a “let’s regulate” solution provides a totally satisfactory resolution.

Unfortunately, our earlier efforts at achieving regulatory solutions did not end happily.To explore the problem in more depth, and to think about solutions, it is helpful to start with a simple but historically realistic model: a 19th-century railroad running past a factory or a farm. The railroad needs factories and farms en masse to provide revenue, but it does not need any single one of them. So, in the pre-truck days, the railroad could set a rate that captured most of the value of the combination of factory+railroad or farm+railroad. This allowed the producer factory or farm a return that paid only marginal costs, forcing it to forgo much if any return on capital.

Furthermore, the railroad could set different rates for every factory along its lines, extracting maximum value from each. A plant making a commodity product with low margins would pay a low rate; one making specialized goods with high demand would pay a high rate. But each could be charged so much as to be left with only the pittance needed to keep it in business.

Now, shift the hypothetical and add a second railroad. The factory's situation improves dramatically, because it can force the railroads to bid against each other. Indeed, if the factory is big enough, it can reverse the bargaining power, and use the railroad's own high-fixed, low-marginal-cost structure against it, driving rail rates down to the railroad's marginal cost, and forcing it to eat capital costs and retain only a pittance. If the shipper is big enough, such as John D. Rockefeller's Standard Oil, it can force the railroad to charge Standard's competitors a high price and then rebate part of the revenue to Standard, a twofer of disrupting the competitors and reducing the railroad's returns.

Some obvious ways exist for the participants to try to break out of these dilemmas, and they can work pretty well. A factory owner can build his factory where roads, rails, and water all coincide, taking his pick of transport modes, for example. Or he doesn't have to build a factory at all unless the railroad guarantees reasonable rates.

But these strategies work only approximately. Sometimes the location of production facilities is constrained by geography. And in many cases we are talking about serious long-term capital investments. If the factory is located first, how does it persuade the railroad to build? Does it promise never to use a different railroad? Not to use water transportation? Again, these are long-term investments under conditions of great uncertainty, and it is difficult to write the contracts that would be required. A significant problem with the mantra of "do it by contract" is that it requires contract writers with an unlimited legal budget and a level of foresight that would be the envy of a psychic.

Another possible solution to the problem is integration. If the railroad owns the factory, then allocating the total revenues it receives from the factory+railroad becomes a question of internal accounting. This approach, too, has precedents; it is a variation of the land grants given to railroads in the 19th century. These grants enabled them to finance construction by cashing in on the increase in the value of the land triggered by the building of the railroad, and by the ability to sell land to raise crops that the railroad could then haul.

But there are also some problems with integration. If the railroad owns one factory, then what about the competitors of that factory? The railroad has an incentive to charge them more or provide them

with inferior service. Google's chief legal officer, David Drummond, thinks that a modern example is Microsoft's attempted acquisition of Yahoo: "Could a combination of the two take advantage of a PC software monopoly to unfairly limit the ability of consumers to freely access competitors' email, IM, and Web-based services? Policymakers around the world need to ask these questions—and consumers deserve satisfying answers." (Actually, Microsoft does have pretty good answers, called its Interoperability Principles, which are discussed below.)

A great concern of the contemporary tech world is that innovators need access to platforms on reasonable and predictable terms if the full creativity of society is to be tapped.Besides, if society values innovation, a world of large integrated firms is a poor way to promote it. For many reasons, such as excess bureaucracy and fear of cannibalizing existing businesses, large entities tend to be stodgy. The railroads enabled the great catalog mail-order businesses that transformed rural America, but they did not invent them. A great concern of the contemporary tech world is that innovators need access to platforms on reasonable and predictable terms if the full creativity of society is to be tapped.

What about the opposite approach—forbid integration, and impose a legal rule against discrimination on price or other terms? This approach breaks down as a matter of elementary economics, which says that the capacity of a network/platform is not infinite, that it must be rationed and traffic prioritized, and that the price system is the best way to accomplish these ends.

Besides, the dependent companies themselves need price discrimination. If I own a factory, and another shipper can pay only a widow's mite for transportation, then I want the railroad to give that shipper a deal so that it contributes to the railroad's capital costs, even if it pays less than I do. If the railroad does not cut price far enough to capture this firm's business, then all the costs will fall on me. Of course, if the other factory is a competitor of mine, then my calculus changes; I want to share the costs, but not at the price of seeing a competitor get a cost break.

What about the competitive solution of ensuring that there are many railroads? This runs into the problem of setting prices in investment-heavy industries. Excess capacity develops, and then competition drives the price down to the marginal cost of the most efficient. Over time, this puts the others out of business and restores a monopoly situation, at which point prices are raised and the dependents are once again in trouble (until, at some point, some new technology saves them). While economists may look at this great wheel of life with equanimity, actual businesses and governments are less sanguine.

Besides, monopoly, rather than revived competition, may be the logical outcome. In some cases the existence and persistence of multiple entities are not a realistic option.

Historically, regulation was based on the concept of “natural monopoly,” which is an economic concept based on a high-fixed, low-marginal-cost structure that makes it inefficient to duplicate basic infrastructure. In a recent variation on theories of natural monopoly based on fixed costs, economists now discuss winner-take-all markets, which coalesce around a single technology, such as Windows on the desktop computer or Blu-Ray as the hi-def DVD disk because of the desire of users to interoperate. Economic theory says not to worry because new technologies will disrupt such empires in the long run; but dependent companies are not, unsurprisingly, interested in waiting for the long run, so there will be political pressure for countervailing action right now.

Say a good word about ‘common carrier law’ in front of a telecom or software executive, and they immediately reach for some holy water and garlic.The same basic scenarios between networks/platforms and dependents play out repeatedly, not just with shippers of physical goods but with developers of software applications for Windows, with Internet companies dependent on telecom firms, with search engines, and with cell phone manufacturers using patent-protected chips. Indeed, in the tech area, the whole “openness” philosophical movement—whether in the context of software or net neutrality or anti-copyright—traces at least part of its lineage to this 19th-century problem of shippers and railroads, and to the desire to avoid dependency. Old railroad battles over long-haul versus short-haul rates, oil tariff rebates, price discrimination, lines built solely for blackmail, and other issues of those thrilling days of yesteryear are re-enacted in the guise of Google battling Verizon and Comcast over net neutrality, or content creators battling Google over the division of search revenues, or the EU forcing Microsoft to open up server technology.

Inevitably, pressure develops for the political system to address the problems. Unfortunately, our earlier efforts at achieving regulatory solutions did not end happily. Many railroads went bankrupt in the late 19th century, and rural populists and investment bankers eventually cooperated to create a comprehensive regulatory regime under the Interstate Commerce Commission. Policy then oscillated between encouraging monopoly and exploiting the railroads for political gain. There were subsequent periods of flowering, but in the long term innovation withered and the railroads declined.

Telecommunications were also regulated and the history of the Federal Communications Commission became a sad saga of suppression of innovation and wasted spectrum. George Mason University law professor Thomas Hazlett, former chief economist at the FCC, has pointed out the technologies that

were substantially retarded by the agency's spectrum policies. These included FM broadcasting, which was delayed for 26 years. Many analysts have noted that the FCC's implementation of the 1996 Telecom Act attempted to turn telecom facilities into a giant commons, which depressed investment in new facilities in favor of investment in lawyers and economists who could game the rules.

Nor has the world beyond America's borders learned. If the EU continues to attack tech companies that achieve market power through innovation because it judges they are not being sufficiently sharing with competitors, the companies are not likely to prosper.

So how do we avoid repeating these experiences? A common complaint from academics, potential competitors, and complementary businesses is that regulatory regimes allow cartelization and monopoly by the networks/platforms. True enough, but, as the history of the railroads and telecommunications shows, the pressure to protect complementary dependent industries can also lead to reduced investment and innovation in the network/platform companies, to the detriment of all.

There is no rule of the world that says that all problems have neat solutions, and this is a set for which no theoretical answers emerge. But a number of practical approaches look useful.

First, it is important to understand what is not relevant. Antitrust law is fairly useless. It focuses on competition among the network/ platform companies themselves, and deals with their relations with dependent companies only in the limited context of "tying," when a dominant company sells a complementary product and forces the customer to buy it. Because antitrust doctrine focuses on harm to consumers—not to intermediate dependent producers—it becomes very difficult to see why tying should be of concern at all.

Say a good word about 'common carrier law' in front of a telecom or software executive, and they immediately reach for some holy water and garlic.To better understand this, consider that consumers do not care if they buy their copy paper from the same company that sells the machines. And if diners in New York have access to meat from Argentina and Poland as well as from Nebraska, antitrust should not, in theory, care if a railroad is extracting most of the value of the meat shipped from Nebraska. There is no consumer harm, because competition from overseas will keep prices down. But since the Nebraska farmers, other paper producers, and the U.S. political system all care a great deal about these situations, the antitrust system plays legal and economic Twister to argue that these are actually antitrust problems.

The EU has a doctrine called “abuse of a dominant position” under Section 82 of the EU Treaty which might, in theory, address the relationship between platform companies and producers of complements. Its language is broad enough to cover what is or is not permissible for dominant companies in dealing with providers of complements as well as with competitors.

However, the EU, under the tutelage of the U.S. Department of Justice, uses Section 82 to replicate—and lately, in the various contentious tech industry cases, to extend—American thinking about actions by dominant companies designed to protect their existing market power, as opposed to how they use this power. It does not look at the actions of such companies in dealing with providers of complements, unless by some intellectual gymnastics a provider can be characterized as a potential competitor of the dominant firm.

If antitrust thinking is of little help, it might seem logical to turn to common carrier regulation, the elaborate regimes developed by the Interstate Commerce Commission, the FCC, and many other federal and state regulatory bodies to control railroads, telecommunications, utilities, and other network industries. This, too, is dangerous, because over the decades these bodies of rules became detailed and intrusive, and constantly distorted by political influence and regulatory whim. Say a good word about “common carrier law” in front of a telecom or software executive, and they immediately reach for some holy water and garlic.

On the other hand, the concept of “common carrier” actually developed earlier, under common law, long before the Industrial Revolution. And some of the fundamental and limited common carrier principles invented in this pre-regulatory era are quite interesting—they consisted of openness, availability, advance notice of terms, and inhibitions (but not prohibitions) on price discrimination.

Reading 19th-century law cases is not usually considered good preparation for analyzing issues of the Internet Age, but in fact it is a useful mental exercise. If one is forced to think with the Supreme Court in Munn v. Illinois (1877), which judged state power to impose price caps and other regulation on grain elevators, it is clear that many of today’s issues were raised over a century ago. For a preview of the current debate over net neutrality, examine the Court’s 1892 discussion of whether “party rates” for groups such as theatrical troupes violated a prohibition on price discrimination. The answer was “no,” with the reservation that the decision might change if the railroad differentiated among competing producers.

The ‘openness’ philosophical movement traces part of its lineage to the 19th-century problem of shippers and railroads, and to the desire to avoid dependency.These questions are more complex than laissez-faire advocates would have you believe, and there is some wisdom in those old cases and their rules. A dependent company’s biggest problems are ambush, in which its resources are committed and then the network/platform company raises the rates; uncertainty, as when an innovator is in the impossible position of asking a large carrier, “If we invented an X, what would you charge to attach it?”; negotiation costs of trying to get an answer from a large network company’s bureaucracy; and price discrimination, in which a competitor gets lower costs. Advance notice, transparency, and guarantees that no other customer will get a better deal from the network/platform company (so-called “most-favored-nation clauses”) are powerful tools.

It is equally important to recognize the limitations of common carrier–like approaches. “No integration” may seem attractive, but it deprives consumers of the advantages of efficient packages. In general, integration often has some powerful benefits, especially in allowing consumers to buy packages of integrated products and services. No one wants to buy an automobile one part at a time, and the same is true increasingly of computer programs. We want plug and play. Should Apple be required to unbundle its hardware/operating system/applications package? It is precisely the smooth integration that makes the Mac appealing to so many.

Also, over time, a no-integration rule that a network or platform must be nothing but a stand-alone dumb pipe triggers a cascade of problems that must lead to regulation, price control, and technological stasis. Imagine a telecom future in which every household has ready access to cable, fiber-optic lines, broadband over power lines, and satellite. Assume that the basic infrastructure has been put in place and the only gap is connection to the premises. Now, suppose that each of these is a dumb pipe, forced to carry all content. The obvious result would be a war that drove prices down to marginal cost, which would starve the carriers of the ability to service their capital or maintain their plant, and send them into bankruptcy. Perhaps the only way to maintain competition among carriers in the long term, and thus avoid the evils of regulatory structures, is to allow integration.

It is not easy to find an optimum path between the limited doctrines of historic common law common carrier doctrines and full-blown regulatory regimes. However, network/platform companies can influence their own fate by contractual obligations or public announcements, especially if these constitute binding obligations, either practically or legally.

It took the great and only partly successful deregulatory effort that started in the 1970s to unchain the economy from the problems caused by the solutions of an earlier era.In 2005, the FCC did the telecom

companies a favor by issuing four principles supporting openness on the Internet. The principles said consumers are entitled to access the lawful Internet content of their choice, to run applications and use services of their choice, to connect their choice of legal devices that do not harm the network, and to have competition among network providers, application and service providers, and content providers. In sum, Internet service providers (ISPs) will not be allowed to exploit their position to exclude others' content or devices.

The FCC had no particular authority to adopt these principles, and their legal status remains uncertain. They are important though because the fear of dominance by ISPs is great, and telecom companies might well forestall unpleasant regulation by promising not to assert their maximum position. In fact, no telecom company dares contradict the four principles, and a large part of their response to demands for net neutrality rests on the proposition that the four principles make any such regulations unnecessary.

On this issue, the telecoms' strategic willingness to give up part of what might be their formal rights under existing law is very wise. The action recognizes that something is needed that lies between the old regulatory models and total telecom discretion, and that there is no reason for the telecoms to wait for it to be imposed legislatively. They have little to lose, because any effort to claim a right of total control over Internet traffic carries heavy risk of a reaction that deprives the carriers of all discretion.

Another example of a company's strategic wisdom in ceding formal legal rights is Microsoft's 2006 adoption of its Windows Principles, designed to assure applications developers of their ability to capture the financial benefits of inventiveness. Microsoft acted under the lash of the Justice Department, admittedly, but the result is important. For one thing, it makes clear that the real axis of the antitrust case against Microsoft revolved around its relations with complements, not with the bogus issue of creating competitors to the Windows operating system. Microsoft extended this philosophy recently with its Interoperability Principles, which, while they must have been long in the works, might have the happy effect of blunting Google's arguments about the Yahoo deal.

The development of the open-source Linux operating system serves the same function as Microsoft's Principles—that of assuring developers of complements that they cannot be exploited after they have committed their resources. Tech expert George Gilder once attributed the rise of the open-source software movement to the fact that in the late 1990s Microsoft's share of the total value of Windows+application systems sharply increased, thus irking the developers, who became receptive to the idea of a new platform.

Other examples of private efforts to solve the network/platform problem can be found, especially in the tech industry, where the concept of granting licenses based on “reasonable and nondiscriminatory” (RAND) terms has taken hold. In the search for solutions, it is better to look at these new contractual models than at the old regulatory structures.

The telecoms’ willingness to give up part of what might be their formal rights under existing law is very wise.But a sense of history is also important, because, as philosopher George Santayana said, “Those who cannot learn from history are doomed to repeat it.” Historian Crane Brinton added that, unfortunately, those who can learn from history are doomed to repeat it with them. The great regulatory regimes of the 19th and 20th centuries were established because of real and difficult problems centered around the nature of high-fixed, low-marginal-cost network/platform providers and their relationships with the producers of complements that depend on these providers.

These regulatory regimes failed in many ways, and the great deregulation movement was immensely important. But the people who put the regimes in place were not stupid. We should look at their regulatory solutions as a natural experiment that did not quite work, and develop a new synthesis that incorporates the merits of both laissez-faire thesis and regulatory antithesis.

James V. DeLong is vice president and senior analyst of the Convergence Law Institute, LLC, and special counsel in the Washington, D.C. office of Kamlet Shepherd & Reichert, LLP.

One version of this article was published by Reason Magazine, and a slightly different version by the University of Chicago Law School Record--their alumni magazine.)

The Swedes Get It Right

When the Swedish Academy awarded the Nobel Prize in Economics to Ronald Coase this year, it was a surprise for two different groups of people. The larger group consisted of people who had either never heard of Coase, or heard of him only as the author of something called the "Coase Theorem," generally presented as a theoretical curiousity of no practical importance. The second and much smaller group consisted of people who were familiar with the importance of Coase's work--and assumed that the Swedish Academy was not.

Some people get the Nobel prize for complicated and technical work that is difficult for an outsider to understand. Coase is at the other extreme. His contribution to economics has largely consisted of thinking through certain questions more carefully and correctly than anyone else, and in the process demonstrating that answers accepted by virtually the entire profession were false. One side effect of his work was a new field of economics: economic analysis of law, the attempt to use economic theory to understand legal systems. While there would probably be something called economic analysis of law if Coase had not existed, it would be a very different field.

One of Coase's important contributions to economics was to rewrite the theory of externalities--the analysis of situations, such as pollution, where one person's actions impose costs (or benefits) on another. His ideas are sufficiently simple to be understood by a layman, as I will try to demonstrate in the next few pages, and sufficiently deep so that they have not yet been entirely absorbed by the profession; to a considerable extent what is still taught in the textbooks is the theory as it existed before Coase.

To understand the significance of Coase's contribution to the theory of externalities, it is useful to start with the theory as it existed before Coase published "The Problem of Social Cost," the essay that first introduced the Coase Theorem to economics. The basic argument went as follows:

In an ideal economic system, goods worth more than they cost to produce get produced, goods worth less than they cost to produce do not; this is part of what economists mean by economic efficiency. In a perfectly competitive private property system, producers pay the value of the inputs they use when they buy them from their owners (wages to workers in exchange for their labor, rent to land owners for the use of their land, etc.) and receive the value of what they produce when they sell it. If a good is worth more than it costs to produce, the producer receives more than he pays

and makes a profit; if the good is worth less than it costs to produce he takes a loss. So goods that should be produced are and goods that should not be produced are not.

This only works if producers must pay all of the costs associated with production. Suppose that is not the case. Suppose, for example, that a steel producer, in addition to using iron ore, coal, etc., also "uses" clean air. In the process of producing a ton of steel he puts ten pounds of sulfur dioxide into the air, imposing (say) $100 worth of bad smells, sore throats, and corrosion on people down wind. Since he does not pay for that cost, he does not include it in his profit and loss calculations. As long as the price he sells his steel for at least covers his costs it is worth making steel. The result is inefficient: Some goods may be produced even though their cost, including the resulting pollution, is greater than their value.

It is inefficient in another respect as well. The steel producer may be able to reduce the amount of pollution by various control devices--air filters, low sulfur coal, high smokestacks--at a cost. Calculated in terms of the net effect on everyone concerned, it is worth eliminating pollution as long as the cost is less than the pollution damage prevented--in our example, as long as it costs less than $10 to prevent a pound of sulfur dioxide emission. But the steel producer, in figuring out how to maximize his profit, includes in his calculations only the costs he must pay. So long as he does not bear the cost of the pollution, he has no incentive to prevent it. So the fact that air pollution is an external cost results in both an ineffiently high level of steel production (it may be produced even when it is not worth producing) and an inefficiently low level of pollution control.

There are two obvious solutions. One is direct regulation--the government tells the steel company how much it is allowed to pollute. The other is emission fees--referred to by economists as Pigouvian taxes (named after A. C. Pigou, the economist whose ideas I am describing).

Under a system of Pigouvian taxes, the government charges the steel company for the damage done by its pollution--$10 per pound in this example. By doing so it converts the external cost into an internal cost--internalizes the externality. In deciding how much steel to produce and what price to sell it at, the company will now include the cost of its pollution--paid as an emission fee--along with other costs. In deciding how much pollution control equipment to buy, the company balances the cost of control against its benefits, and buys the optimal amount. So a system of emission fees can produce both an efficient amount of steel and an efficient amount of pollution control.

In order to achieve that result, the government imposing the fees must be able to measure the cost imposed by pollution. But, unlike direct regulation, the use of emission fees does not require the government to measure the cost of preventing

pollution--whether by installing air filters or by producing less steel. That will be done by the steel company, acting in its own interest.

I have just described the theory of externalities as it existed before Coase. Its conclusion is that, as long as externalities exist and are not internalized via Pigouvian taxes, the result is inefficient. The inefficiency is eliminated by charging the polluter an emission fee equal to the damage done by his pollution. In some real world cases it may be difficult to measure the amount of the damage, but, provided that that problem can be solved, using Pigouvian taxes to internalize externalities produces the efficient outcome.

That analysis was accepted by virtually the entire economics profession prior to Coase's work in the field. It is wrong--not in one way but in three. The existence of externalities does not necessarily lead to an inefficient result. Pigouvian taxes, even if they can be correctly calculated, do not in general lead to the efficient result. Third, and most important, the problem is not really externalities at all--it is transaction costs.

I like to present Coase's argument in three steps: *Nothing works, Everything works, It all depends.*

Nothing Works

The first step is to realize that an external cost is not simply a cost produced by the pollutor and born by the victim. In almost all cases, the cost is a result of decisions by both parties. I would not be coughing if your steel mill were not pouring out sulfur dioxide. But your steel mill would do no damage to me if I did not happen to live down wind from it. It is the joint decision--yours to pollute and mine to live where you are polluting--that produces the cost.

Suppose that, in a particular case, the pollution does $100,000 a year worth of damage and can be eliminated at a cost of only $80,000 a year (from here on, all costs are per year). Further assume that the cost of shifting all of the land down wind to a new use unaffected by the pollution--growing timber instead of renting out summer resorts, say-- is only $50,000. If we impose an emission fee of a hundred thousand dollars a year, the steel mill stops polluting and the damage is eliminated--at a cost of $80,000. If we impose no emission fee the mill keeps polluting, the owners of the land stop advertising for tenants and plant trees instead, and the problem is again solved--at a cost of $50,000. In this case the result without Pigouvian taxes is efficient--the problem is eliminated at the lowest possible cost--and the result with Pigouvian taxes in inefficient.

Moving the victims may not be a very plausible solution in the case of air pollution; it seems fairly certain that even the most draconian limitations on emissions in southern California would be less expensive than evacuating that end of the state. But the problem of externalities applies to to a wide range of different situations, in many of which it is far from obvious which party can avoid the problem at lower cost, and in some of which it is not even obvious which one we should call the victim.

Consider the question of airport noise. One solution is to reduce the noise. Another is to soundproof the houses. A third is to use the land near airports for noisy factories instead of housing. There is no particular reason to think that one of those solutions is always best. Nor is it entirely clear whether the "victim" is the landowner who finds it difficult to sleep in his new house with jets going by overhead or the airline forced by a court or a regulatory agency to adopt expensive sound control measures in order to protect the sleep of people who chose to build their new houses in what used to be wheat fields--directly under the airport's flight path.

Consider a simpler case, where the nominal offender is clearly not the lowest cost avoider. The owner of one of two adjoining tracts of land has a factory, which he has been running for twenty years with no complaints from his neighbors. The purchaser of the other tract builds a recording studio on the side of his property immediately adjacent to the factory. The factory, while not especially noisy, is too noisy for something located two feet from the wall of a recording studio. So the owner of the studio demands that the factory shut down, or else pay damages equal to the full value of the studio. There are indeed "external costs" associated with operating a factory next to a recording studio--but the efficient solution is building the studio at the other end of the lot, not building the studio next to the factory and then closing down the factory.

So Coase's first point is that externalities are a joint product of "pollutor" and "victim," and that a legal rule that arbitrarily assigns blame to one of the parties only gives the right result if that party happens to be the one who can avoid the problem at the lower cost. Pigou's solution is correct only if the agency making the rules already knows which party is the lower cost avoider. In the more general case, nothing works--whichever party the blame is assigned to, by government regulators or by the courts, the result may be inefficient if the other party could prevent the problem at a lower cost.

One of the arguments commonly offered in favor of using Pigouvian taxes instead of direct regulation is that the regulator does not have to know the cost of pollution control in order to produce the efficient outcome--he just sets the tax equal to damage done, and lets the pollutor decide how much pollution to buy at that price. But one of the implications of Coase's argument is that the regulator can only guarantee the

efficient outcome if he knows enough about the cost of control to decide which party should be considered the pollutor (and taxed) and which should be considered the victim.

Everything Works

The second step in Coase's argument is to observe that, as long as the parties involved can readily make and enforce contracts in their mutual interest, neither direct regulation nor Pigouvian taxes are necessary in order to get the efficient outcome. All you need is a clear definition of who has a right to do what and the market will take care of the problem.

To see how that works, let us go back to the case of the steel mill and the resorts. Suppose first that the mill has a legal right to pollute. In that case, as I originally set up the problem, the efficient result occurs immediately. The lowest cost avoiders are the owners of the land downwind; they shift from operating resorts to growing timber.

What if, instead, the legal rule is that the people downwind have a right not to have their air polluted? The result will be exactly the same. The mill could eliminate the pollution at a cost of $80,000 a year. But it is cheaper to pay the landowners some amount, say $60,000 a year, for permission to pollute. The landowners will be better off, since that is more than the cost to them of changing the use of the land, and the steel mill will be better off, since it is less than the cost of eliminating the pollution. So it will pay both parties to make some such agreement.

Now suppose we change the numbers in the example, to make pollution control the more efficient option--say lower its cost to $20,000. In that case, whether or not the mill has the right to pollute, it will find that it is better off not polluting. If it has the right to pollute, the landowners will pay more than the $20,000 cost of pollution control in exchange for a guarantee that it will not exercise its right. If it does not have the right to pollute, the most the steel mill will be willing to offer the landowners for permission to pollute is $20,000, and the landowners will turn down that offer.

The generalization of this example is straightforward:

If transaction costs are zero--if, in other words, any agreement that is in the mutual benefit of the parties concerned gets made--then any initial definition of property rights leads to an efficient outcome.

It is this result that is sometims referred to as the "Coase Theorem." It leads immediately to the final stage of the argument.

It All Depends (On Transaction Costs)

Why is it, if Coase is correct, that we still have pollution in Los Angeles? One possible answer is that the pollution is efficient--that the damage it does is less than the cost of preventing it. A more plausible answer is that much of the pollution is inefficient, but that the transactions necessary to eliminate it are prevented by prohibitively high transaction costs.

Let us return to the steel mill. Suppose the mill has the right to pollute, but that doing so is inefficient--pollution control is cheaper than either putting up with the pollution or changing the use of the land down wind. Further suppose that there are a hundred landowners down wind.

With only one landowner, there would be no problem--he would offer to pay the mill for the cost of the pollution control equipment, plus a little extra to sweeten the deal. But a hundred landowners face what economists call a public good problem. If ninety of them put up the money and ten do not, the ten get a free ride--no pollution and no cost for pollution control. Each landowner has an incentive to refuse to pay, figuring that his payment is unlikely to make the difference between success and failure in the attempt to bribe the steel mill to eliminate its pollution. If the attempt is going to fail even with him, then it makes no difference whether or not he contributes. If it is going to succeed even without him, then refusing to contribute gives him a free ride. Only if his contribution makes the difference does he gain by agreeing to contribute.

There are a variety of ways in which such problems may sometimes be solved, but none that can always be expected to work. The problem becomes harder the larger the number of people involved. With many millions of people living in southern California, it is hard to imagine any plausible way in which they could voluntarily raise the money to pay all pollutors to reduce their pollution.

This is one example of the sort of problem referred to under the general label of "transaction costs." Another would occur if we reversed the assumptions, making pollution (and timber) the efficient outcome but giving the landowners the right to be pollution free. If there were one landowner the steel mill could buy from him the right to pollute. With a hundred, the mill must buy permission from all of them. Any one has an incentive to be a holdout--to refuse his permission in the hope of getting paid off with a large fraction of the money the mill will save from not having to control its pollution. If too many landowners try that approach the negotiations will break down, and the parties will never get to the efficient outcome.

Seen from this perspective, one way of stating Coase's insight is that the problem is not really due to externalities at all, but to transaction costs. If there were externalities

but no transaction costs there would be no problem, since the parties would always bargain to the efficient solution. When we observe externality problems (or other forms of market failure) in the real world, we should ask not merely where the problem comes from, but what the transaction costs are that prevent it from being bargained out of existence.

Coase, Meade, and Bees

Ever since Coase published "The Problem of Social Cost," economists unconvinced by his analysis have argued that the Coase Theorem is merely a theoretical curiousity, of little or no practical importance in a world where transaction costs are rarely zero. One famous example was in an article by James Meade (who later received a Nobel prize for his work on the economics of international trade).

Meade offered, as an example of the sort of externality problem for which Coase's approach offered no practical solution, the externalities associated with honey bees. Bees graze on the flowers of various crops, so a farmer who grows crops that produce nectar benefits the beekeepers in the area. The farmer receives none of the benefit himself, so he has an inefficiently low incentive to grow such crops. Since bees cannot be convinced to respect property rights or keep contracts, there is, Meade argued, no practical way to apply Coase's approach. We must either subsidize farmers who grow nectar rich crops (a negative Pigouvian tax) or accept inefficiency in the joint production of crops and honey.

It turned out that Meade was wrong. In two later articles, supporters of Coase demonstrated that contracts between beekeepers and farmers had been common practice in the industry since early in this century. When the crops were producing nectar and did not need pollenization, beekeepers paid farmers for permission to put their hives in the farmers' fields. When the crops were producing little nectar but needed pollenization (which increases yields), farmers paid beekeepers. Bees may not respect property rights but they are, like people, lazy, and prefer to forage as close to the hive as possible.

The fact that a Coasian approach solves that particular externality problem does not imply that it will solve all such problems. But the observation that an economist as distinguished as Meade assumed Coase's approach was of no practical significance in a context where it was actually standard practice suggests that the range of problems to which the Coasian solution is relevant may be much greater than many would at first guess.

Coase, Property, and the Economic Analysis of Law

"The Problem of Social Cost" provides more than merely a revolutionary rethinking of the question of externalities. It also suggests a new and interesting approach to the problem of defining property rights.

A court, in settling disputes involving property, or a legislature in writing a law code to be applied to such disputes, must decide just which of the rights associated with land are included in the bundle we call "ownership." Does the owner have the right to prohibit airplanes from crossing his land a mile up? How about a hundred feet? How about people extracting oil from a mile under the land? What rights does he have against neighbors whose use of their land interferes with his use of his? If he builds his recording studio next to his neighbor's factory, who is at fault? If he has a right to silence in his recording studio, does that mean that he can forbid the factory from operating, or only that he can sue to be reimbursed for his losses? It seems simple to say that we should have private property in land, but ownership of land is not a simple thing.

The Coasian answer to this set of problems is that the law should define property in such a way as to minimize the costs associated with the sorts of incompatible uses we have been discussing--factories and recording studios, or steel mills and resorts. The first step in doing so is to try to define rights in such a way that, if right A is of most value to someone who also holds right B, they come in the same bundle. The right to decide what happens two feet above a piece of land is of most value to the person who also holds the right to use the land itself, so it is sensible to include both of them in the bundle of rights we call "ownership of land." On the other hand, the right to decide who flies a mile above a piece of land is of no special value to the owner of the land, hence there is no good reason to include it in that bundle.

If, when general legal rules were being established, we somehow knew, for all cases, what rights belonged together, the argument of the previous paragraph would be sufficient to tell us how property rights ought to be defined. But that is very unlikely to be the case. In many situations a right, such as the right not to have noises of more than X decibels made over a particular piece of property, may be of substantial value to two or more parties--the owner of the property and the owner of the adjacent factory in my earlier example, for instance. There is no general legal rule that will always assign it to the right one.

In this case, the argument underlying the Coase Theorem comes into play. If we assign the right initially to the wrong person, the right person, the one to whom it is of most value, can still buy it. So one of the considerations in the initial definition of property rights is doing it in such a way as to minimize the transaction costs associated with fixing, via private contracts, any initially inefficient definition.

An example may make this clearer. Suppose that, in the pollution case discussed earlier, damages from pollution are easy to measure and the number of people downwind is large. In that case, the efficient rule is probably to give downwind landowners a right to collect damages from the pollutor, but not a right to forbid him from polluting. Giving the right to the landowners avoids the public good problem that we would face if the landowners (in the case where pollution is inefficient) had to raise the money to pay the steel mill not to pollute. Giving them a right to damages rather than giving each landowner the right to an injunction forbidding the steel mill from polluting avoids the holdout problem that the mill would face (in the case where pollution is efficient) in buying permission from all of the landowners.

A full explanation of how Coase's argument can be applied to figuring out what the law ought to be (more precisely, what legal rules lead to the best outcome from the standpoint of economic efficiency) would require a much longer article--perhaps a book. I hope I have said enough to make clear the basic idea, and enough to show the unique and extraordinary nature of one of Ronald Coase's principal contributions to economics. He started with a simple insight, based in part on having read cases in the common law of nuisance--the branch of law that deals with problems such as noisy factories next door to recording studios. He ended by demonstrating that what everyone else in the profession thought was the correct analysis of the problem of externalities was wrong, and, in the process, opening up a whole new approach to the use of economics to analyze law.

There is at least one more thing worth saying about "The Problem of Social Cost." Economists, then and (to some degree) now, tend to jump from the observation that the market produces an inefficient result in some situation to the conclusion that the government ought to intervene to fix the problem. Part of what Coase showed was that, for some problems, there is no legal rule, no form of regulation, that will generate a fully efficient solution. He thus anticipated public choice economists, such as James Buchanan (another Nobel winner), in arguing that the real choice was not between an inefficient market and an efficient government solution but rather among a variety of inefficient alternatives, private and governmental. In Coase's words: "All solutions have costs and there is no reason to suppose that government regulation is called for simply because the problem is not well handled by the market or the firm."

References

Cheung, Steven N. S., "The Fable of the Bees: An Economic Investigation," *Journal of Law and Economics* XVI (1973), 11-33.

Coase, R.H. , The Problem of Social Cost, *Journal of Law and Economics* 3, 1-44 (1960).

Friedman, D., *The Machinery of Freedom*, 2nd Edn., Open Court: La Salle, 1989, Chapters 41-43.

Johnson, David B., "Meade, Bees, and Externalities," *Journal of Law and Economics* XVI (1973), 35-52.

Meade, J. E., "External Economies and Diseconomies in a Competitive Situation," 52 *Economic Journal* 54 (1952).

Pigou, A.C., *Wealth and Welfare* (1912) and *The Economics of Welfare* (1920).

Posner, R., *Economic Analysis of Law*, 3rd Edn., Little Brown & Co. Boston, 1986.

1. Institutions and the Performance of Economies Over Time

DOUGLASS C. NORTH

1. Introduction

The discipline of economics is made up of a static body of theory that explores the efficiency of resource allocation at an instant of time and under the restrictive assumptions of frictionless markets. Recent research has explored the nature of the frictions by incorporating institutions, transaction costs, and political economy into economic analysis thereby providing the theory with a bridge to the real world of real economies. But the first constraint of static analysis severely hinders our ability to analyze and improve the performance of economies in a world of continuous change. And, in fact, the employment of static theory as a source of policy recommendation in a setting of dynamic change is a prescription for the policies producing unanticipated and undesirable results. In this essay I intend to provide an approach to the study of the process of economic change. There is still much that we do not understand about the process but this essay provides an analytical framework that does, I believe, highlight the problems that must be confronted in order to understand and improve economic performance. I first describe the intentional nature of human interaction in a world of pervasive uncertainty (2) before going on to describe the process of economic change (3). I conclude with drawing some implications from this approach to the process of change which highlight the lacunae in our understanding of this process (4).

2. Interactions in a World of Uncertainty

[1]In contrast to standard theory that draws its inspiration from physics, modeling the process of change must derive its inspiration from evolutionary biology but in contrast to Darwinian theory in which the selection mechanisms are not informed by beliefs about the eventual consequences, human evolution is guided by the perceptions of the players in which choices—decisions—are made in the light of these perceptions with the intent of producing outcomes downstream that will reduce the uncertainty of the organizations—political, economic, and social—in

[1] This section is drawn from my essay "Five Propositions about Institutional Change", in Knight, J. and Sened, I., *Exploring Social Institutions*, Michigan: The University Press, 1995.

C. Ménard and M. M. Shirley (eds.), Handbook of New Institutional Economics, 21–30.
© 2005 *Springer. Printed in the Netherlands.*

pursuit of their goals. Institutional change, therefore, is a deliberate process shaped by the perceptions of the actors about the consequences of their actions. The immediate vehicle by which the actors attempt to shape their environment is by altering the institutional framework in order to improve their (and their organizations') competitive position. Let me state five propositions that describe this process:

1. The continuous interaction between institutions and organizations in the economic setting of scarcity and hence competition is the key to institutional change.
2. Competition forces organizations continually to invest in new skills and knowledge to survive. The kind of skills and knowledge individuals and their organizations acquire will shape evolving perceptions about opportunities and hence choices that will incrementally alter institutions.
3. The institutional framework provides the incentive structure that dictates the kinds of skills and knowledge perceived to have the maximum payoff.
4. Perceptions are derived from the mental constructs of the players.
5. The economies of scope, complementarities, and network externalities of an institutional matrix make institutional change overwhelmingly incremental and path dependent.

Let Me Expand on These Propositions

1. Institutions are the rules of the game—both formal rules, informal norms and their enforcement characteristics. Together they define the way the game is played. Organizations are the players. They are made up of groups of individuals held together by some common objectives. Economic organizations are firms, trade unions, cooperatives, etc.; political organizations are political parties, legislatures, regulatory bodies; educational organizations are universities, schools, vocational training centers. The immediate objective of organizations may be profit maximizing (for firms) or improving reelection prospects (for political parties); but the ultimate objective is survival because all organizations live in a world of scarcity and hence competition.
2. New or altered opportunities may be perceived to be a result of exogenous changes in the external environment which alter relative prices to organizations, or a consequence of endogenous competition among the organizations of the polity and the economy. In either case the ubiquity of competition in the overall economic setting of scarcity induces entrepreneurs and the members of their organizations to invest in skills and knowledge. Whether through learning by doing on the job or the acquisition of formal knowledge, improving the efficiency of the organization relative to that of rivals is the key to survival.

   While idle curiosity surely is an innate source of acquiring knowledge among human beings, the rate of accumulating knowledge is clearly tied to the pay-offs. Secure monopolies, be they organizations in the polity or in the economy, simply do not have to improve to survive. But firms, political parties, or even institutions of higher learning faced with rival organizations must

strive to improve their efficiency. When competition is muted (for whatever reasons) organizations will have less incentive to invest in new knowledge and in consequence will not induce rapid institutional change. Stable institutional structures will be the result. Vigorous organizational competition will accelerate the process of institutional change.

3. There is no implication in proposition 2 of evolutionary progress or economic growth—only of change. The institutional matrix defines the opportunity set, be it one that makes income redistribution the highest pay-off in an economy or one that provides the highest payoffs to productive activity. While every economy provides a mixed set of incentives for both types of activity, the relative weights (as between redistributive and productive incentives) are crucial factors in the performance of economies. The organizations that come into existence will reflect the payoff structure. More than that, the direction of their investment in skills and knowledge will equally reflect the underlying incentive structure. If the highest rate of return in an economy comes from piracy we can expect that the organizations will invest in skills and knowledge that will make them better pirates. Similarly if there are high returns to productive activities we will expect organizations to devote resources to investing in skill and knowledge that will increase productivity (the new growth economics literature can become relevant at this point).

   The immediate investment of economic organizations in vocational and on the job training obviously will depend on the perceived benefits; but an even more fundamental influence on the future of the economy is the extent to which societies will invest in formal education, schooling, the dissemination of knowledge, and both applied and pure research which will mirror the perceptions of the entrepreneurs of political and economic organizations.
4. The key to the choices that individuals make is their perceptions about the payoffs, which are a function of the way the mind interprets the information it receives. The mental constructs individuals form to explain and interpret the world around them are partly a result of the genetic evolution of the mind, partly of their cultural heritage, partly a result of the local everyday problems, they confront and must solve, and, partly a result of non-local learning. The mix among these sources in interpreting one's environment obviously varies as between for example a Papuan tribesman on the one hand and an economist in the United States on the other (although there is no implication that the latter's perceptions are independent of his or her cultural heritage).

   The implication of the foregoing paragraph is that individuals from different backgrounds will interpret the same evidence differently; they may, in consequence, make different choices. If the information feedback of the consequences of choices were complete then individuals with the same utility function would gradually correct their perceptions and over time converge to a common equilibrium; but as Frank Hahn has succinctly put it, "There is a continuum of theories that agents can hold and act upon without ever encountering events which lead them to change their theories." (Hahn, 1987, p. 324). The result is that multiple equilibria are possible due to different choices by agents with identical tastes.

5. The viability, profitability, and indeed survival of the organizations of a society typically depend on the existing institutional matrix. That institutional structure has brought them into existence; and their complex web of interdependent contracts and other relationships has been constructed on it. Two implications follow. Institutional change is typically incremental and is path dependent.

   This institutional change is occurring in a world of pervasive uncertainty or ambiguity which by definition is one in which one cannot derive a probability distribution of possible outcomes-such as is possible with decision making in the face of risk (in the Knightian definitions). This uncertainty persists because the "human landscape" in which humans are interacting is continually undergoing change—change induced in part by non-human action (for example changes in climate, natural disasters) but primarily by the human actors themselves.

Humans attempt to reduce that uncertainty (or convert it into risk) by learning. The cumulative learning of a society embodied in language, beliefs, myths, ways of doing things—in short the culture of a society—not only determines societal performance at a moment of time but through the way in which it constrains the choices of the players contributes to the nature of the process through time. Humans scaffold both the mental models they possess—belief systems—and the external environment—institutions. The focus of our attention, therefore, must be on human learning, on what is learned and how it is shared among the members of a society, on the incremental process by which the beliefs and preferences change through time, and on the way in which they shape the performance of economies through time.

We can describe that performance by innumerable statistics on its demographic, economic, technological, and institutional features; but what we really need to know is what is the interplay between all these features that makes it work. The foundations of the interplay are three: the demography, which describes the quantity and quality of human beings; the stock of knowledge that the society possesses, which determines the human command over nature; and the institutional framework that determines the rules of the game. The demographic characteristics include not only the fertility, mortality, and migration characteristics and the labor force composition, but also the stock of human capital (derived from the stock of knowledge). The stock of knowledge includes not only the scientific knowledge that a society possesses, its distribution in the society, and its application to solving problems of scarcity, but also the beliefs that the society holds that influence the choices made. That stock of knowledge determines the potential upper bound of the well-being of the society. The institutional framework determines the incentive structure of the society. It is the interplay between these three that shapes the features of the economy. We know very little about this interaction, although we do have some limited hypotheses about parts of the interaction. Self conscious modeling of this interaction at a moment of time, much less over time, has not been part of the agenda of economists, development economists, or economic historians.

But with this background we are now ready to explore the process of economic change.

3. Process of Economic Change

A bare-bones description of that process is straightforward. The "reality" of a political-economic system is never known to anyone, but humans do construct elaborate beliefs about the nature of that "reality"—beliefs that are both a positive model of the way the system works and a normative model of how it should work. The belief system may be broadly held within the society, reflecting a consensus of beliefs; or widely disparate beliefs may be held, reflecting fundamental divisions in perception about the society. The dominant beliefs—those of political and economic entrepreneurs in a position to make policies—produce over time an elaborate structure of institutions—both formal rules and informal norms—that determines economic/political performance. The resultant institutional matrix imposes severe constraints on the choice set of entrepreneurs when they seek to introduce new or modified institutions in order to improve their economic or political positions. The resultant path dependence typically makes change incremental. But change is continually occurring (although the rate will depend on the degree of competition among organizations and their entrepreneurs) as entrepreneurs enact policies to improve their competitive position—policies that result in alterations of the institutional matrix described in the previous section. The result is revised perceptions of reality, and in new efforts by entrepreneurs to improve their position. The process of change is never ending. Change can also come from non-human induced changes in the human landscape, such as natural disasters; but overwhelmingly it is humans themselves who are incrementally altering the human landscape, as even the most cursory overview of human history will attest.

It is one thing to be able to provide a summary description of the process of economic change; it is something else to provide sufficient content to this description to give us an understanding of this process. What do we mean by reality? How do beliefs get formed? How do they change? What is the relationship between beliefs and institutions?

I have nothing to add to the age old question of philosophers—what is reality? But I do have a direct pragmatic interest in just what it is that we are trying to model in our theories, beliefs, and ideologies. The pragmatic concern is with the degree to which our beliefs coincide with "reality". To the extent that they do then there is some prospect that the policies we enact will produce the intended result. The model is always a very imperfect reflection of how the economy really works. In some cases the defects are fatal, as in the case of the communist economies that disintegrated in 1989.

Beliefs and the way they evolve are at the heart of understanding the process of change. For the most part economists, with a few important exceptions like Hayek, have ignored the role of ideas in making choices. While the rationality assumption has served economists well for a limited range of issues in

micro-theory, it has devastating limitations in dealing with the process of economic change. The way we perceive the world and construct our explanations about that world requires that we delve into how the mind and brain work—the subject matter of cognitive science. We are some distance from a theory of learning that would account for how the mind works but we can at least outline the nature of the process.

The first level of learning entails developing a structure by which to make sense of the varied signals received by the senses. The initial architecture of the structure is genetic, but its subsequent development is a result of the experiences of the individual. This architecture can be thought of as generating an event space which gets us to interpret the data provided by the world. The experiences can be classified into two kinds—those from the physical environment and those from the socio-cultural linguistic environment (Hutchins and Hazlehurst, 1992). The event space structure consists of categories—classifications that gradually evolve from earliest childhood on in order to organize our perceptions and keep track of our memory of analytic results and experiences. Building on these categories, we form mental models to explain and interpret the environment, typically in ways relevant to some goals (Holland et al., p. 22). Both the categories and the mental models will evolve to reflect the feedback derived from new experiences—feedback that may strengthen and confirm our initial categories and models or that may lead to modifications. Thus, the event space may be continually redefined with experience, including contact with others' ideas.

Learning which preserves the categories and concepts intact but provides changed ideas about details and the applicability of the existing knowledge is the second level of learning. Together, learning within a given set of concepts and learning which changes the structure of concepts and mental models suggest an approach to the dynamics of learning.

The belief systems that evolve from learning induce political and economic entrepreneurs in a position to make choices that shape micro and macro economic performance to erect an elaborate structure of rules, norms, conventions and beliefs embodied in constitutions, property rights, and informal constraints; these in turn shape economic performance. This "scaffolding" not only constrains the choice set at a moment of time but is the source of path dependence. Thus when political or economic entrepreneurs seek to alter some aspect of economic performance they make choices that are constrained not only by the standard constraints of technology and income but also by this scaffolding. The process of institutional change described above is intended to alter performance in a particular direction. The aggregate of such institutional changes is continually altering the way the economy works. In turn that leads to gradual alterations of the models we devise in a never ending process of economic change.

Throughout history humans have typically gotten it (at least partly) wrong in 1) their understanding of the way the economy works, 2) the synthetic frameworks they construct, or 3) the policies they enact (at best blunt instruments to serve their purposes) which produce unanticipated consequences. We may

write economic history as a great success story of the enormous increase in material well-being which has reflected the secular growth in the stock of knowledge. But it is also a vast panorama of decisions that have produced death, famine, starvation, defeat in warfare, economic decline and stagnation, and indeed the total disappearance of civilizations. And even the decisions made in the success stories have typically been an admixture of luck intermingled with shrewd judgments and unanticipated outcomes. Take American economic history. From the earliest attempts at settlement, through the colonial era, to the perceptions leading to the revolutionary war the colonists had it, at best, half right. The Constitution, surely a classic of shrewd judgment, was aided by chance (the events of the 1780s), luck (the boycott of the Convention by the anti-federalists), and unanticipated decisions (the development of the independent judiciary and the Marshall court).

I wish to emphasize the limits to our understanding because there is a certain amount of hubris evident in the annual surveys by the World Bank and in the writing of orthodox economists who think that now we have it right. But it is important that we understand that even if we did have it right for one economy it would not necessarily be right for another economy and even if we have it right today it would not necessarily be right tomorrow. I am not suggesting that we haven't learned a good deal about determinants of economic performance. We have; but the implications of my brief survey of the sequence of steps from our understanding of an economy, to the scaffold we erect, to the policies we then enact to alter economic performance are that there are innumerable junctures where we can and do get it wrong. Crucial junctures, critical to the issues of improving the performance of economies, have resulted from the way scaffolds have evolved and policies were formed as well as the way time has affected the formation of beliefs.

Scaffolds include the political structure that specifies how we develop and aggregate political choices, the property rights structure that defines the formal incentives in the economy, and the informal constraints of norms, conventions and internally held beliefs. They have evolved over many generations, reflecting, as Hayek has reminded us (1960), the trial and error process which has sorted out those behavioral patterns that have worked from those that have failed. Because the experience of every society has been unique, they will differ for each economy. They constrain the choice set not only because the organizations of that economy have been built on the foundations of that institutional structure and therefore their survival depends on its continuance but also, and perhaps more fundamentally, because the belief system that is a complementary part of that scaffolding tends to change very slowly. This scaffolding is what makes path dependence so important. When the scaffolding crumbles, as it did in eastern Europe in 1989, the problems of constructing a new framework have exposed our limited understanding of the process of change.

Equally crucial are the policies that we enact to alter the performance of an economy. Even when we have a "correct" understanding of the economy and the (more or less) "correct" theory about its operation, the policies at our disposal

are very blunt instruments. They consist of alterations in the formal rules only, when in fact the performance of an economy is an admixture of the formal rules; the informal norms, and their enforcement characteristics. Changing only the formal rules will produce the desired results only when the informal norms that are complementary to that rule change and enforcement is either perfect or at least consistent with the expectations of those altering the rules.

Finally, time is important because it is the dimension in which human learning occurs and there is no implication in the foregoing brief description of the process of learning that suggests that we get it right. Indeed throughout history we have gotten it wrong far more often than we have gotten it right. The rise and fall of communism in the twentieth century is only a recent illustration. It is probably correct that if "reality" stayed constant the feedback from the policies we enacted would gradually lead us to get it right, but change and therefore persistent uncertainty is our lot which guarantees that we will continue to get it wrong at least part of the time.

4. Implications and Lacunae

The implications of the foregoing brief outline of the process of economic change are straightforward. If our objective is to improve the long run performance of economies we are in possession of the essential characteristics of successful economies. The best single predictor of the growth of an economy remains its investment rate. The new growth economics literature highlights some of the specific features of successful development. What is glaringly missing from this literature is the incentive structure to realize these objectives. But we do know a good deal about the institutional foundations of successful development. A number of recent empirical studies have made clear the importance of the institutional matrix (see Knack and Keefer, 1995 for a good summary). That matrix broadly comprises the incentive structure which will determine the quantity and quality of investment. What is still missing is how to get there. The key is the way path dependence will constrain the process of institutional and economic change.

The implication of the foregoing analysis is that path dependence can and will produce a wide variety of patterns of development depending on the cultural heritage and specific historical experience of the economy. Indeed the success of TVEs (township and village enterprises—a form of organization that is neither a firm nor a cooperative) in China does not fit our preconceptions about successful institutional/organizational structures and has been a sobering reminder of how much we still have to learn about the process. A description of that process in China from the enactment of the household responsibility system traces a unique path which has produced (so far) rapid economic growth (although even that success is tempered by growing problems of TVEs). I would hope that this paper puts to rest for good any simplistic general nostras such as "big bang" or "shock therapy" theories to magically overcome lack of development.

If path dependence can help us to understand the variety of development patterns, it also speaks forcefully to the constraints that the scaffolds erected in an economy impose on institutional change. The historically derived constraints are supported not only by the existing organizations that oppose change but also by the belief system that has evolved to produce those constraints. The rate and direction of change will be determined by the "strength" of the existing organizations and belief system.

The demise of communism in Eastern Europe in 1989 reflected a collapse of the existing belief system and consequent weakening of the supporting organizations. Policy makers were confronted not only by the problems of restructuring an entire society but also by the blunt instrument that is inherent in policy changes that can only alter the formal rules but cannot alter the accompanying norms and even have had only limited success in inducing enforcement of policies. The relative success of policy measures—such as the auctioning of state assets and the reestablishment of a legal system—in the Czech Republic compared to Russia resulted from the heritage of informal norms that made for the relatively harmonious establishment of the new rules (although even here the system for shifting assets from public to private hands in the Czech Republic produced some adverse and unanticipated downstream consequences).

One of the shortcomings of research is the lack of attention paid to the polity and the problem of aggregating choices through the political system. We simply have no good models of polities in third world, transition, or other economies. The interface between economics and politics is still in a primitive state in our theories but its development is essential if we are to implement policies consistent with intentions.

5. Conclusion

Let me conclude by talking again about time. If you accept the crude schematic outline of the process of change I laid out in section 3 above, it is clear that change is an ongoing continuous affair and that typically our institutional prescriptions reflect the learning from past experience. But there is no guarantee that the past experiences are going to equip us to solve new problems. Indeed an historic dilemma of fundamental importance has been the difficulties of economies shifting from a political economy based on personal exchange to one based on impersonal exchange. An equally wrenching change can be the movement from a "command" economy to a market economy. In both cases the necessity to restructure institutions—both economic and political—has been a major obstacle to development; it still is the major obstacle for third world and transition economies. The belief system that has evolved as a result of the cumulative past experiences of a society has not equipped the members to confront and solve the new problems.

We are just beginning systematically to explore the process of economic change. Our laboratory is not only our history but, particularly, what we are

learning in the ongoing efforts to improve the performance of third world and transition economies. We have made some progress but we still have a long way to go.

References

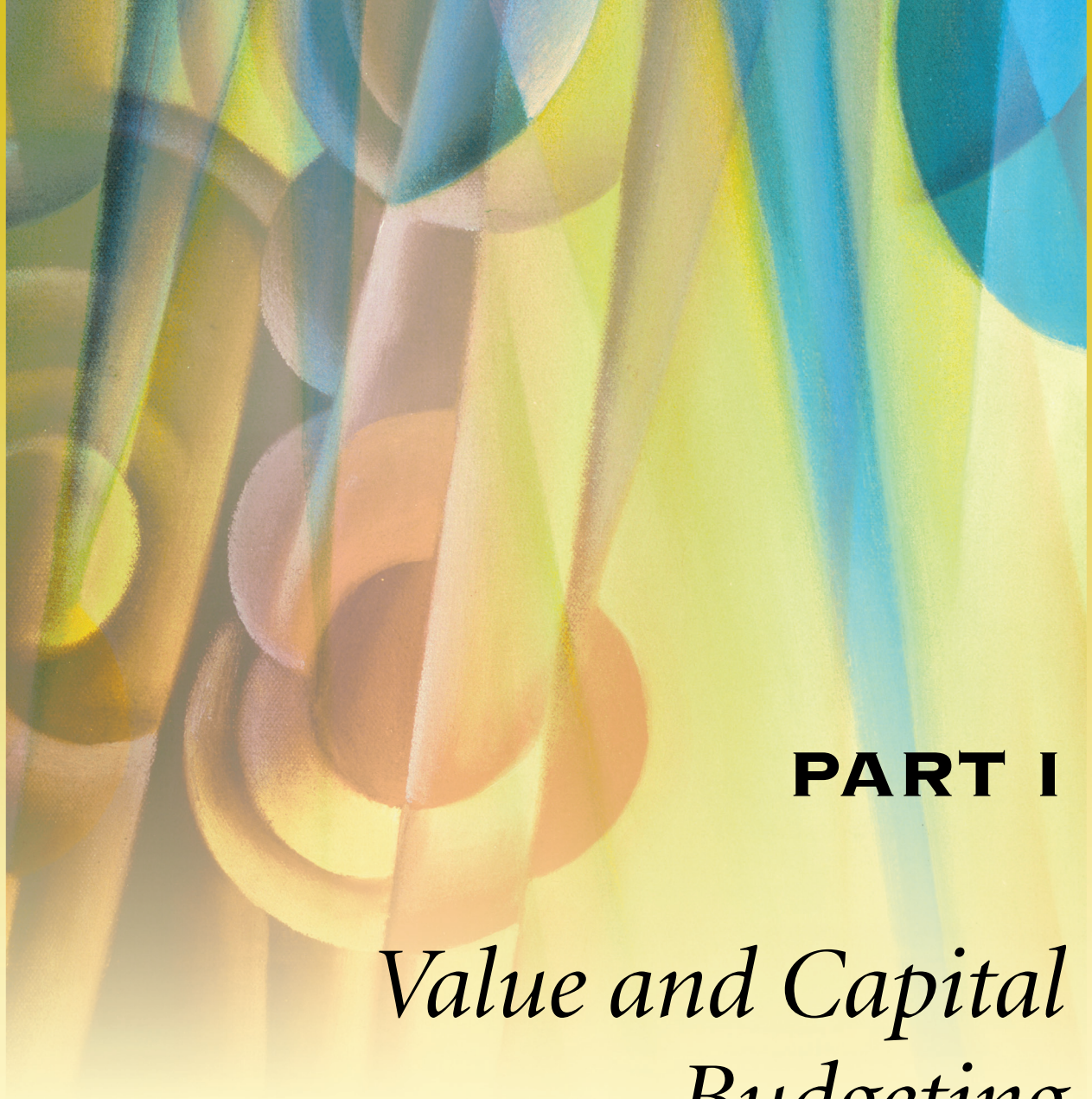
Hahn, Frank H. 1987. "Information, Dynamics, and Equilibrium". *Scottish Journal of Political Economy* 34(4): 321–34.

Hayek, F. A. 1960. *The Constitution of Liberty*. Chicago: The University of Chicago Press.

Holland, John H., Keith J. Idalyoak, Richard E. Nisbett, and Paul R. Thagard. 1986. *Induction: Processes of Inference, Learning, and Discovery*. Cambridge, MA: MIT Press.

Hutchins, Edwin and Brian Hazlehurst. 1992. "Learning in the Cultural Process" in C. G. Langston, C. Taylor, I. Farmer, and S. Rasmussen, (eds.), *Artificial Life II.* Redwood City, CA.: Addison-Wesley, pp. 682–706.

Knack, Stephen, and Philip Keefer. 1995. "Institutional and Economic Performance: Cross Country Tests Using Alternative Institutional Measures". *Economics and Politics* 7(3): 207–227.



PART I

Value and Capital Budgeting

IN A PERFECT MARKET UNDER RISK NEUTRALITY

The two primary goals of this first part of the book (Chapters 2–6) are to explain how to work with rates of return and how to decide whether to accept or reject investment projects. We assume in this part that there are no taxes, no transaction costs, no disagreements, and no limits as to the number of sellers and buyers in the market. This is the so-called perfect market. I will explain later why a perfect market makes your life a lot easier.

WHAT YOU WANT TO LEARN IN THIS PART

- In Chapter 2, we start with the simplest possible scenario. In addition to the perfect market, we assume that there is no uncertainty: You know everything. And we assume that all rates of return in the economy are the same: A 1-year investment pays the same and perfectly known rate of return per annum as a 10-year investment. Under these assumptions, you learn how 1-year returns translate into multiyear returns and when you should accept or reject a project. The chapter introduces the important concept of "present value."

  Typical questions: If you earn 5% per year, how much will you earn over 10 years? If you earn 100% over 10 years, how much will you earn per year? What is the value of a project that will deliver $1,000,000 in 10 years? Should you buy this project if it cost you $650,000? What inputs do you need to decide this?

- In Chapter 3, you learn how to value particular kinds of projects—perpetuities and annuities—if the economy-wide interest rate remains constant. You then learn how to apply the formulas to the valuation of stocks and bonds. The popular Gordon dividend growth model for valuing stocks assumes that dividends are a simple growing perpetuity cash flow stream, which makes it a perfect application of the perpetuity formula. Mortgages and other bonds are good applications of pricing using the annuities formulas.

  Typical questions: If a firm pays $1/share dividends next year, growing by 3% per year forever, then what should its stock price be? What is the monthly payment for a $300,000 mortgage bond if the interest rate is 4% per year?

- In Chapter 4, you learn more about capital budgeting methods. Although net present value (NPV) is the correct method, at least one other common method often comes to the correct result: the internal rate of return. In the real world, a number of other, plainly incorrect, methods are also in wide use. You should know why you should be wary of them. This chapter also tells you what CFOs actually rely on.

  Typical questions: If a project has one investment outflow and two return inflows, how would you compute a "rate of return"? Can you accept projects whose rates of return are above their cost of capital? How bad is it when you use incorrect estimates—as you inevitably will—in your calculations? What are the big problems with a rule that accepts those projects that return money most quickly?

- In Chapter 5, you abandon the assumption that annual rates of return are the same regardless of the length of time of your investment. For example, 1-year investments may pay 2% per year, while 10-year investments may pay 5% per year. The scenario of time-varying rates of return is more realistic, but the questions that you want to answer still remain the same as those in Chapter 2. (The chapter then also explains more advanced aspects of bonds, such as the Treasury yield curve.)

  Typical questions: If you earn 5% in the first year and 10% in the second year, how much will you earn over both years? What is the meaning of a 4% annualized interest rate? What is the meaning of a 4% yield-to-maturity? How can you value projects if appropriate rates of return depend on different time horizons?

- In Chapter 6, you abandon the assumption that you know the future. To be able to study uncertainty in the real world, you must first learn how to describe it. This

is done with statistics, the necessary aspects of which are explained here, too. The chapter then introduces risk neutrality, which is an assumption that can make it easier to understand some concepts in finance under uncertainty. Perhaps the two most important concepts are the difference between promised and expected rates of return and the difference between debt and equity. Under uncertainty, a project may not return the promised amount. Because of the possibility of default, the *stated* rate of return must be higher than the *expected* rate of return. Although you are interested in the latter, it is almost always only the former that you are quoted (promised). It is important that you always draw a sharp distinction between promised (stated) rates of return and expected rates of return. The second concept that this chapter explains is the difference between debt and equity—corporate claims that have a meaningful difference only under uncertainty.

Typical questions: If there is a 2% chance that your borrower will not return the money, how much extra interest should you charge? From an investment perspective, what is the difference between debt and equity? What is financing priority? What is a residual claim?

Looking ahead, Part II will continue with uncertainty scenarios in which investors are risk averse. Part III will explain what happens when financial markets or decision rules are not perfect.



2

The Time Value of Money and Net Present Value

THE MOTHER OF ALL FINANCE

We begin with the concept of a rate of return—the cornerstone of finance. You can always earn interest by depositing your money today into the bank. This means that money today is more valuable than the same amount of money next year. This concept is called the *time value of money*—$1 in present value is better than $1 in future value.

Investors make up just one side of the financial markets. They give money today in order to receive money in the future. Firms make up the other side. The process firms use to decide what to do with their money—which projects to take and which projects to pass up—is called *capital budgeting*. You will learn that there is one clear best method for making this critical decision. The firm should translate all *future* cash flows—both inflows and outflows—into their equivalent *present values* today, and then add them up to find the *net present value*, or NPV. The firm should take all projects that have positive net present values and reject all projects that have negative net present values.

This all sounds more complex than it is, so we'd better get started.

2.1 OUR BASIC SCENARIO: PERFECT MARKETS, CERTAINTY, CONSTANT INTEREST RATES

As promised, we begin with the simplest possible scenario. In finance, this means that we assume that we are living in a so-called **perfect market**:

We start with a so-called perfect market.

- There are no taxes.
- There are no transaction costs (costs incurred when buying and selling).

Welch third pages 2008/8/15 13:46 p. 13 (chap02) Windfall Software, JLS ZzTEX 13.8

- There are no differences in information or opinions among investors.
- There are so many buyers and sellers (investors and firms) in the market that the presence or absence of just one (or a few) individuals does not have an influence on the price.

The perfect market allows us to focus on the basic concepts in their purest forms, without messy real-world factors complicating the exposition. We will use these assumptions as our sketch of how financial markets operate, though not necessarily how firms' product markets work. You will learn in Chapter 10 how to operate in a world that is not perfect. (This will be a lot messier.)

In early chapters only, we add even stronger assumptions.

In this chapter, we will make three additional assumptions (that are not required for a market to be considered "perfect") to further simplify the world:

- There is no risk or uncertainty. You have perfect foresight.
- There is no inflation.
- The interest rate per period is the same.

Of course, this financial utopia is unrealistic. However, the tools that you will learn in this chapter will also work in later chapters, where the world becomes not only progressively more realistic but also more difficult. Conversely, if any tool does not give the right answer in our simple world, it would surely make no sense in a more realistic world.

SOLVE NOW!

Q 2.1 What are the four perfect market assumptions?

2.2 LOANS AND BONDS

Finance jargon: interest, loan, bond, fixed income, maturity.

The material in this chapter is easiest to explain in the context of bonds and loans. A **loan** is the commitment of a borrower to pay a predetermined amount of cash at one or more predetermined times in the future (the final one called **maturity**), usually in exchange for cash up front today. Loosely speaking, the difference between the money lent and the money paid back is the **interest** that the lender earns. A **bond** is a particular kind of loan, so named because it "binds" the borrower to pay money. Thus, for an investor, "buying a bond" is the same as "extending a loan." Bond buying is the process of giving cash today and receiving a promise for money in the future. Similarly, from the firm's point of view, it is "giving a bond," "issuing a bond," or "selling a bond." Loans and bonds are also sometimes called **fixed income**, because they "promise" a fixed amount of payments to the holder of the bond.

Why learn bonds first? Because they are easiest.

You should view a bond as just another type of investment project—money goes in, and money comes out. In Chapter 5, you will learn more about Treasuries, which are bonds issued by the U.S. Treasury. The beauty of such bonds is that you know what the cash flows will be. Besides, much more capital in the economy is tied up in bonds and loans than is tied up in stocks, so understanding bonds well is very useful in itself.

Interest rates: limited upside. Rates of return: arbitrary upside.

You already know that the net return on a loan is called interest, and that the rate of return on a loan is called the **interest rate**—though we will soon firm up your

knowledge about interest rates. One difference between an interest payment and a noninterest payment is that the former usually has a maximum payment, whereas the latter can have unlimited upside potential. However, not every rate of return is an interest rate. For example, an investment in a lottery ticket is not a loan, so it does not offer an interest rate, just a rate of return. In real life, its payoff is uncertain—it could be anything from zero to an unlimited amount. The same applies to stocks and many corporate projects. Many of our examples use the phrase "interest rate," even though the examples almost always work for any other rates of return, too.

Bond: defined by payment next year. Savings: defined by deposit this year.

Is there any difference between buying a bond for \$1,000 and putting \$1,000 into a bank savings account? Yes, a small one. The bond is defined by its future promised payoffs—say, \$1,100 next year—and the bond's value and price today are based on these future payoffs. But as the bond owner, you know exactly how much you will receive next year. An investment in a bank savings account is defined by its investment today. The interest rate can and will change every day, so you do not know what you will end up with next year. The exact amount depends on future interest rates. For example, it could be \$1,080 (if interest rates decrease) or \$1,120 (if interest rates increase).

A bank savings account is like a sequence of 1-day bonds.

If you want, you can think of a savings account as a sequence of consecutive 1-day bonds: When you deposit money, you buy a 1-day bond, for which you know the interest rate this one day in advance, and the money automatically gets reinvested tomorrow into another bond with whatever the interest rate will be tomorrow.

SOLVE NOW!

Q 2.2 Is a deposit into a savings account more like a long-term bond investment or more like a series of short-term bond investments?

2.3 RETURNS, NET RETURNS, AND RATES OF RETURN

Defining return and our time. Our convention is that 0 means "right now."

The most fundamental financial concept is that of a return. The payoff or (dollar) **return** of an investment is simply the amount of cash (C) it returns. For example, an investment project that returns \$12 at time 1 has

$$\text{Return at Time 1} = \$12$$
$$\text{Return}_1 \quad = C_1$$

The subscript is an instant in time, usually abbreviated by the letter t. When exactly time 1 occurs is not important: It could be tomorrow, next month, or next year. But if we mean "right now," we use the subscript 0.

Defining net return and rate of return.

The net payoff, or **net return**, is the difference between the return and the initial investment. It is positive if the project is profitable and negative if it is unprofitable. For example, if the investment costs \$10 today and returns \$12 at time 1 with nothing in between, then it earns a net return of \$2. Notation-wise, we really should use two

subscripts on returns—the time when the investment starts (0) and when it ends (1). This would make it something like "Net Return$_{0,1}$." Yikes! Let's just omit the first subscript on such flows when it is zero.

$$\text{Net Return from Time 0 to Time 1} = \$12 - \$10 = \$2$$

$$\text{Net Return}_1 = C_1 - C_0$$

The **rate of return** is the net return expressed as a percentage of the initial investment.

$$\text{Rate of Return from Time 0 to Time 1} = \frac{\$2}{\$10} = 20\%$$

$$r_1 = \frac{\text{Net Return from Time 0 to Time 1}}{\text{Purchase Price at Time 0}}$$

Often, it is convenient to calculate this as

$$r_1 = \frac{\$12 - \$10}{\$10} = \frac{\$12}{\$10} - 1 = 20\%$$

$$r_1 = \frac{C_1 - C_0}{C_0} = \frac{C_1}{C_0} - 1 \tag{2.1}$$

Rates of return are used so often that they have their own unique letter, r. Percent (the symbol %) is a unit of 1/100. 20% is the same as 0.20.

How to compute returns with interim payments. Capital gains versus returns.

Many investments have interim payments. For example, many stocks pay interim cash **dividends**, many bonds pay interim cash **coupons**, and many real estate investments pay interim rent. How would you calculate the rate of return then? One simple method is to just add interim payments to the numerator. Say an investment costs \$92, pays a dividend of \$5 (at the end of the period), and then is worth \$110. Its rate of return is

$$r = \frac{\$110 + \$5 - \$92}{\$92} = \frac{\$110 - \$92}{\$92} + \frac{\$5}{\$92} = 25\%$$

$$r_1 = \frac{C_1 + \text{All Dividends from 0 to 1} - C_0}{C_0} = \underbrace{\frac{C_1 - C_0}{C_0}}_{\text{Percent Price Change}} + \underbrace{\frac{\text{All Dividends}}{C_0}}_{\text{Dividend Yield}}$$

When there are intermittent payments and final payments, then returns are often broken down into two additive parts. The first part, the price change or **capital gain**, is the difference between the purchase price and the final price, *not* counting interim payments. Here, the capital gain is the difference between \$110 and \$92, that is, the \$18 change in the price of the investment. It is often quoted in percent of the price, which would be \$18/\$92 or 19.6% here. The second part is the amount received in interim payments. It is the dividend or coupon or rent, here \$5. When it is divided by the price, it has names like **dividend yield**, **current yield**, **rental yield**, or **coupon yield**, and these are also usually stated in percentage terms. In our example, the dividend yield is \$5/\$92 $\approx$ 5.4%. Of course, if the interim yield is high, you might be experiencing a negative capital gain and still have a positive rate of return. For

example, a bond that costs \$500, pays a coupon of \$50, and then sells for \$490, has a **capital loss** of \$10 (which comes to a −2% capital yield) but a rate of return of (\$490 + \$50 − \$500)/\$500 = +8%. You will almost always work with rates of return, not with capital gains. The only exception is when you have to work with taxes, because the IRS treats capital gains differently from interim payments. (We will cover taxes in Section 10.4.)

➤ Taxes on capital gains, Section 10.4, p. 323

Most of the time, people (incorrectly but harmlessly) abbreviate a rate of return or net return by calling it just a return. For example, if you say that the return on your \$10,000 stock purchase was 10%, you obviously do not mean you received a unitless 0.1. You really mean that your rate of return was 10% and you received \$1,000. This is usually benign, because your listener will know what you mean. Potentially more harmful is the use of the phrase *yield*, which, strictly speaking, means *rate of return*. However, it is often misused as a shortcut for dividend yield or coupon yield (the percent payout that a stock or a bond provides). If you say that the yield on your stock was 5%, then some listeners may interpret it to mean that you earned a total rate of return of 5%, whereas others may interpret it to mean that your stock paid a dividend yield of 5%.

People often use incorrect terms, but the meaning is usually clear, so this is harmless.

Interest rates should logically always be positive. After all, you can always earn 0% if you keep your money under your mattress—you thereby end up with as much money next period as you have this period. Why give your money to someone today who will give you less than 0% (less money in the future)? Consequently, interest rates are indeed almost always positive—the rare exceptions being both bizarre and usually trivial.

(Nominal) interest rates are usually nonnegative.

Here is another language problem: What does the statement "the interest rate has just increased by 5%" mean? It could mean either that the previous interest rate, say, 10%, has just increased from 10% to 10% · (1 + 5%) = 10.5%, or that it has increased from 10% to 15%. Because this is unclear, the **basis point** unit was invented. A basis point is simply 1/100 of a percent. If you state that your interest rate has increased by 50 basis points, you definitely mean that the interest rate has increased from 10% to 10.5%. If you state that your interest rate has increased by 500 basis points, you definitely mean that the interest rate has increased from 10% to 15%.

Basis points avoid an ambiguity in the English language: 100 basis points equals 1%.

IMPORTANT: 100 basis points constitute 1%. Basis points avoid "percentage ambiguities."

SOLVE NOW!

Q 2.3 A project offers a return of \$1,050 for an investment of \$1,000. What is the rate of return?

Q 2.4 A project offers a net return of \$25 for an investment of \$1,000. What is the rate of return?

Q 2.5 Is 10 the same as 1,000%?

Q 2.6 You purchase a stock for \$40 per share today. It will pay a dividend of \$1 next month. If you can sell it for \$45 right after the dividend is paid,

ANECDOTE Interest Rates over the Millennia

Historical interest rates are fascinating, perhaps because they look so similar to today's interest rates. Nowadays, typical interest rates range from 2% to 20% (depending on other factors). For over 2,500 years, from about the thirtieth century B.C.E. to the sixth century B.C.E., normal interest rates in Sumer and Babylonia hovered around 10–25% per annum, though 20% was the legal maximum. In ancient Greece, interest rates in the sixth century B.C.E. were about 16–18%, dropping steadily to about 8% by the turn of the millennium. Interest rates in ancient Egypt tended to be about 10–12%. In ancient Rome, interest rates started at about 8% in the fifth century B.C.E. but began to increase to about 12% by the third century A.C.E. (a time of great upheaval). When lending resumed in the late Middle Ages (twelfth century), personal loans in England fetched about 50% per annum, though they tended to hover between 10–20% in the rest of Europe. By the Renaissance, commercial loan rates had fallen to 5–15% in Italy, the Netherlands, and France. By the seventeenth century, even English interest rates had dropped to 6–10% in the first half, and to 3–6% in the second half. Mortgage rates tended to be lower yet. Most of the American Revolution was financed with French and Dutch loans at interest rates of 4–5%.

what would be its dividend yield, what would be its capital gain (also quoted as a capital gain yield), and what would be its total holding rate of return?

Q 2.7 If the interest rate of 9% increases to 12%, how many basis points did it increase?

Q 2.8 If the interest rate of 10% decreased by 20 basis points, what is the new interest rate?

2.4 THE TIME VALUE OF MONEY, FUTURE VALUE, AND COMPOUNDING

Because you can earn interest, a given amount of money today is worth more than the same amount of money in the future. After all, you could always deposit your money today into the bank and thereby get back more money in the future. This is an example of the concept of the **time value of money**, which says that a dollar today is worth more than a dollar tomorrow. This is one of the most basic and important concepts in finance.

2.4A THE FUTURE VALUE OF MONEY

Here is how to calculate future payoffs given a rate of return and an initial investment.

How much money will you receive in the future if the rate of return is 20% and you invest \$100 today? Turn around the rate of return formula (Formula 2.1) to determine how money will grow over time given a rate of return:

$$20\% = \frac{\$120 - \$100}{\$100} \Leftrightarrow \$100 \cdot (1 + 20\%) = \$100 \cdot 1.2 = \$120$$

$$r_1 = \frac{C_1 - C_0}{C_0} \Leftrightarrow C_0 \cdot (1 + r_1) = C_1$$

The \$120 next year is called the **future value** (**FV**) of \$100 today. Thus, future value is the value of a present cash amount at some point in the future. It is the time value of

money that causes the future value, \$120, to be higher than its present value (PV), \$100. Using the abbreviations FV and PV, you could also have written the above formula as

$$r_1 = \frac{\text{FV} - \text{PV}}{\text{PV}} \Leftrightarrow \text{FV} = \text{PV} \cdot (1 + r)$$

(If we omit the subscript on the r, it means a 1-period interest rate from now to time 1, i.e., r_1.) Please note that the time value of money is not the fact that the prices of goods may change between today and tomorrow (that would be inflation). Instead, the time value of money is based exclusively on the fact that your money can earn interest. Any amount of cash today is worth more than the same amount of cash tomorrow. Tomorrow, it will be the same amount plus interest.

➤ Section 5.2, "Inflation," p. 97

SOLVE NOW!

Q 2.9 A project has a rate of return of 30%. What is the payoff if the initial investment is \$250?

2.4B COMPOUNDING AND FUTURE VALUE

Now, what if you can earn the same 20% year after year and reinvest all your money? What would your 2-year rate of return be? Definitely *not* 20% + 20% = 40%! You know that you will have \$120 in year 1, which you can reinvest at a 20% rate of return from year 1 to year 2. Thus, you will end up with

Interest on interest (or rate of return on rate of return) means rates cannot be added.

$$\$100 \cdot (1 + 20\%)^2 = \$100 \cdot 1.2^2 = \$120 \cdot (1 + 20\%) = \$120 \cdot 1.2 = \$144$$

$$C_0 \cdot (1 + r)^2 \quad = \quad C_1 \cdot (1 + r) \quad = \quad C_2$$

This \$144—which is, of course, again a future value of \$100 today—represents a total 2-year rate of return of

$$r_2 = \frac{\$144 - \$100}{\$100} = \frac{\$144}{\$100} - 1 = 44\%$$

$$\frac{C_2 - C_0}{C_0} \quad = \quad \frac{C_2}{C_0} - 1 \quad = \quad r_2$$

This is more than 40% because the original net return of \$20 in the first year earned an additional \$4 in interest in the second year. You earn interest on interest! This is also called **compound interest**. Similarly, what would be your 3-year rate of return? You would invest \$144 at 20%, which would provide you with

$$C_3 = \$144 \cdot (1 + 20\%) = \$144 \cdot 1.2 = \$100 \cdot (1 + 20\%)^3 = \$100 \cdot 1.2^3 = \$172.80$$

$$C_3 = \quad C_2 \cdot (1 + r) \quad = \quad C_0 \cdot (1 + r)^3 \quad = \quad C_3$$

Your 3-year rate of return from time 0 to time 3, call it r_3, would thus be

$$r_3 = \frac{\$172.80 - \$100}{\$100} = \frac{\$172.80}{\$100} - 1 = 72.8\%$$

$$\frac{C_3 - C_0}{C_0} \quad = \quad \frac{C_3}{C_0} - 1 \quad = \quad r_3$$

This formula translates the three sequential 1-year rates of return into one 3-year **holding rate of return**—that is, what you earn if you hold the investment for the entire period. This process is called **compounding**, and the formula that does it is the "one-plus formula":

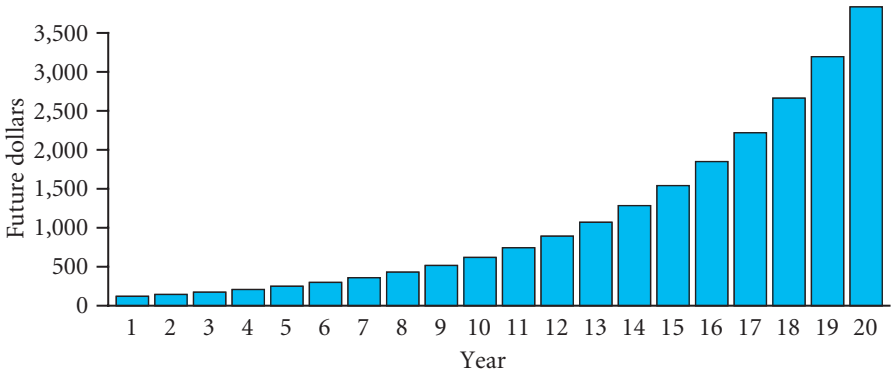
$$(1 + 72.8\%) = (1 + 20\%) \cdot (1 + 20\%) \cdot (1 + 20\%)$$

$$(1 + r_3) \quad = \quad (1 + r) \cdot (1 + r) \cdot (1 + r)$$

or, if you prefer it shorter,

$$1.728 = 1.2^3$$

Figure 2.1 shows how your \$100 would grow if you continued investing it at a rate of return of 20% per annum. The function is exponential—that is, it grows faster and faster as interest earns more interest.



Period	Start value	1 + one-year rate	End value	Total factor on \$100	Total rate of return $r_t = (1 + r)^t - 1$
0 to 1	\$100	(1 + 20%)	\$120.00	1.2	20.0%
1 to 2	\$120	(1 + 20%)	\$144.00	1.2 · 1.2 = 1.44	44.0%
2 to 3	\$144	(1 + 20%)	\$172.80	1.2 · 1.2 · 1.2 = 1.728	72.8%
⋮					

Money grows at a constant rate of 20% per annum. If you compute the graphed value at 20 years out, you will find that each dollar invested right now is worth \$38.34 in 20 years. The money at first grows in a roughly linear pattern, but as more and more interest accumulates and itself earns more interest, the graph accelerates steeply upward.

FIGURE 2.1 Compounding over 20 Years at 20% per Annum

IMPORTANT: The compounding formula translates sequential future rates of return into an overall holding rate of return:

$$\underbrace{(1+r_t)}_{\substack{\text{Multiperiod Holding}\\ \text{Rate of Return}}} = \underbrace{(1+r^t)}_{\substack{\text{Multiperiod Holding}\\ \text{Rate of Return}}} = \underbrace{(1+r)}_{\substack{\text{Current 1-Period}\\ \text{Spot Rate of Return}}} \cdot \underbrace{(1+r)}_{\substack{\text{Next 1-Period}\\ \text{Rate of Return}}} \cdots \underbrace{(1+r)}_{\substack{\text{Final 1-Period}\\ \text{Rate of Return}}}$$

The first rate is called the spot rate because it starts now (on the spot).

The compounding formula is so common that you must memorize it.

Another example of a payoff computation.

You can use the compounding formula to compute all sorts of future payoffs. For example, an investment project that costs \$212 today and earns 10% each year for 12 years will yield an overall holding rate of return of

$$r_{12} = (1 + 10\%)^{12} - 1 = (1.1^{12} - 1) \approx 213.8\%$$

$$(1 + r^t) - 1 \qquad = \quad r_{12}$$

Your \$212 investment today would therefore turn into a future value of

$$C_{12} = \$212 \cdot (1 + 10\%)^{12} = \$212 \cdot 1.1^{12} \approx \$212 \cdot (1 + 213.8\%) \approx \$665.35$$

$$C_0 \cdot (1 + r)^{12} \qquad = \quad C_{12}$$

"Uncompounding": Turn around the formula to compute individual holding rates.

Now suppose you wanted to know what constant two 1-year interest rates (r) would give you a 2-year rate of return of 50%. It is not 25%, because $(1 + 25\%) \cdot (1 + 25\%) - 1 = 1.25^2 - 1 = 56.25\%$. Instead, you need to solve

$$(1 + r) \cdot (1 + r) = (1 + r)^2 = 1 + 50\% = 1.50$$

The correct answer is

$$r = \sqrt[2]{1.50} - 1 \approx 22.47\%$$

$$= \sqrt[t]{1 + r_t} - 1 = \quad r$$

Check your answer: $(1 + 22.47\%) \cdot (1 + 22.47\%) = 1.2247^2 \approx (1 + 50\%)$. If the 12-month interest rate is 213.8%, what is the 1-month interest rate?

$$(1 + r)^{12} \qquad \approx \qquad 1 + 213.8\%$$

$$r = \sqrt[12]{1 + 213.8\%} - 1 = (1 + 213.8\%)^{1/12} - 1 \approx 10\%$$

➤ Exponentiation, Book Appendix, p. A-1

You can determine fractional time interest rates via compounding, too.

Interestingly, compounding works even over fractional time periods. Say the overall interest rate is 5% per year, and you want to find out what the rate of return over half a year would be. Because $(1 + r_{0.5})^2 = (1 + r_1)$, you would compute

$$(1 + r_{0.5}) = (1 + r_1)^{0.5} = (1 + 5\%)^{0.5} \approx 1 + 2.4695\% = 1.024695$$

ANECDOTE Life Expectancy and Credit

Your life expectancy may be 80 years, but 30-year bonds existed even in an era when life expectancy was only 25 years—at the time of Hammurabi, around 1700 B.C.E. (Hammurabi established the Kingdom of Babylon and is famous for the Hammurabi Code, the first known legal system.) Moreover, four thousand years ago, Mesopotamians already solved interesting financial problems. A cuneiform clay tablet contains the oldest known interest rate problem for prospective students of the financial arts. The student must figure out how long it takes for 1 mina of silver, growing at 20% interest per year, to reach 64 minae. Because the interest compounds in an odd way (20% of the principal is accumulated until the interest is equal to the principal, and then it is added back to the principal), the answer to this problem is 30 years, rather than 22.81 years. This is not an easy problem to solve—and it even requires knowledge of logarithms!

Check—compounding 2.4695% over two (6-month) periods indeed yields 5%:

$$(1 + 2.4695\%) \cdot (1 + 2.4695\%) = 1.024695^2 \approx (1 + 5\%)$$

$$(1 + r_{0.5}) \cdot (1 + r_{0.5}) \quad = (1 + r_{0.5})^2 = (1 + r_1)$$

You need logs to determine the time needed to get x times your money.

If you know how to use logarithms, you can also determine with the same formula how long it will take at the current interest rate to double or triple your money. For example, at an interest rate of 3% per year, how long would it take you to double your money?

$$(1 + 3\%)^x = (1 + 100\%) \Leftrightarrow x = \frac{\log(1 + 100\%)}{\log(1 + 3\%)} = \frac{\log(2.00)}{\log(1.03)} \approx 23.5$$

$$(1 + r)^t = (1 + r_t) \Leftrightarrow t = \frac{\log(1 + r_t)}{\log(1 + r)}$$

How Bad Are Mistakes?

ADDING OR COMPOUNDING INTEREST RATES?

Adding rather than compounding can make forgivably small mistakes in certain situations—but don't be ignorant of what you are doing.

Unfortunately, when it comes to interest rates in the real world, many users are casual, sometimes to the point where they are outright wrong. Some people mistakenly add interest rates instead of compounding them. When the investments, the interest rates, and the time periods are small, the difference between the correct and incorrect computation can be minor, so this practice can be acceptable, even if it is wrong. For example, when interest rates are 10%, compounding yields

$$(1 + 10\%) \cdot (1 + 10\%) - 1 = 1.1^2 - 1 = 21\%$$

$$(1 + r) \cdot (1 + r) - 1 \quad = r_2$$

$$= 1 + r + r + r \cdot r - 1$$

which is not exactly the same as the simple sum of two r's, which comes to 20%. The difference between 21% and 20% is the "cross-term" $r \cdot r$. This cross-product is especially unimportant if both rates of return are small. If the interest rate were both

1%, the cross-term would be 0.0001. This is indeed small enough to be ignored in most situations, and therefore a forgivable approximation. However, when you compound over many periods, you will accumulate more and more cross-terms, and eventually the quality of your approximation will deteriorate.

SOLVE NOW!

Q 2.10 If the 1-year rate of return is 20% and interest rates are constant, what is the 5-year holding rate of return?

Q 2.11 If you invest $2,000 today and it earns 25% per year, how much will you have in 15 years?

Q 2.12 What is the holding rate of return for a 20-year investment that earns 5%/year each year? What would a $200 investment grow to?

Q 2.13 A project lost one-third of its value each year for 5 years. What was its total holding rate of return? How much is left if the original investment was $20,000?

Q 2.14 If the 5-year holding rate of return is 100% and interest rates are constant, what is the (compounding) annual interest rate?

Q 2.15 What is the quarterly interest rate if the annual interest rate is 50%?

Q 2.16 If the per-year interest rate is 5%, what is the 2-year total interest rate?

Q 2.17 If the per-year interest rate is 5%, what is the 10-year total interest rate?

Q 2.18 If the per-year interest rate is 5%, what is the 100-year total interest rate? How does this compare to 100 times 5%?

Q 2.19 At a constant rate of return of 6% per annum, how many years does it take you to triple your money?

2.4C HOW BANKS QUOTE INTEREST RATES

Banks add to the confusion, quoting interest rates using strange but traditional conventions.

Banks and many other financial institutions use a number of conventions for quoting interest rates that may surprise you.

An **annual percentage yield** (**APY**) is the simple rate of return. (It is what our book calls an interest rate. Your bank sometimes calls this an **annual equivalent rate** (**AER**) or an **effective annual rate**.) If you invest $100, and the APY is 10%, you end up with $110 at the end of the year.

The **interest rate** stated without qualification is not really a rate of return, but just a method of quoting. The true daily interest rate is this annual interest quote divided by 365 (or 360 by another convention). For example, if your bank quotes you an annual interest rate of 10%, it means that the daily interest rate is $10\%/365 \approx 0.0274\%$. This is also why your bank may call this the **annual rate, compounded daily**. Therefore, if you leave your money in the bank for 1 year, you earn a true

$$\text{Actual Rate of Return} = [1 + (10\%/365)]^{365} - 1 \approx 10.52\%$$

In sum, at a quoted bank interest rate of 10%, $100 turns into $110.52 after 1 year.

An annual percentage rate (APR) is the rate that a bank is required to quote on loans it extends, according to the Consumer Credit Act of 1980. This act requires lenders to quote an "annual rate, compounded monthly," thus rendering APR as a number similar to a plain interest quote (not an APY). For example, if the quote to you is 10% per annum, then the lender will collect $(1 + 10\%/12)^{12} - 1 \approx 10.47\%$ per year on the money lent to you. For every $100 you borrow, you will have to pay the bank $10.47 every year. However, in contrast to the simple interest quote, APR not only has a different compounding interval, but is also required to reflect other closing costs and fees in order to aid consumers. Yet even though APR is supposedly a standardized measure, there are still enough variations in common use that comparing APRs may not always be comparing apples to apples.

A certificate of deposit (CD) is a longer-term investment vehicle than a savings account deposit. If your bank wants you to deposit your money in a CD, do you think it will put the more traditional interest rate quote or the APY on its sign in the window? Because the APY of 10.52% looks larger and thus more appealing to depositors than the traditional 10% interest rate quote, most banks advertise the APY for deposits. If you want to borrow money from your bank, do you think your loan agreement will similarly emphasize the APY? No. Most of the time, banks leave this number to the fine print and focus on the APR (or the traditional interest rate quote) instead.

Interest rates are not intrinsically difficult but they can be tedious, and definitional confusions abound in their world. My best advice when money is at stake: If in doubt, ask how the interest rate is computed! Even better, ask for a simple illustrative calculation.

SOLVE NOW!

Q 2.20 If you earn an (effective) interest rate of 12% per annum, how many basis points do you earn in interest on a typical calendar day?

Q 2.21 If the bank quotes an interest rate of 12% per annum (not as an effective interest rate), how many basis points do you earn in interest on a typical day?

Q 2.22 If the bank states an *effective* interest rate of 12% per annum, and there are 52.15 weeks, how much interest do you earn on a deposit of $100,000 over 1 week?

Q 2.23 If the bank quotes interest of 12% per annum, and there are 52.15 weeks, how much interest do you earn on a deposit of $100,000 over 1 week?

Q 2.24 If the bank quotes interest of 12% per annum, and there are 52.15 weeks, how much interest do you earn on a deposit of $100,000 over 1 year?

Q 2.25 If the bank quotes an interest rate of 6% per annum, what does a deposit of $100 in the bank come to after 1 year?

Q 2.26 If the bank quotes a loan APR rate of 8% per annum, compounded monthly, and there are no fees, what do you have to pay back in 1 year if you borrow $100 from the bank?

2.5 PRESENT VALUES, DISCOUNTING, AND CAPITAL BUDGETING

Capital budgeting: Should you budget capital for a project?

Now turn to the flip side of the future value problem: If you know how much money you will have next year, what does this correspond to in value *today*? This is especially important in a corporate context, where the question is, "Given that Project X will return $1 million in 5 years, how much should you be willing to pay to undertake this project today?" The process entailed in answering this question is called **capital budgeting** and is at the heart of corporate decision making. (The origin of the term was the idea that firms have a "capital budget," and that they must allocate capital to their projects within that budget.)

Start again with the rate of return formula

➤ Formula 2.1, p. 16

The "present value formula" *is* nothing but the rate of return definition—inverted to translate future cash flows into (equivalent) today's dollars.

$$r_1 = \frac{C_1 - C_0}{C_0} = \frac{C_1}{C_0} - 1$$

You only need to turn this formula around to answer the following question: If you know the prevailing interest rate in the economy (r_1) and the project's future cash flows (C_1), what is the project's value to you *today*? In other words, you are looking for the **present value** (**PV**)—the amount a future sum of money is worth today, given a specific rate of return. For example, if the interest rate is 10%, how much would you have to save (invest) to receive $100 next year? Or, equivalently, if your project will return $100 next year, what is the project worth to you today? The answer lies in the present value formula, which translates future money into today's money. You merely need to rearrange the rate of return formula to solve for the present value:

$$C_0 = \frac{\$100}{1 + 10\%} = \frac{\$100}{1.1} \approx \$90.91$$

$$C_0 = \frac{C_1}{1 + r_1} = \mathrm{PV}(C_1)$$

Check this—investing $90.91 at an interest rate of 10% will indeed return $100 next period:

$$10\% \approx \frac{\$100 - \$90.91}{\$90.91} = \frac{\$100}{\$90.91} - 1 \Leftrightarrow (1 + 10\%) \cdot \$90.91 \approx \$100$$

$$r_1 = \frac{C_1 - C_0}{C_0} = \frac{C_1}{C_0} - 1 \Leftrightarrow (1 + r_1) \cdot C_0 = C_1$$

Discounting translates future cash into today's equivalent.

This is the **present value formula**, which uses a division operation known as **discounting**. (The term "discounting" indicates that we are reducing a value, which is exactly what we are doing when we translate future cash into current cash.) If you wish, you can think of discounting—the conversion of a future cash flow amount into its equivalent present value amount—as the *reverse* of compounding.

Present value varies inversely with the cost of capital.

Thus, the present value (PV) of next year's $100 is $90.91—the value today of future cash flows. Let's say that this $90.91 is what the project costs. If you can borrow or lend at the interest rate of 10% elsewhere, then you will be indifferent between

receiving $100 next year and receiving $90.91 for your project today. In contrast, if the standard rate of return in the economy were 12%, your specific project would not be a good deal. The project's present value would be

$$\text{PV}(C_1) = \frac{\$100}{1 + 12\%} = \frac{\$100}{1.12} \approx \$89.29$$

$$C_0 = \frac{C_1}{1 + r_1} = \text{PV}(C_1)$$

which would be less than its cost of $90.91. But if the standard economy-wide rate of return were 8%, the project would be a great deal. Today's present value of the project's future payoff would be

$$\text{PV}(C_1) = \frac{\$100}{1 + 8\%} = \frac{\$100}{1.08} \approx \$92.59$$

$$C_0 = \frac{C_1}{1 + r_1} = \text{PV}(C_1)$$

which would exceed the project's cost of $90.91. It is the present value of the project, weighed against its cost, that should determine whether you should undertake a project today or avoid it. The present value is also the answer to the question, "How much would you have to save at current interest rates today if you wanted to have a specific amount of money next year?"

The PV formula with 2 periods.

Let's extend the time frame in our example. If the interest rate were 10% per period, what would $100 in 2 periods be worth today? The value of the $100 is then

$$\text{PV}(C_2) = \frac{\$100}{(1 + 10\%)^2} = \frac{\$100}{1.21} \approx \$82.64 \tag{2.2}$$

$$\text{PV}(C_2) = \frac{C_2}{(1 + r)^2} = C_0$$

Note the 21%. In 2 periods, you could earn a rate of return of $(1 + 10\%) \cdot (1 + 10\%) - 1 = 1.1^2 - 1 = 21\%$ elsewhere, so this is your appropriate comparable rate of return.

The interest rate can be called the "cost of capital."

This discount rate—the rate of return, r, with which the project can be financed—is often called the **cost of capital**. It is the rate of return at which you can raise money elsewhere. In a perfect market, this cost of capital is also the **opportunity cost** that you bear if you fund your specific investment project instead of the alternative next-best investment elsewhere. Remember—you can invest your money at this rate in another project instead of investing it in this one. The better these alternative projects in the economy are, the higher will be your cost of capital, and the lower will be the value of your specific investment project with its specific cash flows. An investment that promises $1,000 next year is worth less today if you can earn 50% rather than 5% elsewhere. A good rule is to always add mentally the word "opportunity" before "cost of capital"—it is always your **opportunity cost of capital**. (In this part of our book, I will just tell you what the economy-wide rate of return is—here 10%—

for borrowing or investing. In later chapters, you will learn how this rate of return is determined.)

IMPORTANT: Always think of the r in the present value denominator as your "opportunity" cost of capital. If you have great opportunities elsewhere, your projects have to be discounted at high discount rates. The discount rate, the cost of capital, and the required rate of return are really all just names for the same factor.

When you multiply a future cash flow by its appropriate **discount factor**, you end up with its present value. Looking at Formula 2.2, you can see that this discount factor is the quantity

The discount factor is a simple function of the cost of capital.

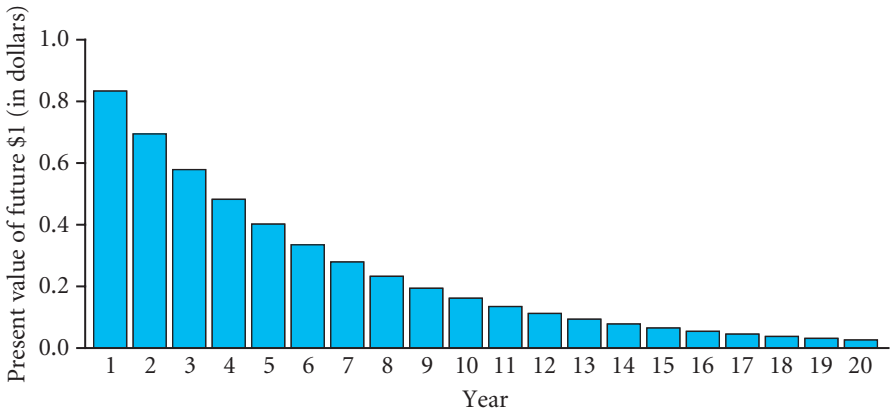
$$\left(\frac{1}{1+21\%}\right) \approx 0.8264$$

$$\left(\frac{1}{1+r_2}\right)$$

In other words, the discount factor translates 1 dollar in the future into the equivalent amount of dollars today. In the example, at a 2-year 21% rate of return, a dollar in 2 years is worth about 83 cents today. Because interest rates are usually positive, discount factors are usually less than 1—a dollar in the future is worth less than a dollar today. (Sometimes, and less correctly, people call this the **discount rate**, but this name should be used for r_t instead.)

Figure 2.2 shows how the discount factor declines when the cost of capital is 20% per annum. After about a decade, any dollar the project earns is worth less than 20 cents to you today. If you compare Figure 2.1 to Figure 2.2, you should notice how each is the "flip side" of the other.

The discount rate is higher for years farther out, so the discount factor is lower.



Each bar is $1/(1 + 20\%) \approx 83.3\%$ of the size of the bar to its left. After 20 years, the last bar is 0.026 in height. This means that $1 in 20 years is worth 2.6 cents in money today.

FIGURE 2.2 Discounting over 20 Years at a Cost of Capital of 20% per Annum

IMPORTANT: The cornerstones of finance are the following formulas:

$$\text{Rate of Return:} \qquad r_t = \frac{C_t - C_0}{C_0} = \frac{C_t}{C_0} - 1$$

Rearrange the formula to obtain the future value:

$$\text{Future Value:} \qquad \text{FV}_t = C_t = C_0 \cdot (1 + r_t) = C_0 \cdot (1 + r)^t$$

The process of obtaining r_t is called compounding, and it works through the "one-plus" formula:

$$\text{Compounding:} \qquad \underbrace{(1 + r_t)}_{\substack{\text{Total Holding} \\ \text{Rate of Return}}} = \underbrace{(1 + r)}_{\substack{\text{First Period} \\ \text{Rate of Return}}} \cdot \underbrace{(1 + r)}_{\substack{\text{Second Period} \\ \text{Rate of Return}}} \cdots \underbrace{(1 + r)}_{\substack{\text{Third Period} \\ \text{Rate of Return}}}$$

Rearrange the formula again to obtain the present value:

$$\text{Present Value:} \qquad \text{PV} = C_0 = \frac{C_t}{(1 + r_t)} = \frac{C_t}{(1 + r)^t}$$

The process of translating C_t into C_0—that is, the multiplication of a future cash flow by $1/(1 + r_t)$—is called discounting. The discount factor is:

$$\text{Discount Factor:} \qquad \frac{1}{(1 + r_t)} = \frac{1}{(1 + r)^t}$$

It translates 1 dollar at time t into its equivalent value today.

Bonds' present values and the prevailing interest rates move in opposite directions.

Remember how bonds are different from savings accounts? The former is pinned down by its promised fixed future payments, while the latter pays whatever the daily interest rate is. This induces an important relationship between the value of bonds and the prevailing interest rates—*they move in opposite directions*. For example, if you have a bond that promises to pay $1,000 in 1 year, and the prevailing interest rate is 5%, the bond has a present value of $1,000/1.05 ≈ $952.38. If the prevailing interest rate suddenly increases to 6% (and thereby becomes your new opportunity cost of capital), the bond's present value becomes $1,000/1.06 ≈ $943.40. You would have lost $8.98, which is about 0.9% of your original $952.38 investment. The value of your fixed-bond payment in the future has gone down, because investors can now do better than your 5% by buying new bonds. They have better opportunities elsewhere in the economy. They can earn a rate of return of 6%, not just 5%, so if you wanted to sell your bond now, you would have to sell it at a discount to leave the next buyer a rate of return of 6%. If you had delayed your investment, the sudden change to 6% would have done nothing to your investment. On the other hand, if the prevailing interest rate suddenly drops to 4%, then your bond will be more valuable. Investors would be willing to pay $1,000/1.04 ≈ $961.54, which is an immediate $9.16 gain. The inverse relationship between prevailing interest rates and bond prices is general and worth noting.

IMPORTANT: The price and the implied rate of return on a bond with fixed payments move in opposite directions. When the price of the bond goes up, its implied rate of return goes down. When the price of the bond goes down, its implied rate of return goes up.

SOLVE NOW!

Q 2.27 A project has a cost of capital of 30%. The final payoff is \$250. What should it cost today?

Q 2.28 A bond promises to pay \$150 in 12 months. The annual true interest rate is 5% per annum. What is the bond's price today?

Q 2.29 A bond promises to pay \$150 in 12 months. The bank quotes you interest of 5% per annum, compounded daily. What is the bond's price today?

Q 2.30 If the cost of capital is 5% per annum, what is the discount factor for a cash flow in 2 years?

Q 2.31 Interpret the meaning of the discount factor.

Q 2.32 What are the units on rates of return, discount factors, future values, and present values?

Q 2.33 Would it be good or bad for you, in terms of the present value of your liabilities, if your opportunity cost of capital increased?

Q 2.34 The price of a bond that offers a safe promise of \$100 in 1 year is \$95. What is the implied interest rate? If the bond's interest rate suddenly jumped up by 150 basis points, what would the bond price be? How much would an investor gain/lose if she held the bond while the interest rate jumped up by these 150 basis points?

2.6 NET PRESENT VALUE

Present values are alike and thus can be added, subtracted, compared, and so on.

An important advantage of present value is that all cash flows are translated into the same unit: cash today. To see this, say that a project generates \$10 in 1 year and \$8 in 5 years. You cannot add up these different future values to come up with \$18—it would be like adding apples and oranges. However, if you translate both future cash flows into their present values, you *can* add them. For example, if the interest rate was 5% per annum (so $(1 + 5\%)^5 = (1 + 27.6\%)$ over 5 years), the present value of these two cash flows together would be

$$\text{PV}(\$10 \text{ in 1 year}) = \frac{\$10}{1.05} \approx \$9.52$$

$$\text{PV}(\$8 \text{ in 5 years}) = \frac{\$8}{1.05^5} \approx \$6.27$$

$$\text{PV}(C_t) = \frac{C_t}{(1+r)^t}$$

Therefore, the total value of the project's future cash flows *today* (at time 0) is \$15.79.

The definition and use of NPV.

The **net present value** (**NPV**) of an investment is the present value of all its future cash flows minus the present value of its cost. It is really the same as present value, except that the word "net" up front reminds you to add and subtract *all* cash flows, including the *upfront* investment outlay today. The NPV calculation method is always the same:

1. Translate all future cash flows into today's dollars.
2. Add them all up. This is the present value of all future cash flows.
3. Subtract the initial investment.

NPV is the most important method for determining the value of projects. It is a cornerstone of finance. Let's assume that you have to pay \$12 to buy this particular project with its \$10 and \$8 cash flows. In this case, it is a positive NPV project, because

$$\text{NPV} = -\$12 + \frac{\$10}{1.05} + \frac{\$8}{1.05^5} \approx \$3.79$$

$$C_0 + \frac{C_1}{1+r_1} + \frac{C_5}{(1+r)^5} = \text{NPV}$$

(For convenience, we omit the 0 subscript for NPV, just as we did for PV.)

Think about what NPV means, and how it can be justified.

There are a number of ways to understand net present value.

- One way is to think of the NPV of \$3.79 as the difference between the market value of the future cash flows (\$15.79) and the project's cost (\$12)—this difference is the "value added."
- Another way to think of your project is to compare its cash flows to an equivalent set of bonds that exactly *replicates* them. In this instance, you would want to purchase a 1-year bond that promises \$10 next year. If you save \$9.52—at a 5% interest rate—you will receive \$10. Similarly, you could buy a 5-year bond that promises \$8 in year 5 for \$6.27. Together, these two bonds exactly replicate the project cash flows. The **law of one price** tells you that your project should be worth as much as this bond project—the cash flows are identical. You would have had to put away \$15.79 today to buy these bonds, but your project can deliver these cash flows at a cost of only \$12—much cheaper and thus better than your bond alternative.

Yet another way to justify NPV: opportunity cost.

- There is yet another way to think of NPV. It tells you how your project compares to the alternative opportunity of investing in the capital markets. These opportunities are expressed in the denominator through the discount factor. What would you get if you took your \$12 and invested it in the capital markets instead of in your project? Using the future value formula, you know that you could earn a 5% rate of return from now to next year, and 27.6% from now to 5 years. Your \$12 would grow into \$12.60 by next year. You could take out the same \$10 cash flow that your project gives you and be left with \$2.60 for reinvestment. Over the next 4 years, at the 5% interest rate, this \$2.60 would grow into \$3.16. But your project would do better for you, giving you \$8. Thus, your project achieves a higher rate of return than the capital markets alternative would achieve.

The conclusion of this argument is not only the simplest but also the best capital budgeting rule: If the NPV is positive, as it is here, you should take the project. If it is negative, you should reject the project. If it is zero, it does not matter.

The correct capital budgeting rule: Take all positive NPV projects.

IMPORTANT:

- The **net present value formula** is

$$\begin{aligned}\text{NPV} &= C_0 + \text{PV}(C_1) + \text{PV}(C_2) + \text{PV}(C_3) + \text{PV}(C_4) + \cdots \\ &= C_0 + \frac{C_1}{1+r_1} + \frac{C_2}{1+r_2} + \frac{C_3}{1+r_3} + \frac{C_4}{1+r_4} + \cdots \\ &= C_0 + \frac{C_1}{(1+r)} + \frac{C_2}{(1+r)^2} + \frac{C_3}{(1+r)^3} + \frac{C_4}{(1+r)^4} + \cdots\end{aligned}$$

  The subscripts are time indexes, C_t is the net cash flow at time t (positive for inflows, negative for outflows), and r_t is the relevant interest rate for investments from now to time t. With constant interest rates, $r_t = (1+r)^t - 1$.

- The **net present value capital budgeting rule** states that you should accept projects with a positive NPV and reject projects with a negative NPV.
- Taking positive NPV projects increases the value of the firm. Taking negative NPV projects decreases the value of the firm.
- NPV is definitively the best method for capital budgeting—the process by which you should accept or reject projects.

The NPV formula is so important that you must memorize it.

Let's work another NPV example. A project costs \$900 today, yields \$200/year for 2 years, then \$400/year for 2 years, and finally requires a cleanup expense of \$100. The prevailing interest rate is 5% per annum. These cash flows are summarized in Table 2.1. Should you take this project?

Let's work a project NPV example.

1. You need to determine the cost of capital for tying up money for 1 year, 2 years, 3 years, and so on. The compounding formula is

   First, determine your multiyear costs of capital.

$$(1 + r_t) = (1 + r)^t = (1.05)^t = 1.05^t$$

   So for money right now, the cost of capital r_0 is $1.05^0 - 1 = 0$; for money in 1 year, r_1 is $1.05^1 - 1 = 5\%$; for money in 2 years, r_2 is $1.05^2 - 1 = 10.25\%$. And so on.
2. You need to translate the cost of capital into discount factors. Recall that these are 1 divided by 1 plus your cost of capital. A dollar in 1 year is worth $1/(1 + 5\%) = 1/1.05 \approx 0.9524$ dollars today. A dollar in 2 years is worth $1/(1 + 5\%)^2 = 1/1.05^2 \approx 0.9070$. And so on.
3. You can now translate the future cash flows into their present value equivalents by multiplying the payoffs by their appropriate discount factors. For example, the \$200 cash flow at time 1 is worth about $0.9524 \cdot \$200 \approx \$190.48$.

TABLE 2.1 HYPOTHETICAL PROJECT CASH FLOW TABLE

Time	Project Cash Flow	Interest Rate		Discount Factor	Present Value
		Annualized	Holding		
t	C_t	r	r_t	$\frac{1}{(1+r)^t}$	$\text{PV}(C_t)$
Today	−$900	5.00%	0.00%	1.000	−$900.00
Year +1	+$200	5.00%	5.00%	0.9524	+$190.48
Year +2	+$200	5.00%	10.25%	0.9070	+$181.41
Year +3	+$400	5.00%	15.76%	0.8638	+$345.54
Year +4	+$400	5.00%	21.55%	0.8227	+$329.08
Year +5	−$100	5.00%	27.63%	0.7835	−$78.35
				Net Present Value (Sum):	$68.16

As a manager, you must estimate your project cash flows. The appropriate interest rate (also called cost of capital in this context) is provided to you by the opportunity cost of your investors—determined by the supply and demand for capital in the broader economy, where your investors can place their capital instead. The "Project Cash Flow" and the left interest rate column are the two input columns. The remaining columns are computed from these inputs. The goal is to calculate the final column.

4. Because present values are additive, you then sum up all the terms to compute the overall net present value. Make sure you include the original upfront cost as a negative.

Consequently, the project NPV is $68.16. Because this is a positive value, you should take this project.

If the upfront cost was higher, you should not take the project.

However, if the upfront expense was $1,000 instead of $900, the NPV would be negative (−$31.84), and you would be better off investing the money into the appropriate sequence of bonds from which the discount factors were computed. In this case, you should have rejected the project.

SOLVE NOW!

Q 2.35 Work out the present value of your tuition payments for the next 2 years. Assume that the tuition is $30,000 per year, payable at the start of the year. Your first tuition payment will occur in 6 months, and your second tuition payment will occur in 18 months. You can borrow capital at an interest rate of 6% per annum.

Q 2.36 Write down the NPV formula from memory.

Q 2.37 What is the NPV capital budgeting rule?

Q 2.38 Determine the NPV of the project in Table 2.1, if the per-period interest rate were 8% per year, not 5%. Should you take this project?

Q 2.39 You are considering a 3-year lease for a building, where you have to make one payment now, one in a year, and a final one in 2 years.

(a) Would you rather pay $1,000,000 up front, then $500,000 each in the following two years; or would you rather pay $700,000 each year?

(b) If the interest rate is 10%, what equal payment amount (rather than $700,000) would leave you indifferent? (This is also called the equivalent annual cost (EAC).)

➤ Section 3.4, "Projects With Different Lives and Rental Equivalents," p. 60

Q 2.40 Use a spreadsheet to answer the following question: Car dealer A offers a car for $2,200 up front (first payment), followed by $200 lease payments over the next 23 months. Car dealer B offers the same lease at a flat $300 per month (i.e., your first upfront payment is $300). Which lease do you prefer if the interest rate is 0.5% per month?

2.6A APPLICATION: ARE FASTER-GROWING FIRMS BETTER BARGAINS?

The firm's price should incorporate the firm's attributes.

Let's work another NPV problem applied to companies overall. Would it make more sense to invest in companies that grow quickly rather than slowly? If you wish, you can think of this question loosely as asking whether you should buy stocks in a fast-growing company like Google or in a slow-growing company like Procter & Gamble. The answer will be that this choice does not matter in a perfect market. Whether a company is growing quickly or slowly is already incorporated in the firm's price today, which is just the present value of the firm's cash flows that will accrue to the owners. Therefore, neither is the better deal.

Should you invest in a fast-grower or a slow-grower?

For example, consider company "Grow" (G) that will produce over the next 3 years

$$G_1 = \$100 \qquad G_2 = \$150 \qquad G_3 = \$250$$

and company "Shrink" (S) that will produce

$$S_1 = \$100 \qquad S_2 = \$90 \qquad S_3 = \$80$$

Is G not a better company to buy than S?

Let's find out: Compute the values.

There is no uncertainty involved, and both firms face the same cost of capital of 10% per annum. The price of G today is

$$\text{PV}(G) = \frac{\$100}{1.1^1} + \frac{\$150}{1.1^2} + \frac{\$250}{1.1^3} \approx \$402.70 \tag{2.3}$$

and the price of S today is

$$\text{PV}(S) = \frac{\$100}{1.1^1} + \frac{\$90}{1.1^2} + \frac{\$80}{1.1^3} \approx \$225.39$$

Your investment dollar grows at the same 10% rate. Your investment's growth rate is disconnected from the cash flow growth rate.

If you invest in G, then next year you will have $100 cash and own a company with $150 and $250 cash flows coming up. G's value at time 1 (so PV now has subscript 1) will thus be

$$\text{PV}_1(G) = \$100 + \frac{\$150}{1.1^1} + \frac{\$250}{1.1^2} \approx \$442.98$$

Your investment will have earned a rate of return of $\$442.98/\$402.70 - 1 \approx 10\%$. If you instead invest in S, then next year you will receive $100 cash and own a company

with "only" \$90 and \$80 cash flows coming up. *S*'s value will thus be

$$\text{PV}_1(S) = \$100 + \frac{\$90}{1.1^1} + \frac{\$80}{1.1^2} \approx \$247.93$$

Your investment will have earned a rate of return of \$247.93/\$225.39 − 1 ≈ 10%. In either case, you will earn the fair rate of return of 10%. Whether cash flows are growing at a rate of +50%, −10%, +237.5%, or −92% is irrelevant: *The firms' market prices today already reflect their future growth rates*. There is no necessary connection between the growth rate of the underlying project cash flows or earnings and the growth rate of your investment money (i.e., your expected rate of return).

Any sudden wealth gains would accrue to existing shareholders, not to new investors.

Make sure you understand the thought experiment here: This statement that higher-growth firms do not necessarily earn a higher rate of return does not mean that a firm in which managers succeed in increasing the future cash flows at no extra investment cost will not be worth more. Such firms will indeed be worth more, and the current owners will benefit from the rise in future cash flows, but this will also be reflected immediately in the price at which you can purchase this firm. This is an important corollary worth repeating. If General Electric has just won a large defense contract (like the equivalent of a lottery), shouldn't you purchase GE stock to participate in the windfall? Or if Wal-Mart managers do a great job and have put together a great firm, shouldn't you purchase Wal-Mart stock to participate in this windfall? The answer is that you cannot. The old shareholders of Wal-Mart are no dummies. They know the capabilities of Wal-Mart and how it will translate into cash flows. Why should they give you, a potential new shareholder, a special bargain for something to which you contributed nothing? Just providing more investment funds is not a big contribution—after all, there are millions of other investors equally willing to provide funds at the appropriately higher price. It is competition—among investors for providing funds and among firms for obtaining funds—that determines the expected rate of return that investors receive and the cost of capital that firms pay. There is actually a more general lesson here. Economics tells you that you must have a scarce resource if you want to earn above-normal profits. Whatever is abundant and/or provided by many will not be tremendously profitable.

SOLVE NOW!

Q 2.41 Assume that company *G* pays no interim dividends, so you receive \$536 at the end of the project. What is *G*'s market value at time 1, 2, and 3? What is your rate of return in each year? Assume that the cost of capital is still 10%.

Q 2.42 Assume that company *G* pays out the full cash flows (refer to the text example) in earnings each period. What is *G*'s market value at time 1, 2, and 3? What is your rate of return in each year?

Q 2.43 One month ago, a firm suffered a large court award against it that will force it to pay compensatory damages of \$100 million next January 1. Are shares in this firm a bad buy until January 2?

SUMMARY

This chapter covered the following major points:

- A perfect market assumes no taxes, no transaction costs, no opinion differences, and the presence of many buyers and sellers.
- A bond is a claim that promises to pay an amount of money in the future. Buying a bond is extending a loan. Issuing a bond is borrowing. Bond values are determined by their future payoffs.
- One hundred basis points are equal to 1%.
- The time value of money means that 1 dollar today is worth more than 1 dollar tomorrow, because of the interest that it can earn.
- Returns must not be averaged, but compounded over time.
- Interest rate quotes are *not* interest rates. For example, stated annual rates are usually not the effective annual rates that your money will earn in the bank. If in doubt, ask!
- The discounted present value (PV) translates future cash values into present cash values. The net present value (NPV) is the sum of all present values of a project, including the investment cost (usually, a negative upfront cash flow today).
- The values of bonds and interest rates move in opposite directions. A sudden increase in the prevailing economy-wide interest rate decreases the present value of a bond's future payouts and therefore decreases today's price of the bond. Conversely, a sudden decrease in the prevailing economy-wide interest rate increases the present value of a bond's future payouts and therefore increases today's price of the bond.
- The NPV formula can be written as

$$\begin{aligned} \text{NPV} &= C_0 + \frac{C_1}{1+r_1} + \frac{C_2}{1+r_2} + \cdots \\ &= C_0 + \frac{C_1}{1+r} + \frac{C_2}{(1+r)^2} + \cdots \end{aligned}$$

  In this context, r is called the discount rate or cost of capital, and $1/(1+r)$ is called the discount factor.
- The net present value capital budgeting rule states that you should accept projects with a positive NPV and reject projects with a negative NPV.
- In a perfect market, firms are worth the present value of their assets. Whether firms grow quickly or slowly does not make them more or less attractive investments in a perfect market because their prices always already reflect the present value of future cash flows.
- In a perfect market, the gains from sudden surprises accrue to old owners, not new capital providers, because old owners have no reason to want to share the spoils.

KEY TERMS

AER, 23
annual equivalent rate, 23
annual percentage rate, 24
annual percentage yield, 23
annual rate, compounded daily, 23
APR, 24
APY, 23
basis point, 17
bond, 14
capital budgeting, 25
capital gain, 16
capital loss, 17
CD, 24
certificate of deposit, 24
compounding, 20
compound interest, 19
cost of capital, 26
coupon, 16
coupon yield, 16
current yield, 16
discount factor, 27
discounting, 25
discount rate, 27
dividend, 16
dividend yield, 16
effective annual rate, 23
fixed income, 14
future value, 18
FV, 18
holding rate of return, 20
interest, 14
interest rate, 14, 23
law of one price, 30
loan, 14
maturity, 14
net present value, 30
net present value capital budgeting rule, 31
net present value formula, 31
net return, 15
NPV, 30
opportunity cost, 26
opportunity cost of capital, 26
perfect market, 13
present value, 25
present value formula, 25
PV, 25
rate of return, 16
rental yield, 16
return, 15
time value of money, 18

SOLVE NOW! SOLUTIONS

Q 2.1 The four perfect market assumptions are no taxes, no transaction costs, no differences in opinions, and no large buyers or sellers.

Q 2.2 A savings deposit is an investment in a series of short-term bonds.

Q 2.3 $r = (\$1{,}050 - \$1{,}000)/\$1{,}000 = 5\%$

Q 2.4 $r = \frac{\$25}{\$1{,}000} = 2.5\%$

Q 2.5 Yes, $10 = 1{,}000\%$.

Q 2.6 The dividend yield would be $\$1/\$40 = 2.5\%$, the capital gain would be $\$45 - \$40 = \$5$, so that its capital gain yield would be $\$5/\$40 = 12.5\%$, and the total rate of return would be $(\$46 - \$40)/\$40 = 15\%$.

Q 2.7 $1\% = 100$ basis points, so an increase of 3% is 300 basis points.

Q 2.8 20 basis points are 0.2%, so the interest rate declined from 10.0% to 9.8%.

Q 2.9 $r = 30\% = (x - \$250)/\$250 \quad \Longrightarrow x = 1.3 \cdot \$250 = \$325$

Q 2.10 $1.20^5 - 1 \approx 148.83\%$

Q 2.11 $\$2{,}000 \cdot 1.25^{15} \approx \$56{,}843.42$

Q 2.12 The total holding rate of return is $1.05^{20} - 1 \approx 165.33\%$, so you would end up with $\$200 \cdot (1 + 165.33\%) \approx \$530.66$.

Q 2.13 Losing one-third is a rate of return of -33%. To find the holding rate of return, compute $[1 + (-1/3)]^5 - 1 \approx -86.83\%$. About $(1 - 86.83\%) \cdot \$20{,}000 \approx \$2{,}633.74$ remains.

Q 2.14 $(1 + 100\%)^{1/5} - 1 \approx 14.87\%$

Q 2.15 $(1 + r_{0.25})^4 = (1 + r_1)$. Thus, $r_{0.25} = \sqrt[4]{1 + r_1} - 1 = 1.5^{1/4} - 1 \approx 10.67\%$.

Q 2.16 $r_2 = (1 + r_{0,1}) \cdot (1 + r_{1,2}) - 1 = 1.05 \cdot 1.05 - 1 = 10.25\%$

Q 2.17 $r_{10} = (1 + r_1)^{10} - 1 = 1.05^{10} - 1 \approx 62.89\%$

Q 2.18 $r_{100} = (1 + r_1)^{100} - 1 = 1.05^{100} - 1 = 130.5 \approx 13{,}050\%$. In words, this is about 130 times the initial investment, and substantially more than 500% (5 times the initial investment).

Q 2.19 Tripling is equivalent to earning a rate of return of 200%. Therefore, solve $(1 + 6\%)^x = (1 + 200\%)$, or $x \cdot \log(1.06) = \log(3.00)$ or $x = \log(3.00)/\log(1.06) \approx 18.85$ years.

Q 2.20 $(1 + r)^{365} = 1.12$. Therefore, $1.12^{(1/365)} - 1 \approx 0.00031054 = 0.031054\% \approx 3.1$ bp/day.

Q 2.21 The bank means to collect $12\%/365 \approx 3.288$ bp/day.

Q 2.22 The true daily interest rate, assuming 365 days, is $1.12^{1/365} - 1 \approx 0.031054\%$. To compute your true rate of return, compound this over 7 days: $(1 + 0.03105\%)^7 = 1.00031054^7 \approx 1.0021758$. (You could also compute the rate of return differently: There are 52.15 weeks in 365 days. Therefore, $r = (1 + 12\%)^{(1/52.15)} - 1 \approx 1.0021758$.) Your \$100,000 will grow into \$100,217.58. You would have earned \$217.58 in interest.

Q 2.23 With 12% in nominal APR interest *quoted*, you earn $12\%/365 \approx 0.032877\%$ per day. Therefore, the weekly rate of return is $(1 + 0.032877\%)^7 - 1 \approx 0.23036\%$. Your \$100,000 will grow into \$100,230.36. Note that you end up with more money from the 12% quoted rate than from the 12% effective rate.

Q 2.24 With 12% in nominal APR interest *quoted*, you earn $12\%/365 \approx 0.032877\%$ per day. Therefore, the annual rate of return is $(1 + 0.032877\%)^{365} - 1 \approx 12.747462\%$. Your \$100,000 will grow into \$112,747.46.

Q 2.25 The bank quote of 6% means that it will pay an interest rate of $6\%/365 \approx 0.0164384\%$ per day. This earns an actual interest rate of $(1 + 0.0164384\%)^{365} - 1 \approx 6.18\%$ per annum. Therefore, each invested \$100 grows to \$106.18, thus earning \$6.18 over the year.

Q 2.26 The bank quote of 8% means that you will have to pay an interest rate of $8\%/12 \approx 0.667\%$ per month. This earns an actual interest rate of $(1 + 0.667\%)^{12} - 1 \approx 8.30\%$ per annum. You will have to pay \$108.30 in repayment for every \$100 you borrowed.

Q 2.27 $r = 30\% = (\$250 - x)/x$. Thus, $x = \$250/1.30 \approx \$192.31$.

Q 2.28 $\$150/(1.05) \approx \$142.86$

Q 2.29 $\$150/[1 + (5\%/365)]^{365} \approx \$142.68$

Q 2.30 $1/[(1.05) \cdot (1.05)] \approx 0.9070$

Q 2.31 It is today's value in dollars for 1 future dollar, that is, at a specific point in time in the future.

Q 2.32 The rate of return and additional factors are unit-less. The latter two are in dollars (though the former is dollars in the future, while the latter is dollars today).

Q 2.33 Good. Your future payments would be worth less in today's money.

Q 2.34 The original interest rate is $\$100/\$95 - 1 \approx 5.26\%$. Increasing the interest rate by 150 basis points is 6.76%. This means that the price should be $\$100/(1.0676) \approx \$93.67$. A price change from \$95 to \$93.67 is a rate of return of $\$93.67/\$95 - 1 \approx -1.40\%$.

Q 2.35 The first tuition payment is worth $\$30{,}000/(1.06)^{1/2} \approx \$29{,}139$. The second tuition payment is worth $\$30{,}000/(1.06)^{3/2} \approx \$27{,}489$. Thus, the total present value is \$56,628.

Q 2.36 If you cannot write down the NPV formula by heart, do not go on until you have it memorized.

Q 2.37 Accept if NPV is positive. Reject if NPV is negative.

Q 2.38 $-\$900 + \$200/(1.08)^1 + \$200/(1.08)^2 + \$400/(1.08)^3 + \$400/(1.08)^4 - \$100/(1.08)^5 \approx \$0.14$. The NPV is positive. Therefore this is a worthwhile project that you should accept.

Q 2.39 For the 3-year building leases:

(a) Your preference depends on the interest rate. If the interest rate is zero, then you would prefer the $2 million sum-total payment to the $2.1 million rent. If the prevailing interest rate is less than 21.5%, it is better to lease. If it is more than 21.5%, you prefer the rent. For example, if it is 40%, the net present cost of the lease is $1.612 million, while the net present cost of the rent is $1.557 million.

(b) At a 10% interest rate, the total net present cost of the lease is $\$1 + \$0.5/1.1 + \$0.5/1.1^2 \approx \$1.868$ million. An equivalent rent contract must solve

$$x + \frac{x}{1.1} + \frac{x}{1.1^2} = \$1.868$$

Multiply by $1.1^2 = 1.21$

$$1.21 \cdot x + 1.1 \cdot x + x = \$1.868 \cdot 1.21$$

$$\Leftrightarrow x \cdot (1.21 + 1.1 + 1) = \$2{,}260.28$$

Therefore, the equivalent rental cost would be $x \approx \$682.864$.

Q 2.40 Lease A has an NPV of −$6,535. Lease B has an NPV of −$6,803. Therefore, lease A is cheaper.

Q 2.41 For easier naming, call 2000 your year 0. The firm's present value in 2000 is $\$536/1.10^3 \approx \$402.70$—but you already knew this. If you purchase this company, its value in 2001 depends on a cash flow stream that is $0 in 2001, $0 in year 2002, and $536 in year 2003. It will be worth $\$536/1.10^2 \approx \$442.98$ in 2001. In 2002, your firm will be worth $\$536/1.10 \approx \$487.27$. Finally, in 2003, it will be worth $536. Each year, you expect to earn 10%, which you can compute from the four firm values.

Q 2.42 Again, call 2000 your year 0. The firm's present value in 2000 is based on dividends of $100, $150, and $250 in the next three years. The firm value in 2000 is the $402.70 from page 33. The firm value in 2001 was also worked out to be $442.98, but you immediately receive $100 in cash, so the firm is worth only $\$442.98 - \$100 = \$342.98$. As an investor, you would have earned a rate of return of $\$442.98/\$402.70 - 1 \approx 10\%$. The firm value in 2002 is $PV_2(G) = \$250/1.1 \approx \$227.27$, but you will also receive $150 in cash, for a total firm-related wealth of $377.27. In addition, you will have the $100 from 2001, which would have grown to $110—for a total wealth of $487.27. Thus, starting with wealth of $442.98 and ending up with wealth of $487.27, you would have earned a rate of return of $\$487.27/\$442.98 - 1 \approx 10\%$. A similar computation shows that you will earn 10% from 2002 ($487.27) to 2003 ($536.00).

Q 2.43 No! The market will already have adjusted the price.

PROBLEMS

The ■ *indicates problems available in* myfinancelab

Q 2.44 What is a perfect market? What were the assumptions made in this chapter that were not part of the perfect market scenario?

Q 2.45 What is the difference between a bond and a loan?

Q 2.46 In the text, I assumed you received the dividend at the end of the period. In the real world, if you received the dividend at the beginning of the period instead of the end of the period, could this change your effective rate of return? Why?

Q 2.47 Your stock costs $100 today, pays $5 in dividends at the end of the period, and then sells for $98. What is your rate of return?

Q 2.48 The interest rate has just increased from 6% to 8%. How many basis points is this?

Q 2.49 Assume an interest rate of 10% per year. How much would you lose over 5 years if you had to give up interest on the interest—that is, if you received 50% instead of compounded interest?

Q 2.50 Over 20 years, would you prefer 10% per annum, with interest compounding, or 15% per annum but without interest compounding? (That is, you receive the interest, but it is put into an account that earns no interest, which is what we call simple interest.)

Q 2.51 A project returned +30%, then −30%. Thus, its arithmetic average rate of return was 0%. If you invested $25,000, how much did you end up with? Is your rate of return positive or negative? How would your overall rate of return have been different if you first earned –30% and then +30%?

Q 2.52 A project returned +50%, then −40%. Thus, its arithmetic average rate of return was +5%. Is your rate of return positive or negative?

Q 2.53 An investment for $50,000 earns a rate of return of 1% in each month of a full year. How much money will you have at year's end?

Q 2.54 There is always disagreement about what stocks are good purchases. The typical degree of disagreement is whether a particular stock is likely to offer, say, a 10% (pessimistic) or a 20% (optimistic) annualized rate of return. For a $30 stock today, what does the difference in belief between these two opinions mean for the expected stock price from today to tomorrow? (Assume that there are 365 days in the year. Reflect on your answer for a moment, and recognize that a $30 stock typically moves about ±$1 on a typical day. This unexplainable up-and-down volatility is often called noise.)

Q 2.55 If the interest rate is 5% per annum, how long will it take to double your money? How long will it take to triple it?

Q 2.56 If the interest rate is 8% per annum, how long will it take to double your money?

Q 2.57 From Fibonacci's *Liber Abaci*, written in the year 1202: "A certain man gave 1 denaro at interest so that in 5 years he must receive double the denari, and in another 5, he must have double 2 of the denari and thus forever. How many denari from this 1 denaro must he have in 100 years?"

Q 2.58 A bank quotes you a loan interest rate of 14% on your credit card. If you charge $15,000 at the beginning of the year, how much will you have to repay at the end of the year?

Q 2.59 Go to the website of a bank of your choice. What kind of quote does your bank post for a CD, and what kind of quote does your bank post for a mortgage? Why?

Q 2.60 What is the 1-year discount factor if the interest rate is 33.33%?

Q 2.61 You can choose between the following rent payments:

(a) A lump sum cash payment of $100,000;
(b) 10 annual payments of $12,000 each, the first occurring immediately;
(c) 120 monthly payments of $1,200 each, the first occurring immediately. (Friendly suggestion: This is a lot easier to calculate on a computer spreadsheet.)
(d) Which rental payment scheme would you choose if the interest rate was an effective 5% per year?
(e) Spreadsheet question: At what interest rate would you be indifferent between the first and the second choice above? (Hint: Graph the NPV of the second project as a function of the interest rate.)

Q 2.62 A project has cash flows of $15,000, $10,000, and $5,000 in 1, 2, and 3 years, respectively. If the prevailing interest rate is 15%, would you buy the project if it costs $25,000?

Q 2.63 Consider the same project that costs $25,000 with cash flows of $15,000, $10,000, and $5,000. At what prevailing interest rate would this project be profitable? Try different interest rates, and plot the NPV on the y-axis, and the interest rate on the x-axis.

Q 2.64 On April 12, 2006, Microsoft stock traded for $27.11 and claimed to pay an annual dividend of $0.36. Assume that the first dividend will be paid in 1 year, and that it then grows by 5% each year for the next 5 years. Further, assume that the prevailing interest rate is 6% per year. At what price would you have to sell Microsoft stock in 5 years in order to break even?

Q 2.65 Assume you are 25 years old. The IAW insurance company is offering you the following retirement contract (called an *annuity*): Contribute $2,000 per year for the next 40 years.

When you reach 65 years of age, you will receive $30,000 per year for as long as you live. Assume that you believe that the chance that you will die is 10% per year after you will have reached 65 years of age. In other words, you will receive the first payment with probability 90%, the second payment with probability 81%, and so on. Assume the prevailing interest rate is 5% per year, all payments occur at year-end, and it is January 1 now. Is this annuity a good deal? (Use a spreadsheet.)

Q 2.66 A project has the following cash flows in periods 1 through 4: −$200, +$200, −$200, +$200. If the prevailing interest rate is 3%, would you accept this project if you were offered an upfront payment of $10 to do so?

Q 2.67 Assume you are a real estate broker with an exclusive contract—the condo association rules state that everyone selling their condominiums must go through you or a broker designated by you. A typical condo costs $500,000 today and sells again every 5 years. This will last for 50 years, and then all bets are off. Your commission will be 3%. Condos appreciate in value at a rate of 2% per year. The interest rate is 10% per annum.

(a) What is the value of this exclusivity rule? In other words, at what price should you be willing to sell the privilege of exclusive condo representation to another broker?
(b) If free Internet advertising was equally effective and if it could replace all real estate brokers so that buyers' and sellers' agents would no longer earn the traditional 6% (3% each), what would happen to the value gain of the condo?

Q 2.68 If the interest rate is 5% per annum, what would be the equivalent annual cost (see Question 2.39) of a $2,000 lease payment up front, followed by $800 for three more years?

Q 2.69 The prevailing discount rate is 15% per annum. Firm *F*'s cash flows start with $500 in year 1 and grow at 20% per annum for 3 years. Firm *S*'s cash flows also start with $500 in year 1 but shrink at 20% per annum for 3 years. What are the prices of these two firms? Which one is the better "buy"?



8

Investor Choice: Risk and Reward

We are still after the same prize: a good estimate of the corporate cost of capital ($\mathcal{E}(\tilde{r})$) in the NPV formula. But before you can understand the opportunity costs of capital for your firm's own projects, you have to understand the other opportunities that your investors have. This means that you must understand better what investors like (reward) and what they dislike (risk), how they are likely to measure their risks and rewards, how diversification works, what portfolios smart investors are likely to hold, and why "market beta" is a good measure of the contribution of an investment asset to the market portfolio's risk.

8.1 MEASURING RISK AND REWARD

Put yourself into the shoes of an investor and start with the most basic questions: How should you measure the risk and reward of your portfolio? As always, we first cook up a simple example and then generalize our insights into a broader real-world context. Let's follow four risky assets (securities), named A through D, plus a risk-free asset named F. These assets could even be portfolios, themselves consisting of many individual portfolios, assets, and so on. (This is essentially what a mutual fund is.)

We work with five assets that have four equally likely outcomes.

There are four equally likely scenarios, named S1 through S4, as in Table 8.1. (If you find it easier to think in terms of historical outcomes, you can pretend that scenario S1 happened at time 1, S2 at time 2, and so forth, and you are now analyzing this historical data. This is not entirely correct, but often a helpful metaphor.) Which investment strategies do you deem better or worse, safer or riskier? It is the goal of this section to analyze the assets and scenarios in Table 8.1 to sharpen your understanding of the concepts and trade-offs of risk and reward.

Historical samples can be viewed as scenarios.

➤ Why this is not entirely correct, Section 8.5A, p. 223

Visuals always help, so Figure 8.1 graphs the returns in Table 8.1. Each scenario is equally likely (the histogram bars are all equally tall), so you can just indicate where each outcome lies on the x-axis. In this histogram plot, you prefer assets that have scenario outcomes farther to the right (they have higher rates of return), outcomes

In a histogram, bars to the right mean higher returns. Bars that are more spread out indicate higher risk.

Welch third pages 2008/8/25 8:38 p. 201 (chap08) Windfall Software, JLS ZzTEX 13.8

TABLE 8.1 RATES OF RETURN ON FIVE INVESTMENT ASSETS

	Assets' Rates of Return r				
Future	Pfio A (M)	Pfio B	Pfio C	Pfio D	Pfio F
In Scenario S1 ♣	−1.0%	+2.0%	−2.0%	+14.0%	+1.0%
In Scenario S2 ♦	+2.0%	+11.0%	+3.0%	+6.0%	+1.0%
In Scenario S3 ♥	+4.0%	−1.0%	+7.0%	0.0%	+1.0%
In Scenario S4 ♠	+11.0%	+4.0%	+12.0%	−12.0%	+1.0%
"Reward" ($\mathcal{E}(\tilde{r})$)	4.0%	4.0%	5.0%	2.0%	1.0%
"Variance" $\mathcal{V}ar(\tilde{r})$	19.5%%	19.5%%	26.5%%	90.0%%	0.0%%
"Risk" ($\mathcal{S}dv(\tilde{r})$)	4.42%	4.42%	5.15%	9.49%	0.00%

We use *pfio* as an abbreviation for *portfolio*. Variance ($\mathcal{V}ar$) and standard deviation ($\mathcal{S}dv$) were explained in Section 6.1B. The four scenarios are "stand-ins" for a much larger and exhaustive set of possible outcomes that could occur. For illustration, we assume that they are the only possible outcomes and that it is perfectly known that they occur with equal probability.

➤ Random variables *are* histograms, Section 6.1A, p. 138

that are *on average* farther to the right (they have higher *expected* rates of return), and outcomes that are more bunched together (they have less risk). Visual inspection shows that investment F has outcomes perfectly bunched at the same spot, so it is not only least risky but also risk-free. It is followed by the risky A and B, then C, and finally, the most risky, D.

8.1A MEASURING REWARD: THE EXPECTED RATE OF RETURN

Measure reward with the expected rate of return.

Although graphical measures are helpful, we also need algebraic formulas with associated numerical measures. A good measure for the **reward** is easy: You can use the **expected rate of return**, which is the probability-weighted average of all possible returns. For example, the mean rate of return for asset A is

$$\begin{aligned}\mathcal{E}(\tilde{r}_A) &= (1/4)\cdot(-1\%) + (1/4)\cdot(+2\%) + (1/4)\cdot(+4\%) + (1/4)\cdot(+11\%)\\ &= +4\%\end{aligned}$$

$$\mathcal{E}(\tilde{r}_A) = \mathcal{P}rob(S1)\cdot(\tilde{r}_{S1}) + \mathcal{P}rob(S2)\cdot(\tilde{r}_{S2}) + \mathcal{P}rob(S3)\cdot(\tilde{r}_{S3}) + \mathcal{P}rob(S4)\cdot(\tilde{r}_{S4})$$

If you invest in A, you would expect to earn a rate of return of 4%. Because each outcome is equally likely, you can compute this faster as a simple average:

$$\mathcal{E}(\tilde{r}_A) = \frac{(-1\%) + (+2\%) + (+4\%) + (+11\%)}{4} = 4\%$$

8.1B MEASURING RISK: THE STANDARD DEVIATION OF THE RATE OF RETURN

Measure risk with the standard deviation of the rate of return.

➤ Measures of risk (standard deviation), Section 6.1B, p. 141

A good measure of risk is less obvious than a good measure of reward. You already got a good peek at the most common risk measures in Section 6.1B, but let's do it again in the context of our specific set of securities. Figure 8.1 shows that A is more spread out than F (i.e., A is more risky than F) and less spread out than D (i.e., A is less risky than

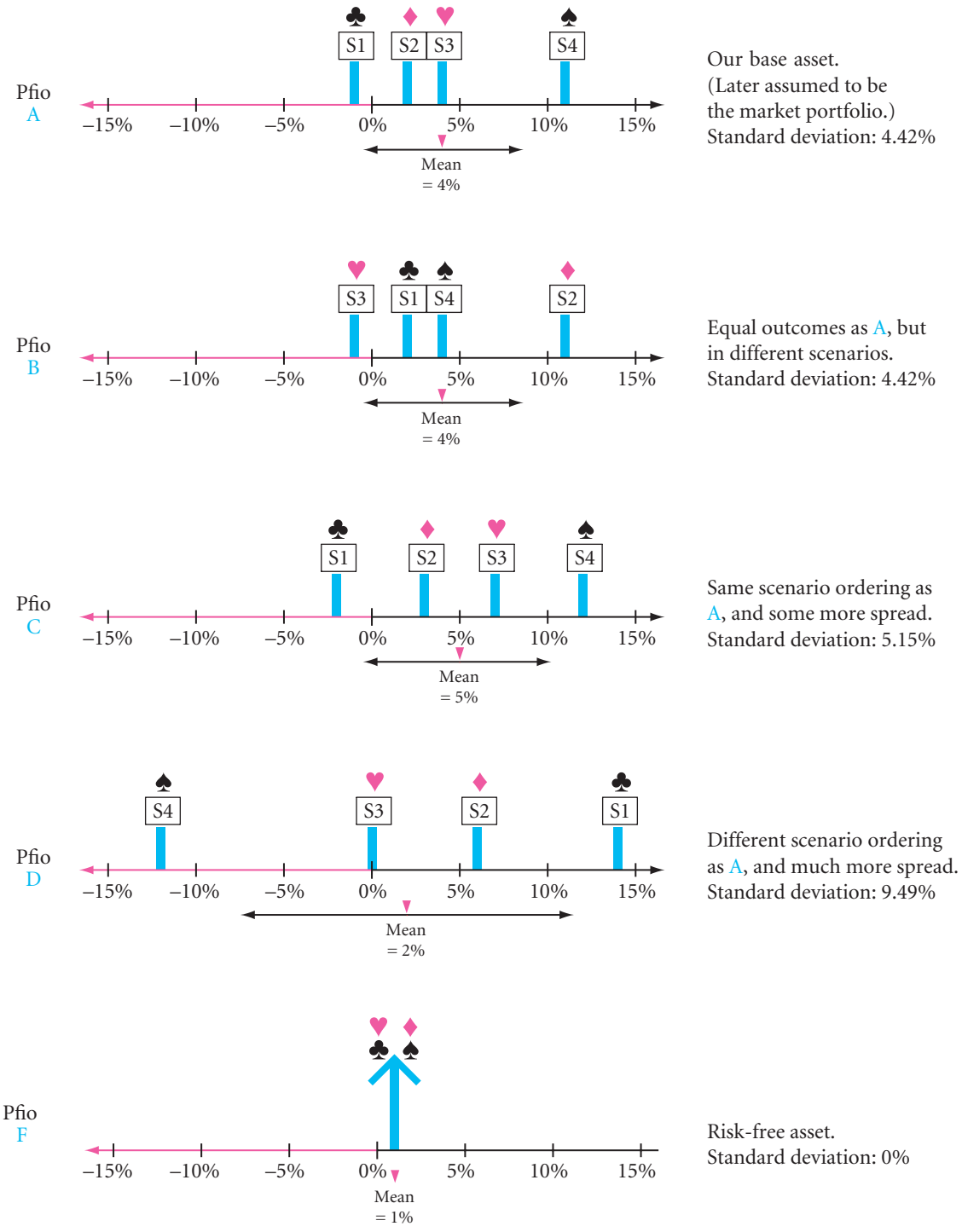


FIGURE 8.1 Graphical Perspectives on Performance

The graphs are standard histograms. Each outcome is equally likely, so each bar is 25% tall—with the exception of the bar in the final graph for the risk-free security, which is 100% tall. ♣ is the rate of return outcome in scenario S1, the ♦ in scenario S2, the ♥ in scenario S3, and the ♠ in scenario S4. The left-right arrows below the axes indicate the standard deviations.

D). A good first intuition is that it would make sense to rate each data point by how far away it is from the center (average). If your average is +4%, an outcome of 3% would be closer to the mean than an outcome of 0%. The former is only 1 unit away from the mean. The latter is 4 units away from the mean. It therefore makes sense to think in such deviations from the mean. Here is how A shapes up in terms of deviations from its mean:

Outcomes	In S1 (♣)	In S2 (♦)	In S3 (♥)	In S4 (♠)
Asset A Rate of Return	−1%	+2%	+4%	+11%
. . . in deviation from its 4% mean	−5%	−2%	0%	+7%

The average deviation from the mean is always 0. It cannot measure risk.

Unfortunately, you cannot compute risk as the average deviation from the mean. It is always zero—for example, the average here is $(-5 - 2 + 0 + 7)/4 = 0$. You must "neutralize" the sign, so that negative deviations count the same as positive deviations. The common fix is to compute the average *squared* deviation from the mean. This is called the **variance**:

$$\begin{aligned}\mathcal{V}ar(\tilde{r}_A) &= \frac{(-1\% - 4\%)^2 + (2\% - 4\%)^2 + (4\% - 4\%)^2 + (11\% - 4\%)^2}{4} \\ &= \frac{(-5\%)^2 + (-2\%)^2 + (0\%)^2 + (+7\%)^2}{4} = 19.5\%\% \\ &= \frac{[r_{S1} - \mathcal{E}(\tilde{r})]^2 + [r_{S2} - \mathcal{E}(\tilde{r})]^2 + [r_{S3} - \mathcal{E}(\tilde{r})]^2 + [r_{S4} - \mathcal{E}(\tilde{r})]^2}{\text{Number of Outcomes}} \\ &= \frac{\text{Sum over All Scenarios } \left[\tilde{r}_S \text{ in Scenario } S - \mathcal{E}(\tilde{r})\right]^2}{\text{Number of Scenarios}}\end{aligned}$$

The variance has units that are intrinsically impossible to interpret for ordinary humans (% *squared* $= 0.01 \cdot 0.01$, written as x%%). Therefore, the variance carries very little intuition, except that a higher variance means more risk.

The standard deviation of the portfolio's rate of return is a common measure of risk.

A measure that has more meaningful units is the **standard deviation**. It is just the square root of the variance,

$$\mathcal{S}dv(\tilde{r}_A) = \sqrt{\mathcal{V}ar(\tilde{r}_A)} = \sqrt{19.5\%\%} \approx 4.42\% \tag{8.1}$$

The standard deviation of the portfolio's rate of return is the most common measure of overall ***portfolio risk***. Looking at Figure 8.1, you can see that this standard deviation of 4.42% seems like a reasonable measure of how far the typical outcome of A is away from the mean of A. The last row in Table 8.1 also lists the standard deviations of B–F. In Figure 8.1, you can see their visual representations: F is risk-free; A and B are equally risky at 4.42%; C is a little more risky at 5.15%; and D is most risky at 9.49%.

IMPORTANT:

- You can measure investment portfolio reward by the expected rate of return on the *overall* portfolio.
- You can measure investment portfolio risk by the standard deviation of the rate of return on the *overall* portfolio.

(Warning: You will not measure the investment risk *contributions* of individual assets inside the portfolio via their standard deviations. This will be explained in Section 8.3B.)

A preview: Smart investors eliminate unnecessary risk. After they have done so, more reward requires taking more risk.

At this point, you should begin to wonder how risk and reward are related in a reasonable world. This will be the subject of much of the next chapter. The brief answer for now is that you can speculate in dumb ways that give you high investment risk with low reward—as anyone who has gambled knows. However, if you are smart, after eliminating all investment mistakes (the low-hanging fruit), you have no choice but to take on more risk if you want to earn higher rewards.

SOLVE NOW!

Q 8.1 What happens if you compute the average deviation from the mean, rather than the average squared deviation from the mean?

Q 8.2 Asset A from Table 8.1 offers −1%, +2%, +4%, and +11% with equal probabilities. Now add 5% to each of these returns. This new asset offers +4%, +7%, +9%, and +16%. Compute the expected rate of return, the variance, and the standard deviation of this new asset. How does it compare to A?

Q 8.3 Compute the risk and reward of C from Table 8.1.

8.2 PORTFOLIOS, DIVERSIFICATION, AND INVESTOR PREFERENCES

In the real world, you are usually not constrained to purchase assets in isolation—you can purchase a little of each. This has the important consequence of reducing your overall portfolio risk. Let's see why.

Portfolios are bundles of multiple assets. Their returns can be averaged.

Start again with investment assets A and B, which offer the same rates of return, but in different future scenarios. If you purchase $100 in either A or B, you would expect to earn $4 with a risk of $4.42. But what if you purchase $50 in A and $50 in B? Call this your investment portfolio P. In this case, your $100 investment would look like this:

Scenario Outcome	S1 (♣)	S2 (♦)	S3 (♥)	S4 (♠)	Average
Return on $50 in A:	$49.50	$51.00	$52.00	$55.50	$52.00
Return on $50 in B:	$51.00	$55.50	$49.50	$52.00	$52.00
⇒ Total return in P:	$100.50	$106.50	$101.50	$107.50	$104.00
Cost = $100 ⇒ Rate of return in P:	0.5%	6.5%	1.5%	7.5%	4.0%

You can do this more quickly by using the returns on A and B themselves. In this case, your portfolio P invests portfolio weight $w_A = 50\%$ into A and $w_B = 50\%$ in B. For example, to obtain the 6.5% in scenario S2, you could have computed the portfolio rate of return from A's 2% rate of return and B's 11% rate as

$$\tilde{r}_P = \tilde{r}_{50\% \text{ in A}, 50\% \text{ in B}} \text{ (in S2)} = 50\% \cdot 2\% \quad + \quad 50\% \cdot 11\% = 6.5\%$$

$$\tilde{r}_{P=(w_1, w_2, \ldots, w_N)} \quad = \quad w_1 \cdot \tilde{r}_1 \quad + \cdots + \quad w_N \cdot \tilde{r}_N$$

Thus, you could have computed P's four scenario rates of return as follows:

$$\text{In S1 } \clubsuit\text{:} \quad r_{\text{P=(50\% in A, 50\% in B) in S1}} = 50\% \cdot (-1\%) + 50\% \cdot (+2.0\%) = 0.5\%$$

$$\text{In S2 } \diamondsuit\text{:} \quad r_{\text{P=(50\% in A, 50\% in B) in S2}} = 50\% \cdot (+2\%) + 50\% \cdot (+11.0\%) = 6.5\%$$

$$\text{In S3 } \heartsuit\text{:} \quad r_{\text{P=(50\% in A, 50\% in B) in S3}} = 50\% \cdot (+4\%) + 50\% \cdot (-1.0\%) = 1.5\%$$

$$\text{In S4 } \spadesuit\text{:} \quad r_{\text{P=(50\% in A, 50\% in B) in S4}} = 50\% \cdot (+11\%) + 50\% \cdot (+4.0\%) = 7.5\%$$

$$r_{\text{P in S}} \quad = \quad w_A \cdot r_{\text{A in S}} \quad + \quad w_B \cdot r_{\text{B in S}}$$

Visually, the A and B combination portfolio called P has lower variability (risk and range) than either A or B.

Now look at these three possible investment portfolios: A, B, and P. The four outcomes are plotted sequentially on the x-axis in Figure 8.2, as if they had occurred in different months. (As noted, this is not a bad way to think about scenarios.) Each box has the name of its portfolio in it. The gray and magenta areas from −1% to +11% are the outcome ranges for single investments in either A or B. However, the investment of half in A and half in B has a much smaller range of outcomes (from 0.5% to 7.5%), as shown by the blue area. This portfolio P simply has less variability and range than either of its two components.

Algebraically, the combination portfolio also has lower risk.

Can you use the algebraic measures to back up your visual perception? The average (expected) rate of return of the combination portfolio P is the same 4% as that of A and B. The risk of the combination portfolio P is lower, however, than the risk of either A or B. In fact, it is

$$\mathcal{V}ar_{\substack{50\% \text{ in A} \\ 50\% \text{ in B}}} = \frac{(0.5\% - 4\%)^2 + (6.5\% - 4\%)^2 + (1.5\% - 4\%)^2 + (7.5 - 4\%)^2}{4}$$

$$= 9.25\%\%$$

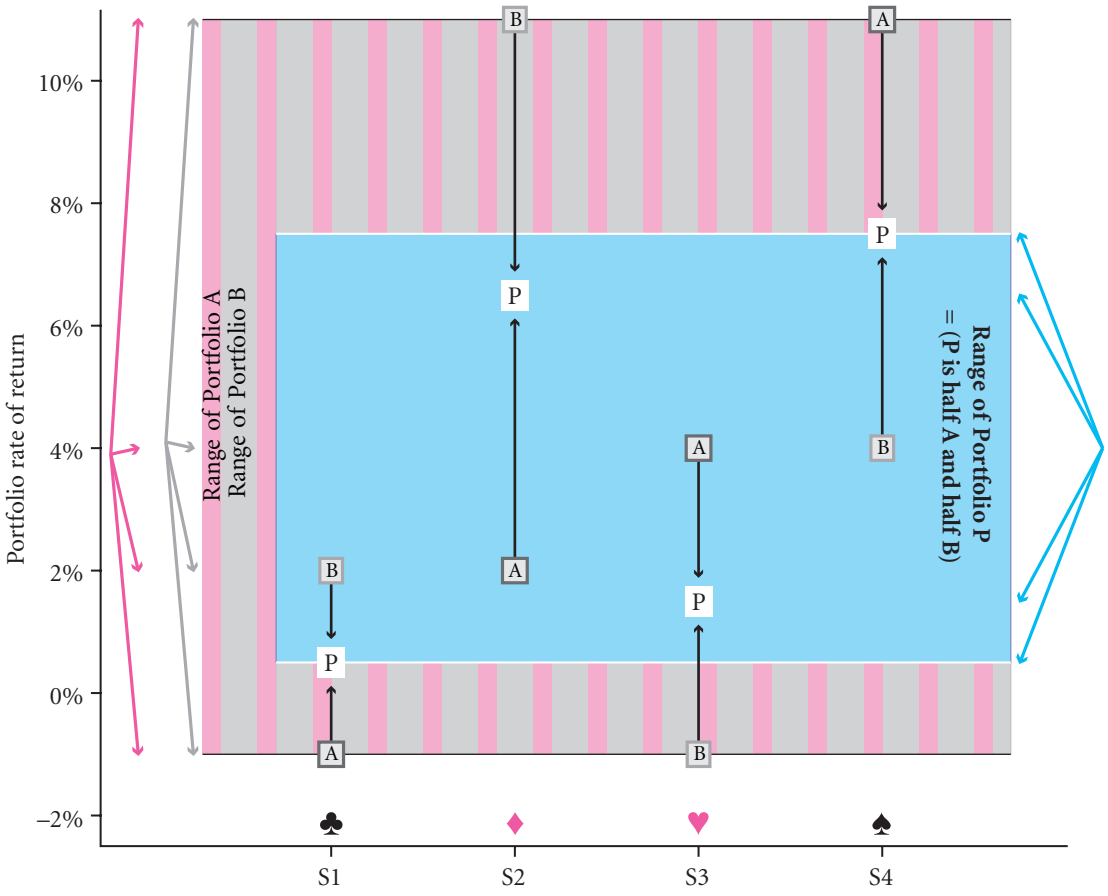
$$= \frac{[r_{S1} - \mathcal{E}(\tilde{r})]^2 + [r_{S2} - \mathcal{E}(\tilde{r})]^2 + [r_{S3} - \mathcal{E}(\tilde{r})]^2 + [r_{S4} - \mathcal{E}(\tilde{r})]^2}{N}$$

$$\Longrightarrow \quad \mathcal{S}dv_{50\% \text{ in A, } 50\% \text{ in B}} = \sqrt{\mathcal{V}ar} = \sqrt{9.25\%\%} \approx 3.04\%$$

An investment in either A or B has a risk of 4.42%. But an investment in half of A and half of B has a risk of only 3.04%! Why?

This is caused by diversification.

The reason is **diversification**—the mixing of different investments within a portfolio that reduces the impact of each one on the overall portfolio performance. More simply put, diversification means that not all of your eggs are in the same basket. If one investment component goes down, the other investment component sometimes



Future	Pfio A	Pfio B	Pfio P = AB
In Scenario S1 ♣	−1.0%	+2.0%	+0.5%
In Scenario S2 ♦	+2.0%	+11.0%	+6.5%
In Scenario S3 ♥	+4.0%	−1.0%	+1.5%
In Scenario S4 ♠	+11.0%	+4.0%	+7.5%
"Reward" ($\mathcal{E}(\tilde{r})$)	4.0%	4.0%	4.0%
"Risk" ($\mathcal{S}dv(\tilde{r})$)	4.42%	4.42%	**3.04%**
"Range" (see figure)	12%	12%	7%

Portfolio P is half A, half B. Because each half-A/half-B point is halfway between A and B, P has lower spread (risk) than either of its components, A and B, by itself. (The risks of A and B were computed as 4.42% in Formula 8.1 on page 204.) Returns on the single-asset portfolios A and B range from −1% to +11%, i.e., 12%. Returns on the combination-asset portfolio P range from +0.5% to +7.5%, i.e., 7%. This is color-coded as blue in the figure. The "spread arrows" on the left and the right also point to the possible outcomes of the three portfolios, indicating the variability visually. The combination portfolio has less spread.

FIGURE 8.2 Rate of Return Outcomes for A, B, and the 50%-50% Combination Portfolio P

happens to go up, or vice versa. The imperfect correlation ("nonsynchronicity") reduces the overall portfolio risk.

8.2A ASSUME INVESTORS CARE ONLY ABOUT RISK AND REWARD

Investors love diversification: the more the better. They could like the market portfolio because it is highly diversified.

This intuition suggests that heavily diversified portfolios—portfolios that invest in many different assets—tend to have lower risk. As a corporate manager, it would be reasonable to assume that your investors are smart. Because diversification helps them reduce their investment risk, you can also reasonably believe that they are indeed holding heavily diversified portfolios. The most heavily diversified portfolio contains a little of every asset. Therefore, we often assume that your investors' portfolio is the overall **market portfolio**, consisting of all available investment opportunities.

If your investors like high reward and low risk and hold the market portfolio, you can work out how your projects affect them.

Why would you want to make any assumptions about your investors' portfolios? The answer is that if you are willing to assume that your investors are holding the market (or something very similar to it), your job as a corporate manager becomes much easier. Instead of asking what each and every one of your investors might possibly like, you can just ask, "When would my investors want to give me their money for investment into my firm's project, given that my investors are currently already holding the broad overall stock market portfolio?" The answer will be as follows:

1. Your investors should like projects that offer more reward (higher expected rates of return).
2. Your investors should like projects that help them diversify away some of the risk in the market portfolio, so that their overall portfolios end up being less risky.

In sum, your corporate managerial task is to take those projects that your investors would like to add to their current (market) portfolios. You should therefore search for projects that have high expected rates of return and high diversification benefits with respect to the market. Let's now turn toward measuring this second characteristic: How can your projects aid your investors' diversification, and how should you measure how good this diversification is?

IMPORTANT:

- Diversification is based on imperfect correlation, or "nonsynchronicity," among investments. It helps smart investors reduce the overall portfolio risk.
- Therefore, as a corporate manager, absent other intelligence, you should believe that your investors tend to hold diversified portfolios. They could even hold portfolios as heavily diversified as the "entire market portfolio."
- As a corporate manager, your task is to think about how a little of your project can aid your investors in terms of its contribution to the risk and reward of their heavily diversified overall portfolios. (You should not think about how risky your project is in itself.)

SOLVE NOW!

Q 8.4 A combination portfolio named AB invests 90% in A and 10% in B.
(a) Compute its risk and reward.
(b) In a bar plot similar to those in Figure 8.1, would this new AB portfolio look less spread out than the P(50%, 50%) portfolio that was worked out in the table in Figure 8.2?

8.3 HOW TO MEASURE RISK CONTRIBUTION

If we are willing to assume that our smart investors are holding all assets in the market, then what projects offer them the best diversification?

8.3A AN ASSET'S OWN RISK IS NOT A GOOD MEASURE FOR ITS RISK CONTRIBUTION TO A PORTFOLIO

Comovement determines risk contribution.

Obviously, diversification does *not* help if two investment opportunities always move in the same direction. For example, if you try to diversify one $50 investment in A with another $50 investment in A (which always has the same outcomes), then your risk does not decrease. On the other hand, if two investment opportunities always move in *opposite* directions, then diversification works extremely well: One is a buffer for the other.

Pretend A is the market, now called M. Is C or D a better addition?

Let's formalize this intuition. For explanation's sake, assume that the stock market portfolio held by your investors is A from Table 8.1, so rename it M (for "market"). Assume that C and D are two projects that your firm could invest in, but you cannot choose both. C offers not only a higher expected rate of return than D (5% versus 2%) but also lower risk (5.15% versus 9.49%). As a manager, acting on behalf of your investors, would you therefore assume that project C is automatically better for them than D?

The combination MC has almost the same risk as M.

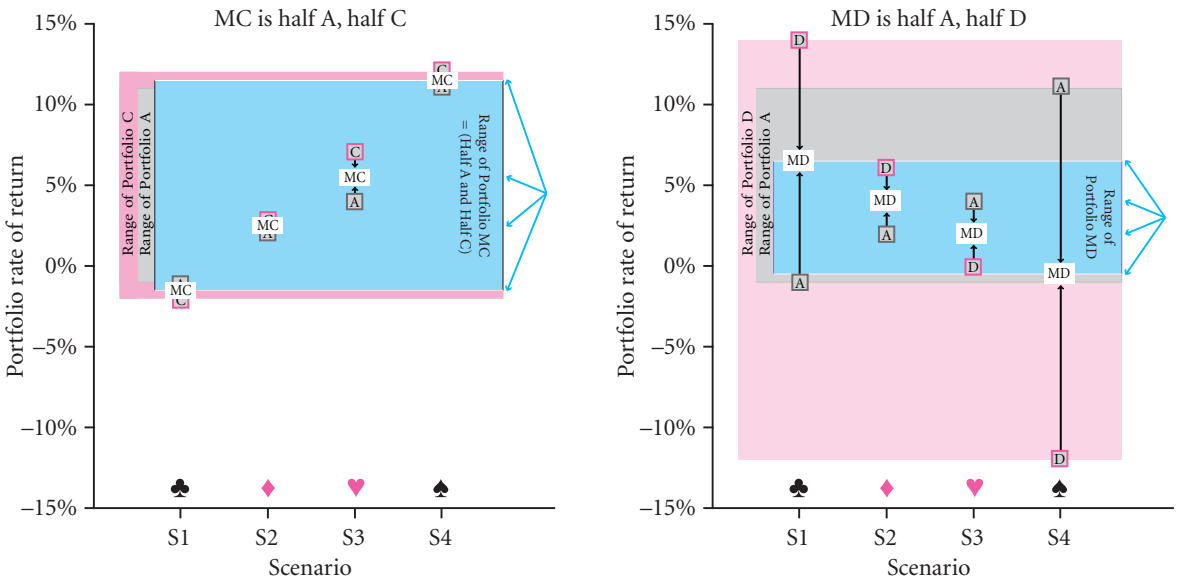
The answer is no. Let's assume that your investors start out with the market portfolio. Figure 8.3 shows what happens if they sell half of their M portfolio to invest in either C or D. You can call these two 50-50 portfolios MC and MD, respectively. Start with MC. If your investors reallocate half their money from M into C, their portfolios would have the following rates of return:

Scenario Outcome	$1 (♣)	S2 (♦)	S3 (♥)	S4 (♠)	Reward	Risk
MC	−1.5%	+2.5%	+5.5%	+11.5%	4.5%	4.74%

The left graph in Figure 8.3 plots the MC rates of return, plus the rates of return for both M and C by themselves. The magenta area is the range of portfolio C, the gray area is the range of portfolio A (i.e., M), and the blue area is the range of the combination portfolio. There is not much change in the risk of your portfolio in moving from a pure M portfolio to the MC portfolio. The risk actually increases from 4.42% to 4.74%.

The combination MD has much lower risk than M.

Now consider the combination of MD, which is the right graph in Figure 8.3. The pink area shows that, by itself, D is a very risky investment. However, if your investors



Portfolio Rates of Return r

Probability	Future	Pfio M (=A)	Pfio C	Pfio D	Pfio MC	Pfio MD
1/4	In Scenario S1 ♣	−1.0%	−2.0%	+14.0%	−1.5%	+6.5%
1/4	In Scenario S2 ♦	+2.0%	+3.0%	+6.0%	+2.5%	+4.0%
1/4	In Scenario S3 ♥	+4.0%	+7.0%	0.0%	+5.5%	+2.0%
1/4	In Scenario S4 ♠	+11.0%	+12.0%	−12.0%	+11.5%	−0.5%
	"Reward" ($\mathcal{E}(\tilde{r})$)	+4.00%	+5.00%	+2.00%	+4.50%	+3.00%
	"Risk" ($\mathcal{S}dv(\tilde{r})$)	4.42%	5.15%	9.49%	4.74%	2.57%
	Range	12%	14%	26%	13%	7%

Although the single-asset portfolio D is much riskier than the single-asset portfolio C, D is much better than C in reducing the risk of the market portfolio. This is because D tends to move opposite to the market. This can also be seen by looking at the color-coded ranges. The magenta range of C in the left figure is smaller than the pink range of D in the right figure. C is simply less risky an investment *in itself* than D. However, the blue range in the MC portfolio on the left is much bigger than the blue range in the MD portfolio on the right. As a portfolio component combined with the market portfolio, D adds much less risk than C. The arrows on the right of each figure also point to the possible outcomes of the combination portfolios, and help indicate their spreads.

FIGURE 8.3 Combining M with either C or D

instead reallocate half of their wealth from M into D, their overall portfolio would have the following rates of return:

Scenario Outcome	S1 (♣)	S2 (♦)	S3 (♥)	S4 (♠)	Reward	Risk
MD	+6.5%	+4.0%	+2.0%	−0.5%	3.0%	2.57%

This is much lower than the range of outcomes on the left (with portfolio C). The MD combination portfolio is simply much safer—even though D by itself is much riskier. If you compare the MC spread with the MD spread, the latter is much smaller. The

algebraic risk measure, the standard deviation, confirms this: Even though D by itself is the riskiest choice, adding it to the M portfolio has reduced your investors' risk from 4.42% to 2.57%. In sum:

Portfolio	Reward	Risk	Note
M (=A) alone	4.00%	4.42%	The market portfolio that your investors were holding.
C alone	5.00%	5.15%	C is less risky than D, if purchased by itself.
D alone	2.00%	9.49%	
MC: half M, half C	4.50%	4.74%	If C is added to M, portfolio risk barely goes down,
MD: half M, half D	3.00%	2.57%	**but if D is added to** M**, portfolio risk goes down a lot!**

You now know that D's own higher standard deviation (9.49%) compared to C's (5.15%) is not a good indication of whether D helps your investors reduce portfolio risk more or less than C. If your investors are primarily holding M, then a very risky project like D can allow them to build lower-risk portfolios. However, if your investors are not holding any assets other than D, they would not care about D's diversification benefits and only about D's own risk. Thus, as a manager, you cannot determine whether your investors would prefer you to invest in C or D unless you know their entire portfolios. (Moreover, it could also depend on how your investors would like you to trade off more overall reward against more overall risk.)

The implication for your project choices as a corporate manager: Everything else equal, D could better reduce portfolio risk for your investors despite its higher own risk.

IMPORTANT: A project's (own) standard deviation is not necessarily a good measure of how it effects the risk of your investors' portfolios. Indeed, it is possible that a project with a very high standard deviation by itself may actually help lower an investor's overall portfolio risk.

SOLVE NOW!

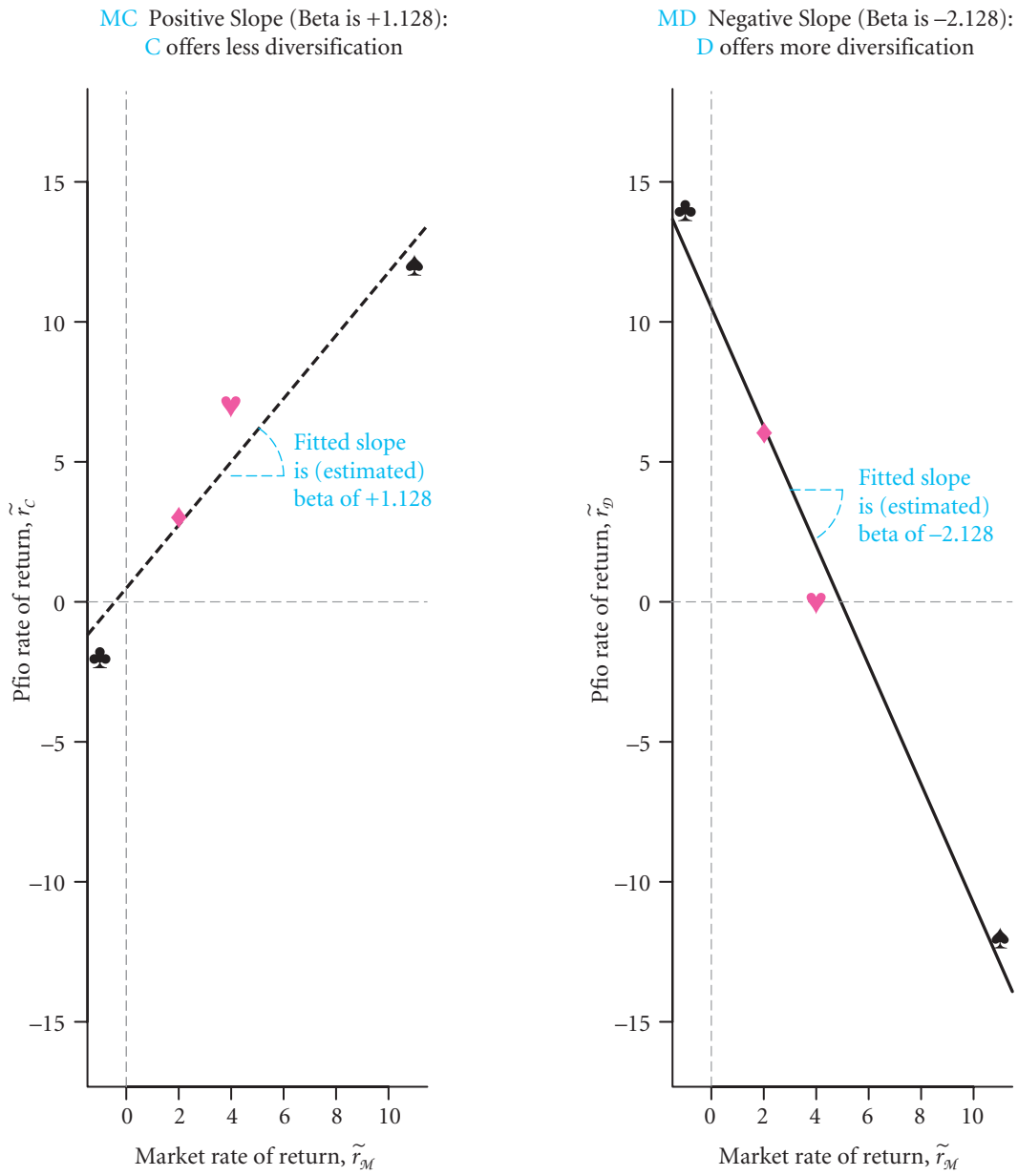
Q 8.5 Confirm the risk and reward calculations for the MC and MD portfolios in the table under Figure 8.3.

8.3B BETA IS A GOOD MEASURE FOR AN ASSET'S RISK CONTRIBUTION TO A PORTFOLIO

Can you guess why portfolio D is so much better than portfolio C in reducing the overall risk when held in combination with the M portfolio? The reason is that D tends to go up when M tends to go down, and vice versa. The same cannot be said for C—it tends to move together with M. You could call this "synchronicity" or "comovement." It is why C does not help investors who are heavily invested in the overall market in their quests to reduce their portfolio risks. Figure 8.4 shows the comovement graphically. If you draw the best-fitting line between M and C, the line slopes up. (It is the same kind of line that you already saw in Section 7.1C.) This means that C tends to be higher when M is higher. If you draw the best-fitting line between M and D, the line slopes down. This means that D tends to be higher when M is lower, and vice versa.

D reduces M's risk because it tends to move in the opposite direction. Comovement can be measured by the slope of a line.

➤ Market beta of PepsiCo, Section 7.1C, p. 185



The four data points in each plot are taken from Table 8.1 on page 202. They are the rates of return on the portfolios M, C, and D. These rates of return are quoted in percentages. In the example, you know that these are the four true possible outcomes. In the real world, if the four points were not the true known outcomes, but just the historical outcomes (sample points), then the slope would not be the true unknown beta, but only the "estimated" beta.

FIGURE 8.4 Possible Outcomes: Rates of Return versus Market Rate of Return

This slope is a common measure of expected comovement or countermovement—how much diversification benefit an investor can obtain from adding a particular new project. A higher slope means more comovement and less diversification; a lower, or even negative, slope means less comovement and more diversification.

The slope of a line is generally called a **beta** because it is common to write the formula for a line as

The line's slope is called beta.

$$y = \alpha + \beta \cdot x \tag{8.2}$$

A beta of 1 is a 45° diagonal line; a beta of 0 is a horizontal line. A positive beta slopes up; a negative beta slopes down; and a beta of infinity is a vertical line. The particular line we care about in finance is the **market beta**. It matters if you can posit that your smart investors are primarily holding the market portfolio. If so, you want to know how the rate of return on your own project comoves with that of the market. This is exactly what the market beta of your project tells you. To find it, draw the rate of return on M on the x-axis (hence the prefix "market" in market beta) and the rate of return on your project (here, either C or D) on the y-axis. Then take a ruler and try to draw the best line between the four points. You will find that the market beta of C is positive (your best line is upward sloping), whereas the market beta of D is negative (your best line is downward sloping). In statistics, you should have learned that you can find the beta by running a linear regression. If you don't remember, no worries: In Section 8.3B, I will show you again how to compute them exactly. For now, take my word that the two best lines are

We want the beta of the asset's rate of return on the y-axis and the market's rate of return on the x-axis.

$$\begin{aligned}
\tilde{r}_{\mathrm{C}} &\approx 0.49\% + (+1.128) \cdot \tilde{r}_{\mathrm{M}} \\
\tilde{r}_{\mathrm{D}} &\approx 10.51\% + (-2.128) \cdot \tilde{r}_{\mathrm{M}} \\
\tilde{r}_{\mathrm{i}} &= \alpha_{\mathrm{i,M}} + \beta_{\mathrm{i,M}} \cdot \tilde{r}_{\mathrm{M}}
\end{aligned} \tag{8.3}$$

The subscripts on the betas remind you what the variables on the x-axis and the y-axis are. The first subscript is on the y-axis, the second is on the x-axis, so $\beta_{\mathrm{C,M}} \approx 1.128$ and $\beta_{\mathrm{D,M}} \approx -2.128$. In fact, market beta plays such an important role in finance that the name "beta" has itself become synonymous for "market beta," and the second subscript is usually omitted. Formula 8.3 is sometimes called the **market model**.

IMPORTANT:

- Diversification works better if the new investment project tends to move in the opposite direction of the rest of the portfolio than if it tends to move in the same direction.
- It is often reasonable to assume that smart investors are already holding the market portfolio and are now considering investing into just a little of one additional asset—your firm's new project.
- If this new investment asset has a negative beta with respect to the market (its "market beta"), it means that it tends to go down when the market goes up, and vice versa. If this new investment asset has a positive beta with respect to the market, it means that it tends to move together with the market. If this new investment asset has a zero beta with respect to the market, it means that it moves independently of the market for all practical purposes.

- The market beta is a good measure of an investment asset's risk contribution for an investor who holds the market portfolio. The lower (or negative) the market beta, the more this investment helps reduce your investor's risk.
- The market beta of an asset can be interpreted as a line slope, where the rate of return on the market is on the *x*-axis and the rate of return on the new asset is on the *y*-axis. The line states how you expect the new asset to perform as a function of how the market will perform.
- You can think of market beta as a measure for "toxicity." In a reasonable equilibrium, holding everything else constant, risk-averse investors who are holding the market portfolio would agree to pay more for assets that have lower market betas. They would pay less for assets with higher market betas.

Warning: All of this beta-related risk measuring is interesting only if your investors are holding (portfolios close to) the overall market.

Before we conclude, some caveats are in order. From your perspective as the manager of a company, perhaps a publicly traded one, it is reasonable to assume that your investors are holding the market portfolio. It is also reasonable to assume that your new project is just a tiny new additional component of your investors' overall portfolios. We will staunchly maintain these assumptions, but you should be aware that they may not always be appropriate. If your investors are *not* holding something close to the market portfolio, then your project's market beta would *not* be a good measure of your projects' risk contributions. In the extreme, if your investors are holding *only* your project, market beta would not measure the project's risk contribution at all. This is often the case for entrepreneurs. They often have no choice but to put all their money into one egg in one basket. Such investors care only about the project's standard deviation, not the project's market beta.

Alpha has meaning, too, even though you won't use it just yet.

Although we shall not use it further in this book, the alpha intercept in Formula 8.3 also plays an important role. Together, alpha and beta help determine how attractive an investment is. For example, if the rate of return on the market is 10%, Formula 8.3 tells you that you would expect the rate of return on D to be

$$\mathcal{E}\left[(\tilde{r}_{\mathrm{D}})| \text{ if } \tilde{r}_{\mathrm{M}} = 10\%\right] \approx 10.51\% + (-2.128) \cdot 10\% \approx -10.77\%$$

The higher the alpha, the better the average performance of your investment given any particular rate of return on the market. Just as investment professionals often call the market beta just beta, they often call this specific intercept (here 10.51%) just alpha. (There is one small complication: They usually subtract the risk-free interest rate first from both $\tilde{r}_{\mathrm{D}}$ and $\tilde{r}_{\mathrm{M}}$ in their regressions.)

COMPUTING MARKET BETAS FROM HISTORICAL RATES OF RETURN

You can compute the best-fit beta via a 4-step procedure.

➤ Table 8.1, p. 202

Now that you understand what beta means, how can you actually compute it? Let me show you. Let's return to the assets in Table 8.1. What is the market beta of asset C? I have already told you that this slope is +1.128. To calculate it, I followed a tedious, but not mysterious, recipe. Here is what you have to do:

1. Just as you did for your variance calculations, first translate all returns into deviations from the mean. That is, for M, C, and D, subtract their own means from every realization.

First, de-mean each rate of return.
➤ Variance calculations, Section 6.1B, p. 141

	Original Rates of Return			Net-of-Mean Rates of Return		
Future	Pfio M	Pfio C	Pfio D	Pfio M	Pfio C	Pfio D
In Scenario S1 ♣	−1.0%	−2.0%	+14.0%	−5.0%	−7.0%	+12.0%
In Scenario S2 ♦	+2.0%	+3.0%	+6.0%	−2.0%	−2.0%	+4.0%
In Scenario S3 ♥	+4.0%	+7.0%	0.0%	0.0%	+2.0%	−2.0%
In Scenario S4 ♠	+11.0%	+12.0%	−12.0%	+7.0%	+7.0%	−14.0%
"Reward" ($\mathcal{E}(\tilde{r})$)	4.00%	5.00%	2.00%	0.00%	0.00%	0.00%

(How demeaning!)

2. Compute the variance of the series on the x-axis. This is the variance of the rates of return on M. With the net-of-mean M returns, this is easy:

Take squares and then average. This is the variance.

$$\mathcal{V}ar(\tilde{r}_M) = \frac{(-5\%)^2 + (-2\%)^2 + (0)^2 + (7\%)^2}{4} = 19.5\%\%$$

$$= \frac{\text{Sum over All Scenarios } S\text{: } [\tilde{r}_M \text{ in Scenario } S - \text{Average } \mathcal{E}(\tilde{r}_M)]^2}{N}$$

Because 1% is "multiply by 0.01," 19.5%% could be rewritten as 0.195% or 0.00195. (Note also that you do not need to compute the variances of either C or D to obtain their market betas.)

3. Compute the average product of the net-of-mean variables. In this case, you want to compute the market beta for C, so you work with the rates of return on M and C.

For covariances, multiply net-of-mean returns, then average.

$$\mathcal{C}ov(\tilde{r}_M, \tilde{r}_C) = \frac{(-5\%)\cdot(-7\%) + (-2\%)\cdot(-2\%) + (0)\cdot(2\%) + (7\%)\cdot(7\%)}{4}$$

$$= 22\%\% = 0.22\% \qquad (8.4)$$

$$= \frac{\text{Sum over All Scenarios } S\text{: } [\tilde{r}_M \text{ in Scenario } S - \mathcal{E}(\tilde{r}_M)]\cdot[\tilde{r}_C \text{ in Scenario } S - \mathcal{E}(\tilde{r}_C)]}{N}$$

This statistic is called the **covariance** between the rates of return on M and C.

4. The beta of C with respect to the market M, formally $\beta_{C,M}$ but often abbreviated as β_C, is the ratio of these two quantities,

The beta is the covariance divided by the variance.

$$\beta_C = \beta_{C,M} = \frac{22\%\%}{19.5\%\%} \approx 1.128 \qquad (8.5)$$

$$= \frac{\mathcal{C}ov(\tilde{r}_M, \tilde{r}_C)}{\mathcal{V}ar(\tilde{r}_M)}$$

You can confirm our calculations using a spreadsheet.

This slope of 1.128 (a little more than a perfect 45° diagonal) is exactly the market beta we drew in Figure 8.4. Many spreadsheets and all statistical programs can compute it for you: They call the routine that does this a **linear regression**.

Think of market beta as the characteristic of an asset.

You should always think of the beta of an asset i with respect to a portfolio P as a characteristic measure of your asset i relative to an underlying base portfolio P. The rate of return on P is on the x-axis; the rate of return on i is on the y-axis. As we stated earlier, most often—but not always—the portfolio P is the market portfolio, M, so $\beta_{i,M}$ is often just called the market beta, or even just the beta (and the second subscript is omitted).

The average beta of the market (all stocks) is 1, not 0.

Now think for a moment. What should the average beta of a stock in the economy be? Equivalently, what is the beta of the market portfolio itself? Replace the C in Formula 8.5 with M:

$$\beta_M = \frac{\mathcal{C}ov(\tilde{r}_M, \tilde{r}_M)}{\mathcal{V}ar(\tilde{r}_M)}$$

If you look at the definition of covariance, you can see that the covariance of a variable with itself *is* the variance. (The covariance is a generalization of the variance concept from one variable to two variables.) Therefore, $\mathcal{C}ov(\tilde{r}_M, \tilde{r}_M) = \mathcal{V}ar(\tilde{r}_M)$, and the market beta of the market itself is 1. Graphically, if both the x-axis and the y-axis are graphing the same values, every point must lie on the diagonal. Economically, this should not be surprising, either: the market goes up one-to-one with the market.

Why torture you with computations? So you can play with scenarios.

Now that you know how to compute betas and covariances, you can consider scenarios for your project. For example, you might have a new project for which you would guess that it will have a rate of return of −5% if the market returns −10%; a rate of return of +5% if the market returns +5%; and a rate of return of 30% if the market returns 10%. Knowing how to compute a market beta therefore makes it useful to think of such scenarios. (You can also use this technique to explore the relationship between your projects and some other factors. For example, you could determine how your projects covary with the price of oil to learn about your project's oil risk exposure.)

➤ An oil-price beta, Section 9.8A, p. 292

Practical advice to help you estimate market beta in the real world: Use 3–5 years of daily observations and then adjust.

In the real world, you will sometimes think in terms of such scenarios, but more often you have to compute a market beta from historical rates of return for the overall stock market and for your project (or similar projects). Fortunately, as we noted up front, the beta computations themselves are exactly the same. In effect, when you use historical data, you simply assume that each time period was one representative scenario and proceed from there. Nevertheless, there are some real-world complications you should think about:

1. Should you use daily, weekly, monthly, or annual rates of return? The answer is that the best market beta estimates come from daily or weekly data. Annual data should be avoided (except in a textbook in which space is limited). Monthly data can be used if need be.
2. How much data should you use? Most researchers tend to use 3–5 years of historical rate of return data. This reflects a trade-off between having enough data and not going too far back into ancient history, which may be less relevant. If you have daily data, 3 years works quite well.

3. Is the historical beta a good estimate of the future beta? It turns out that history can sometimes be deceptive, especially if your estimated historical beta is far away from the market's beta average of 1. Fortunately, there are at least two methods to help adjust historical betas so that you get better estimates of future betas:
   (A) **Averaging:** You could rely not just on the historical beta computed from your own project's returns. Instead, you could use the average historical betas for many other projects that are similar to your own (for example, projects from the same industry or in the same size class). Such averages are usually less noisy.
   (B) **Shrinking:** You could "shrink" your historical beta toward the overall market beta of 1. For example, in the simplest such shrinker, you would simply compute an average of the overall market beta of 1 and your historical market beta estimate. If you computed a historical market beta of 4 for your project, you would work with a prediction of future market beta of about $(4 + 1)/2 = 2.5$ for your project.

   Many smart executives start with a statistical beta estimated from historical data and then use their intuitive judgment to adjust it.

SOLVE NOW!

Q 8.6 Return to your computation of market beta of 1.128 in Formula 8.5. We called it $\beta_{C,M}$, or β_C for short. Is the order of the subscripts important? That is, please compute $\beta_{M,C}$ and see whether it is also 1.128.

8.3C WHY NOT CORRELATION OR COVARIANCE?

Covariance and beta (and correlation) always have the same sign.

There is a close family relationship between covariance, beta, and correlation. The beta is the covariance divided by one of the variances. The correlation is the covariance divided by both standard deviations. The denominators are always positive. Thus, if the covariance is positive, so are the beta and the correlation; if the covariance is negative, so are the beta and the correlation; and if the covariance is zero, so are the beta and the correlation. The nice thing about the correlation, which makes it useful in many contexts outside finance, is that it has no scale and is always between −100% and +100%:

- Two variables that always move perfectly in the same direction have a correlation of 100%.
- Two variables that always move perfectly in opposite directions have a correlation of −100%.
- Two variables that are independent have a correlation of 0%.

This makes correlations very easy to interpret. The not-so-nice thing about correlation is that it has no scale and is always between −100% and +100%. This means that two investments, the second being a million times bigger than the first (all project rates of return multiplied by a million), have the same correlation with the stock market. Yet, the second investment would go up or down with any slight tremor in the market by a million times more, which would of course mean that it would contribute

TABLE 8.2 SOME MARKET BETAS AND CAPITALIZATIONS ON MAY 10, 2008

Company	Ticker	Mkt-Cap[a]	Market Beta Yahoo	Market Beta AOL	Company	Ticker	Mkt-Cap[a]	Market Beta Yahoo	Market Beta AOL
AMD	AMD	4	1.96	2.67	Intel	INTC	124	1.73	1.85
Coca-Cola	KO	130	0.52	0.78	PepsiCo	PEP	107	0.24	0.22
Citigroup	C	124	1.71	1.31	J.P. Morgan	JPM	158	0.85	0.91
Goldman Sachs	GS	74	2.24	1.84	Morgan Stanley	MS	51	1.75	1.71
IBM	IBM	170	0.95	0.94	Hewlett-Packard	HPQ	121	1.09	1.46
Dell	DELL	39	1.53	1.36	Sun	JAVA	10	1.01	2.13
Apple Inc	AAPL	162	2.86	2.57	Sony	SNE	45	0.97	1.08
Google	GOOG	180	2.60	2.17	Yahoo	YHOO	36	0.39	1.03
Ford	F	16	2.11	2.13	General Motors	GM	12	1.50	1.64
American Airlines	AMR	2	1.71	2.69	Southwest	LUV	9.5	−0.12	0.13
Exxon Mobil	XOM	469	1.14	1.04	Barrick Gold	ABX	34	−0.20	0.49
Philip Morris	PM	109	0.00	NA	Procter & Gamble	PG	199	0.63	0.57
Textron	TXT	15	1.46	1.81	Boeing	BA	63	1.08	1.22

a. "Mkt-Cap" is the equity market value in billions of dollars. Yahoo explained its betas as follows:

> The Beta is beta of equity. Beta is the monthly price change of a particular company relative to the monthly price change of the S&P 500. The time period for Beta is 5 years when available, and not less than 2.5 years. This value is updated monthly.

Note that Yahoo! *Finance* seems to ignore dividends, but this usually makes little difference. I could not find an explanation for the market betas provided by AOL. Google's market betas were the same as AOL's, but there was no explanation for them, either.

much more risk. The correlation ignores this, which disqualifies it as a serious candidate for a project risk measure. Fortunately, beta takes care of scale—indeed, the beta for the second project would be a million times larger. This is why we prefer beta over correlation as a measure of risk contribution to a portfolio.

8.3D INTERPRETING TYPICAL STOCK MARKET BETAS

Market beta works well when investors are holding the market and adding only a little of your project.

The market beta is the best measure of "diversification help" for an investor who holds the stock market portfolio and considers adding *just a little* of your firm's project. From your perspective as a manager seeking to attract investors, this is not a perfect, necessarily true assumption—but it is a reasonable one. Recall that we assume that investors are smart, so presumably they are holding highly diversified portfolios. To convince your market investors to like your $10 million project, you just need the average investor to want to buy $10 million divided by about $20 trillion (the stock market capitalization), which is 1/2,000,000 of their portfolios. For your investors, your corporate projects are just tiny additions to their market portfolios.

Most financial websites publish market beta estimates.

You can look up the market betas of publicly traded stocks on many financial websites. Table 8.2 lists the betas of some randomly chosen companies in May 2008 from Yahoo! *Finance* and from AOL's finance site. Most company betas are in the

range of around 0 to about 2.5. A beta above 1 is considered risk-increasing for an investor holding the overall stock market (it is riskier than the stock market itself), while a beta below 1 is considered risk-reducing. Betas that are negative are quite rare—in Table 8.2, there happen to be only two such stocks, Southwest and Barrick Gold, according to the Yahoo betas, and none according to the AOL betas.

Beta can be viewed as the marginal change of your project with respect to the market.

Market beta has yet another nice intuitive interpretation: It is the degree to which the firm's value tends to change if the stock market changes. For example, Apple's market beta of approximately 2.7 (somewhere between 2.57 and 2.86) says that if the stock market will return an extra 5% next year (above and beyond its expectations), Apple will return an extra $2.7 \cdot 5\% = 13.5\%$ (above and beyond Apple's expectations). Of course, market beta is not a measure of how good an investment Apple is. (This measure is the alpha [which can be interpreted as an expected rate of return]. In the next section, you will learn a model that relates market beta to the expected rate of return. For now let's assume for illustration that there is no reasonable relationship.) Let's make the absurd assumption that Apple's expected rate of return is -30% and the more reasonable assumption that the market's expected rate of return is 10%. All that Apple's market beta then tells you is that when the market does 1% better than expected (i.e., $(10\% + 1\% = 11\%)$), then Apple would do 2.7% better than expected (i.e., $(-30\% + 2.7 \cdot 1\% = -27.3\%)$). If the market does 0% (i.e., 10% worse than expected), Apple would be expected to do -57% (i.e., 27% worse than expected). And so on. Apple's high market beta is useful because it informs you that if you hold the stock market, adding Apple stock would not help you diversify your market risk very much. Holding Apple stock would amplify any market swings, not reduce them. In any case, Apple's market beta does not tell you whether Apple is priced so high that it is an investment with a negative expected rate of return that you should avoid in the first place.

SOLVE NOW!

Q 8.7 You estimate your project x to return -5% if the stock market returns -10%, and $+5\%$ if the stock market returns $+10\%$. What would you use as the market beta estimate for your project?

Q 8.8 You estimate your project to return $+5\%$ if the stock market returns -10%, and -5% if the stock market returns $+10\%$. What would you use as the market beta estimate for your project?

8.4 EXPECTED RATES OF RETURN AND MARKET BETAS FOR (WEIGHTED) PORTFOLIOS AND FIRMS

Portfolios consist of multiple assets (or of other portfolios). Value-weighted and equal-weighted portfolios are defined.

Let's go back to your managerial perspective of figuring out the risk and return of your corporate projects. Many small projects are bundled together, so it is very common for managers to consider multiple projects already packaged together as one portfolio. For example, you can think of your firm as a collection of divisions that have been packaged together. If division C is worth \$1 million and division D is worth \$2 million, then a firm consisting of C and D is worth \$3 million. C constitutes 1/3 of the portfolio

"Firm" and D constitutes 2/3 of the portfolio "Firm." This kind of portfolio is called a **value-weighted portfolio** because the weights correspond to the market values of the components. (A portfolio that invests \$100 in C and \$200 in D would also be value-weighted. A portfolio that invests equal amounts in the constituents (for example, \$500 in each) is called an **equal-weighted portfolio**.)

What are the expected rate of return and market beta of a portfolio?

Thus, as a manager, you have to know how to work with a portfolio (firm) when you have all the information about all of its underlying component stocks (projects). If I tell you what the expected rate of return on each project is, and what the market beta of each project is, can you tell me what the firm's overall expected rate of return and overall market beta are? Let's try it. Use the C and D stocks from Table 8.1 on page 202, and call CDD the portfolio (or firm) that consists of 1/3 investment in division C and 2/3 investment in division D.

You can average *actual* rates of return.

Actually you already know that individual portfolio rates of return can be averaged. For example, in scenario S4 (♠), investment C has a rate of return of +12%, and investment D has a rate of return of −12%. Consequently, the overall investment CDD has a rate of return of

$$r_{\text{CDD, in S4 (♠)}} = 1/3 \cdot (+12\%) + 2/3 \cdot (-12\%) = -4\%$$
$$= w_{\text{C}} \cdot r_{\text{C, in S4}} + w_{\text{D}} \cdot r_{\text{D, in S4}}$$

Let us verify this: Put \$100 into C and \$200 into D. C turns into $1.12 \cdot \$100 = \$112$. D turns into $(1 - 12\%) \cdot \$200 = \$176$. The total portfolio turns into \$288, which is a rate of return of $\$288/\$300 - 1 = -4\%$ on a \$300 investment.

You can average *expected* rates of return.

It is also intuitive that *expected* rates of return can be averaged. In our example, C has an *expected* rate of return of 5%, and D has an *expected* rate of return of 2%. Consequently, your overall firm CDD has an expected rate of return of

$$\mathcal{E}(\tilde{r}_{\text{CDD}}) = 1/3 \cdot (+5\%) + 2/3 \cdot (+2\%) = 3\%$$
$$= w_{\text{C}} \cdot \mathcal{E}(\tilde{r}_{\text{C}}) + w_{\text{D}} \cdot \mathcal{E}(\tilde{r}_{\text{D}})$$

Let us verify this, too. There are four possible outcomes: In S1, your actual rate of return is 8.67%; in S2, it is 5%; in S3, it is 2.33%; and in S4, it is −4%. The average of these four outcomes is indeed 3%.

News flash: You can also average market betas.

But here is a remarkable and less intuitive fact: Market betas—that is, the projects' risk contributions to your investors' market portfolios—can be averaged, too. That is, I claim that the beta of CDD is the weighted average of the betas of C and D. You already computed the latter in Formula 8.3 as +1.128 and −2.128, respectively. Their value-weighted average is

➤ Market betas of C and D, Formula 8.3, p. 213

$$\beta_{\text{CDD}} = 1/3 \cdot (+1.128) + 2/3 \cdot (-2.128) \approx -1.043 \qquad (8.6)$$
$$= w_{\text{C}} \cdot \beta_{\text{C}} + w_{\text{D}} \cdot \beta_{\text{D}}$$

(*But* you cannot average variances or standard deviations!)

You will be asked in Q 8.9 to confirm this. However, do not think for a moment that you can compute value-weighted averages for all statistics. For example, variances and standard deviations cannot be averaged.

IMPORTANT:

- You can think of the firm as a weighted investment portfolio of components, such as individual divisions or projects. For example, if a firm named ab consists only of two divisions, a and b, then its rate of return is always

$$\tilde{r}_{ab} = w_a \cdot \tilde{r}_a + w_b \cdot \tilde{r}_b$$

  where the weights are the relative values of the two divisions. (You can also think of this one firm as a "subportfolio" within a larger overall portfolio, such as the market portfolio.)

- The expected rate of return ("reward") of a portfolio is the weighted average expected rate of return of its components,

$$\mathcal{E}(\tilde{r}_{ab}) = w_a \cdot \mathcal{E}(\tilde{r}_a) + w_b \cdot \mathcal{E}(\tilde{r}_b)$$

  Therefore, the expected rate of return of a firm is the weighted average rate of return of its divisions.

- Like expected rates of return, betas can be weighted and averaged. The beta of a firm—i.e., the firm's "risk contribution" to the overall market portfolio—is the weighted average of the betas of its components,

$$\beta_{ab} = w_a \cdot \beta_a + w_b \cdot \beta_b$$

  Therefore, the market beta of a firm is the weighted average market beta of its divisions.

- You cannot do analogous weighted averaging with variances or standard deviations.

A firm is a portfolio of debt and equity. Thus, the portfolio formulas apply to the firm (with debt and equity as its components), too!

You can think of the firm not only as consisting of divisions, but also as consisting of debt and equity. For example, say your \$400 million firm is financed with debt worth \$100 million and equity worth \$300 million. If you own all debt and equity, you own the firm. What is the market beta of your firm's assets? Well, the beta of your overall firm must be the weighted average beta of its debt and equity. If your \$100 million in debt has a market beta of, say, 0.4 and your \$300 million of equity has a market beta of, say, 2.0, then your firm has a market beta of

$$1/4 \cdot (0.4) \quad + \quad 3/4 \cdot (2.0) \quad = 1.6$$

$$\beta_{\text{Firm}} = \left(\frac{\text{Debt value}}{\text{Firm value}}\right) \cdot \beta_{\text{Debt}} + \left(\frac{\text{Equity value}}{\text{Firm value}}\right) \cdot \beta_{\text{Equity}} \tag{8.7}$$

This 1.6 is called the **asset beta** to distinguish it from the **equity beta** of 2.0 that financial websites report. Put differently, if your firm refinances itself to 100% equity (i.e., \$400 million worth), then the reported market beta of your equity on Yahoo!

Finance would fall to 1.6. The asset beta is the measure of your firm's projects' risk contribution to the portfolio of your investors. It is the relevant measure that will determine the cost of capital that you should use as the hurdle rate for projects that are like the average project in your firm.

SOLVE NOW!

Q 8.9 Let's check that the beta combination formula (Formula 8.6 on page 220) is correct. Let me lead you along:

(a) Write down a table with the rate of return on the market and on portfolio CDD in each of the four possible states. (Hint: In scenario S1 [♣], the rate of return on CDD is 8.67%.) Then forget about C and D altogether. (In this question, you will work only with the market and CDD.)

(b) Compute the average rate of return on the market and on CDD.

(c) Write down a table with the de-meaned market rate of return and CDD rate of return in each of the four possible states. (The mean of the de-meaned returns must be zero.)

(d) Multiply the de-meaned rates of return in each scenario. This gives you four cross-products, each having units of %%. (Hint: In scenario S1 [♣], it is about −28.35%%.)

(e) Compute the average of these cross-products. This is the covariance between CDD and M.

(f) Divide the covariance between CDD and M by the variance of the market. Is it equal to the −1.04 from Formula 8.6?

(g) Which is faster—this route or Formula 8.6?

Q 8.10 Let's confirm that you cannot take a value-weighted average of component variances (and thus of standard deviations) the same way that you can take value-weighted average expected rates of return and value-weighted average market betas.

(a) What would the value-weighted average variance of CDD be?

(b) What is the actual variance of CDD?

Q 8.11 Consider an investment of 2/3 in C and 1/3 in D. Call this new portfolio CCD. Compute the variance, standard deviation, and market beta of CCD. Do this two ways: first from the four individual scenario rates of return of CCD, and then from the statistical properties of C and D itself.

Q 8.12 Assume that a firm will always have enough money to pay off its bonds, so the beta of its bonds is 0. (Being risk free, the rate of return on the bonds is obviously independent of the rate of return on the stock market.) Assume that the beta of the underlying assets is 2. What would financial websites report for the beta of the firm's equity if it changes its current capital structure from all equity to half debt and half equity? To 90% debt and 10% equity?

8.5 SPREADSHEET CALCULATIONS FOR RISK AND REWARD

In real life, you can do calculations faster with a spreadsheet.

Doing all these calculations by hand is tedious. We computed these statistics within the context of just four scenarios, so that you would understand their meanings better. However, you can do this faster in the real world. Usually, you would download reams of real historical rates of return data into a computer spreadsheet, like Excel or OpenOffice. Spreadsheets have all the functionality you need already built in—and you now understand what their functions actually calculate. In practice, you would use the following functions:

- **average**(*range*) computes the average (rate of return).
- **varp**(*range*) computes the (population) variance. If you worked with historical data instead of known scenarios, you would instead use the **var**(*range*) formula. (The latter divides by $N - 1$ rather than by N, which I will explain in a moment.)
- **stdevp**(*range*) computes the (population) standard deviation. If you used historical data instead of known scenarios, you would instead use the **stdev**(*range*) formula.
- **covar**(*range-1,range-2*) computes the population covariance between two series. If Excel was consistent, this function should be called covarp rather than covar.
- **correl**(*range-1,range-2*) computes the correlation between two series.
- **slope**(*range-Y,range-X*) computes a beta. If *range-Y* contains the rates of return of an investment and *range-X* contains the rates of return on the market, then this formula computes the market beta.

Table 8.3 shows a computer spreadsheet that computes everything that you did in this chapter.

8.5A STATISTICAL NUANCES

➤ Will history repeat itself?, Section 7.1E, p. 189

In this chapter, we have continued to presume (just as we did in Section 7.1E) that historical data gives us an unbiased guide to the future when it comes to means, variances, covariances, and betas. Of course, this is a simplification—and remember that it can be a problematic one. I already noted that this is less of a problem for covariances, variances, and betas than it is for means. Rely on historical means as predictors of future expected rates of return only at your own risk!

When working with a *sample*, the (co)variance formula divides by $N-1$. When working with the *population*, the (co)variance formula divides by N.

There is a second, smaller statistical issue that you should be aware of. Statisticians often use a covariance formula that divides by $N - 1$, not N. Strictly speaking, dividing by $N - 1$ is appropriate if you work with historical data. These are just sample draws and not the full population of possible outcomes. With a sample, you do not really know the true mean when you de-mean your observations. The division by a smaller number, $N - 1$, gives a larger but unbiased covariance estimate. It is also often called the *sample covariance*. In contrast, dividing by N is appropriate if you work with "scenarios" that you know to be true and equally likely. In this case, the statistic is often called the *population covariance*. The difference rarely matters in finance, where you usually have a lot of observations—except in our book examples where you have only four scenarios. (For example, dividing by $N = 1{,}000$ and by $N = 1{,}001$ gives almost the same number.)

TABLE 8.3 THE COMPUTER SPREADSHEET

	A	B	C	D	E	F	G	H	I	J	K	L	M	Formula
1	Investor Choice, Sample Spreadsheet								Combinations with A (M)					
2									MB	MC	MD			
3		Base Portfolios							w_B	w_C	w_D		w_{CDD}	
4		A (M)	B	C	D		F		=1/2	=1/2	=1/2		=(1/3, 2/3)	
5	**S1**	–1.0%	2.0%	–2.0%	14.0%		1.0%		=0.5%	=–1.5%	=6.5%		=8.7%	
6	**S2**	2.0%	11.0%	3.0%	6.0%		1.0%		=6.5%	=2.5%	=4.0%		=5.0%	
7	**S3**	4.0%	–1.0%	7.0%	0.0%		1.0%		=1.5%	=5.5%	=2.0%		=2.3%	
8	**S4**	11.0%	4.0%	12.0%	–12.0%		1.0%		=7.5%	=11.5%	=–0.5%		=–4.0%	
9														
10														
11	**Average**	=4.0%	=4.0%	=5.0%	=2.0%		=1.0%		=4.0%	=4.5%	=3.0%		=3.0%	←=average(×5:×8)
12	**Variance**	=0.1950%	=0.1950%	=0.2650%	=0.9000%		=0.0000%		=0.093%	=0.225%	=0.066%		=0.214%	←=varp(×5:×8)
13	**Risk**	=4.42%	=4.42%	=5.15%	=9.49%		=0.000%		=3.041%	=4.743%	=2.574%		=4.625%	←=stdevp(×5:×8)
14														
15														
16	**Market Beta**	=1.000	=–0.051	=1.128	=–2.128		=0.000		=0.474	=1.064	=–0.564		=–1.043	←=slope(×5:×8,B5:B8)
17														
18	**Alpha**	=0.00%	=4.21%	=0.49%	=10.51%		=1.00%		=2.10%	=0.24%	=5.26%		=7.17%	←=intercept(×5:×8,B5:B8)
19	**Correlation**	=100.0%	=–5.1%	=96.8%	=–99.1%		=#DIV/0!		=68.9%	=99.1%	=–96.8%		=–99.6%	←=correl(×5:×8,B5:B8)
20	**Covariance**	=0.20%	=–0.01%	=0.22%	=–0.42%		=0.00%		=0.09%	=0.21%	=–0.11%		=–0.20%	←=covar(×5:×8,B5:B8)

This spreadsheet (also available on the book website) demonstrates the main statistical calculations that are performed in this chapter. Please note that we are using the population variance and population standard deviation formulas, not the sample variance and sample standard deviation formulas. Spreadsheet cells that are formulas contain an '='. In rows 5–8, columns I–K are equal combinations of M (column B) and one other portfolio (B–D), which are in columns C–E, respectively. Column M is a weighted average of columns J and K. Formulas in rows 11–20 are given on the right side.

The only reason why you even needed to know this is that if you use a program that has a built-in variance or standard deviation function, you should not be surprised if you get numbers different from those you have computed in this chapter. In some programs, you can get both functions. In Excel, you can use the *varp* and *stdevp* population statistical functions to get the population statistics, not the *var* and *stdev* functions that would give you the sample statistics.

This is important to keep in mind if you use a spreadsheet to check your work.

Beta is not affected by whether you divide the variance/covariance by N or $N - 1$, because both numerator (covariance) and denominator (variance) are divided by the same number.

For market beta, the divisor cancels out and does not matter.

Furthermore, statisticians distinguish between underlying unknown statistics and statistics estimated from the data. For example, they might call the unknown true mean μ and the sample mean m (or $\bar{x}$). They might call the unknown true beta β^T and the estimated sample beta a beta with a little hat ($\hat{\beta}$). And so on. Our book is casual about the difference for lack of space, but keep in mind that whenever you work with historical data, you are really just working with sample estimates.

My fault: Our notation should have distinguished between true population and estimated sample statistics.

SUMMARY

This chapter covered the following major points:

- The expected rate of return is a measure of expected reward.

$$\mathcal{E}(\tilde{r}_P) = \frac{\text{Sum over All Scenarios: } \left[\text{Return of Pfio P in Each Scenario}\right]}{N}$$

- The variance is (roughly) the average squared deviation from the mean.

$$\mathcal{V}ar(\tilde{r}_P) = \frac{\text{Sum over All Scenarios: } \left[\text{Return of Pfio P in Each Scenario} - \mathcal{E}(\tilde{r}_P)\right]^2}{N - 1}$$

  Sometimes, you may divide by N instead of $N - 1$. (With a lot of data, this makes no difference.) The variance is an intermediate input to the more interesting statistic, the standard deviation.

- The standard deviation is the square root of the variance. The standard deviation of the rate of return of a portfolio is commonly used as the measure of its risk.

$$\mathcal{S}dv(\tilde{r}_P) = \sqrt{\mathcal{V}ar(\tilde{r}_P)}$$

- Diversification reduces the risk of a portfolio.
- We assume that investors are smart enough to hold widely diversified portfolios, which resemble the overall market portfolio. Diversified portfolios offer higher expected rates of return at lower risks compared to undiversified portfolios.
- An individual asset's own risk is not a good measure of its risk contribution to a portfolio.
- Market beta is a good measure of the risk contribution of an individual asset for an investor who holds the market portfolio.

- Market betas for typical stocks range between 0 and 2.5.
- It is a straightforward application of formulas to compute beta, correlation, and covariance. They are closely related and always share the same sign.
- Like expected rates of return, betas can be averaged (using proper weighting). However, variances or standard deviations cannot be averaged.

KEY TERMS

asset beta, 221
beta, 213
covariance, 215
diversification, 206
equal-weighted portfolio, 220
equity beta, 221
expected rate of return, 202
linear regression, 216
market beta, 213
market model, 213
market portfolio, 208
portfolio risk, 204
reward, 202
standard deviation, 204
value-weighted portfolio, 220
variance, 204

SOLVE NOW! SOLUTIONS

Q 8.1 The average deviation from the mean is always 0.

Q 8.2 The mean of portfolio A was 4%. Adding 5% to each return will give you a mean of 9%, which is 5% higher. The variance and standard deviation remain at the same level, the latter being 4.42%. If you think of 5% as a constant c, you have just shown that $\mathcal{E}(\tilde{r} + c) = \mathcal{E}(\tilde{r}) + c$ and $\mathcal{S}dv(\tilde{r} + c) = \mathcal{S}dv(\tilde{r})$.

Q 8.3 The reward of portfolio C is its expected rate of return. This is simply $[(-2\%) + 3\% + 7\% + 12\%]/4 = 5\%$. (We just divide by 4, rather than multiply each term by 1/4, because all outcomes are equally likely.) The variance of C is $[(-7\%)^2 + (-2\%)^2 + (2\%)^2 + (7\%)^2]/4 = 26.5\%\%$. The standard deviation, which is our measure of risk, is $\sqrt{26.5\%\%} \approx 5.15\%$.

Q 8.4 For the combination portfolio of 90% in A and 10% in B:

(a) The reward, that is, the expected rate of return, is $0.9 \cdot 4\% + 0.1 \cdot 4\% = 4\%$. To work out the variance, first compute the rates of return in the four states:

$$\begin{aligned} S1: \quad & 0.9 \cdot (-1\%) + 0.1 \cdot (2\%) = -0.7\% \\ S2: \quad & 0.9 \cdot (2\%) + 0.1 \cdot (11\%) = 2.9\% \\ S3: \quad & 0.9 \cdot (4\%) + 0.1 \cdot (-1\%) = 3.5\% \\ S4: \quad & 0.9 \cdot (11\%) + 0.1 \cdot (4\%) = 10.3\% \end{aligned}$$

The variance is

$$\frac{(-0.7\% - 4\%)^2 + (2.9\% - 4\%)^2 + (3.5\% - 4\%)^2 + (10.3\% - 4\%)^2}{4}$$

$$\approx \frac{22.09\%\% + 1.21\%\% + 0.25\%\% + +39.69\%\%}{4} = 15.81\%\%$$

The standard deviation is $\sqrt{15.81\%\%} \approx 3.98\%$.

(b) Figure 8.2 on page 207 showed that the risk (standard deviation) of the 50%-50% portfolio was 3.04%. The risk (standard deviation) of the 90%-10% portfolio is 3.98%. Thus, the latter portfolio looks more spread out in a bar plot.

Q 8.5 For the MC portfolio, the portfolio combination rates of return in the four scenarios were on the right side of the table in Figure 8.3 on page 210. Let's confirm them first:

$$\begin{aligned}
\text{In S1:}\quad & 0.5 \cdot (-1\%) + 0.5 \cdot (-2\%) = -1.5\% \\
\text{In S2:}\quad & 0.5 \cdot (2\%) + 0.5 \cdot (3\%) = 2.5\% \\
\text{In S3:}\quad & 0.5 \cdot (4\%) + 0.5 \cdot (7\%) = 5.5\% \\
\text{In S4:}\quad & 0.5 \cdot (11\%) + 0.5 \cdot (12\%) = 11.5\%
\end{aligned}$$

The expected rate of return is

$$\mathcal{E}(\tilde{r}_{\text{MC}}) = \frac{-1.5\% + 2.5\% + 5.5\% + 11.5\%}{4} = 4.5\%$$

The variance of this portfolio is

$$\mathcal{V}ar_{\text{MC}} = \frac{(-1.5\% - 4.5\%)^2 + (2.5\% - 4.5\%)^2 + (5.5\% - 4.5\%)^2 + (11.5\% - 4.5\%)^2}{4} = 22.5\%\%$$

Therefore, $\mathcal{S}dv_{\text{MC}} = \sqrt{22.5\%\%} \approx 4.74\%$.
For the MD portfolio,

$$\begin{aligned}
\text{In S1:}\quad & 0.5 \cdot (-1\%) + 0.5 \cdot (14\%) = 6.5\% \\
\text{In S2:}\quad & 0.5 \cdot (2\%) + 0.5 \cdot (6\%) = 4.0\% \\
\text{In S3:}\quad & 0.5 \cdot (4\%) + 0.5 \cdot (0\%) = 2.0\% \\
\text{In S4:}\quad & 0.5 \cdot (11\%) + 0.5 \cdot (-12\%) = -0.5\%
\end{aligned}$$

The expected rate of return is

$$\mathcal{E}(\tilde{r}_{\text{MD}}) = \frac{6.5\% + 4.0\% + 2.0\% - 0.5\%}{4} = 3\%$$

The variance is $\mathcal{V}ar_{\text{MD}} = [(6.5\% - 3\%)^2 + (4.0\% - 3\%)^2 + (2.0\% - 3\%)^2 + (-0.5\% - 3\%)^2]/4 = 6.625\%\%$. Therefore, $\mathcal{S}dv_{\text{MD}} = \sqrt{6.625\%\%} \approx 2.57\%$.

Q 8.6 The order of subscripts on market beta is important. Algebraically, $\beta_{\text{C,M}} = [\text{cov}(\tilde{r}_{\text{C}}, \tilde{r}_{\text{M}})]/[\text{var}(\tilde{r}_{\text{M}})]$, while $\beta_{\text{MC}} = [\text{cov}(\tilde{r}_{\text{C}}, \tilde{r}_{\text{M}})]/[\text{var}(\tilde{r}_{\text{C}})]$. The denominator is different. The easiest way to compute the latter is to pick off the standard deviation of 5.15% from Table 8.1 and square it ($26.52\%\% = 0.2652\%$). Therefore, the beta is

$$\beta_{\text{M,C}} = \frac{\mathcal{C}ov(\tilde{r}_{\text{M}}, \tilde{r}_{\text{C}})}{\mathcal{V}ar(\tilde{r}_{\text{C}})} \approx \frac{0.22\%}{0.2652\%} \approx 0.83$$

This is not the same as $\beta_{\text{C,M}} \approx 1.128$. Fortunately, you will never ever need to compute $\beta_{\text{M,C}}$. I only asked you to do this computation so that you realize that the subscript order is important.

Q 8.7 The market beta of this project is

$$\beta_{x,\text{M}} = \frac{\tilde{r}_{x,2} - \tilde{r}_{x,1}}{\tilde{r}_{\text{M},2} - \tilde{r}_{\text{M},1}} = \frac{(-5\%) - (+5\%)}{(-10\%) - (+10\%)} = +0.5$$

(This is not "half as volatile" because market beta is not a measure of volatility.)

Q 8.8 Using the same formula, the market beta is $[(+5\%) - (-5\%)]/[(-10\% - (+10\%)] = -0.5$.

Q 8.9 To check that Formula 8.6 on page 220 is correct, you must compute the market beta for CDD from the rates of return for the entire firm CDD.

(a) The second and third columns in the following table show the rates of return on the market and on CDD in each of the four states:

	Original Base Rates		Net-of-Mean Rates		
Scenario	$\tilde{r}_M$	$\tilde{r}_{CDD}$	$\tilde{r}_M$	$\tilde{r}_{CDD}$	Cross-product
In S1 (♣)	−1%	8.67%	−5%	5.67%	−28.35%%
In S2 (♦)	2%	5.00%	−2%	2.00%	−4.00%%
In S3 (♥)	4%	2.33%	0%	−0.67%	0.00%%
In S4 (♠)	11%	−4.00%	7%	−7.00%	−49.00%%
Mean	4%	3%	0%	0%	−20.33%%

(b) The average rates of return are in the last row of the table.
(c) The de-meaned rates of return are in the fourth and fifth columns.
(d) The cross-products are in the sixth column.
(e) The average cross-product is in the last row of the sixth column.
(f) Using Formula 8.6, the beta of investment CDD is

$$\beta_{CDD} \approx \frac{\mathcal{C}ov(\tilde{r}_M, \tilde{r}_C)}{\mathcal{V}ar(\tilde{r}_M)} = \frac{-0.2033\%}{0.195\%} \approx -1.04$$

(g) Formula 8.6 is a bit easier than this route. The advantage would be even more obvious if you had a few hundred securities and a few thousand trading days, and you already knew the market beta for each of them individually.

In any case, you have now confirmed that Formula 8.6 yielded the same result. You did not catch me in a lie.

Q 8.10 To confirm that you cannot value-weight variances (and thus standard deviations):

(a) The variance of $\tilde{r}_C$ was 26.5%%. The variance of $\tilde{r}_D$ was 90.0%%. The value-weighted average of one part variance of C and two parts variance of D is $w_C \cdot \tilde{r}_C + w_D \cdot \tilde{r}_D = 1/3 \cdot 26.5\%\% + 2/3 \cdot 90.0\%\% \approx 68.83\%$.
(b) The actual variance of CDD is $\mathcal{V}ar(\tilde{r}_{CDD}) \approx [(5.67\%)^2 + (2\%)^2 + (-0.67\%)^2 + (-7\%)^2]/4 \approx 85.6\%\%/4 \approx 21.4\%\%$.

Q 8.11 The CCD portfolio has rates of return of 3.3333%, 4.00%, 4.6667%, and 4.00% in the four states. De-meaned, this is −0.6667%, 0%, 0.6667%, and 0%. Therefore, the variance of CCD is $[(-0.6667\%)^2 + (0\%)^2 + (0.6667\%)^2 + (0\%)^2]/4 \approx 0.224\%\%$, and its standard deviation is 0.47%. The de-meaned rates of return on M are −5%, −2%, 0, and 7%. The cross-products of the de-meaned CCD rates of return with the de-meaned M rates of return are therefore 3.3333%%, 0, 0, and 0. Therefore, the covariance of CCD and M is $(3.3333\%\% + 0 \cdot 3)/4 \approx 0.8333\%\%$. The variance of the market is 19.5%%. Therefore, the market beta of CCD is $0.833/19.5 \approx 0.0427$. This was the first method. Now the second method: $\beta_{CCD} = w_C \cdot \beta_C + w_D \cdot \beta_D \approx 2/3 \cdot (+1.128) + 1/3 \cdot (-2.128) \approx 0.0427$.

Q 8.12 For a firm whose debt is risk free, the overall firm beta is $\beta_{Firm} = 0.5 \cdot \beta_{Equity} + 0.5 \cdot \beta_{Debt}$. Thus, $0.5 \cdot \beta_{Equity} + 0.5 \cdot 0 = 2$. Solve for $\beta_{Equity} = \beta_{Firm}/0.5 = 4$. For the (90%, 10%) case, the equity beta jumps to $\beta_{Equity} = 2/0.1 = 20$.

PROBLEMS

The Q indicates problems available in myfinancelab

When not otherwise specified in these problems, questions refer to the named portfolios A through F from Table 8.1.

Q 8.13 Multiply each rate of return for A by 2.0. This portfolio offers −2%, +4%, +8%, and +22%. Compute the expected rate of return and standard deviation of this new portfolio. How do they compare to those of the original portfolio A?

Q 8.14 The following were the closing year-end prices of the Japanese stock market index, the Nikkei-225:

1984	11,474	1992	16,925	2000	13,786
1985	13,011	1993	17,417	2001	10,335
1986	18,821	1994	19,723	2002	8,579
1987	22,957	1995	19,868	2003	10,677
1988	29,698	1996	19,361	2004	11,489
1989	38,916	1997	15,259	2005	16,111
1990	24,120	1998	13,842	2006	17,225
1991	22,984	1999	18,934	2007	15,308

Assume that each historical rate of return was exactly one representative scenario (independent sample draw) that you can use to estimate the future. If a Japanese investor had purchased a mutual fund that imitated the Nikkei-225, what would her annual rates of return, compounded rate of return (from the end of 1984 to the end of 2007), average rate of return, and risk have been?

Q 8.15 Compute the value-weighted average of 1/3 of the standard deviation of C and 2/3 of the standard deviation of D. Is it the same as the standard deviation of a CDD portfolio of 1/3 C and 2/3 D, in which your investment rate of return would be $1/3 \cdot \tilde{r}_C + 2/3 \cdot \tilde{r}_D$?

Q 8.16 What are the risk and reward of a combination portfolio that invests 40% in A and 60% in B?

Q 8.17 Consider the following five assets, which have rates of return in six equally likely possible scenarios:

	Scenarios					
	Awful	Poor	Med.	Okay	Good	Great
Asset P1	–2%	0%	2%	4%	6%	10%
Asset P2	–1%	2%	2%	2%	3%	3%
Asset P3	–6%	2%	2%	3%	3%	1%
Asset P4	–4%	2%	2%	2%	2%	20%
Asset P5	10%	6%	4%	2%	0%	–2%

(a) Assume you can only purchase one of these assets. What are their risks and rewards?
(b) Supplement your previous risk-reward rankings of assets P1–P5 with those of combination portfolios that consist of half P1 and half of each of the other 4 portfolios, P2–P5. What are the risks and rewards of these four portfolios?
(c) Assume that P1 is the market. Plot the rates of return for P1 on the x-axis and the return for each of the other stocks on their own y-axes. Then draw lines that you think best fit the points. Do not try to compute the beta—just use the force (and your eyes), Luke. If you had to buy just a little bit of one of these P2–P5 assets, and you wanted to lower your risk, which would be best?

Q 8.18 Assume you have invested half of your wealth in a risk-free asset and half in a risky portfolio P. Is it theoretically possible to lower your portfolio risk if you move your risk-free asset holdings into another risky portfolio Q? In other words, can you ever reduce your risk more by buying a risky security than by buying a risk-free asset?

Q 8.19 Why is it so common to use historical financial data to estimate future market betas?

Q 8.20 Is it wise to rely on historical statistical distributions as our guide to the future?

Q 8.21 Look up the market betas of the companies in Table 8.2. Have they changed dramatically

since May 2008, or have they remained reasonably stable?

Q 8.22 You estimate your project to return −20% if the stock market returns −10%, and +5% if the stock market returns +10%. What would you use as the market beta estimate for your project?

Q 8.23 Go to Yahoo! *Finance*. Obtain 2 years' worth of weekly rates of return for PepsiCo and for the S&P 500 index. Use a spreadsheet to compute PepsiCo's market beta.

Q 8.24 Consider the following assets:

	Scenario		
	Bad	Okay	Good
Market M	−5%	5%	15%
Asset X	−2%	−3%	25%
Asset Y	−4%	−6%	30%

(a) Compute the market betas for assets X and Y.
(b) Compute the correlations of assets X and Y with M.
(c) Assume you were holding only M. You now are selling off 10% of your M portfolio to replace it with 10% of either X or Y. Would an MX portfolio or an MY portfolio be riskier?
(d) Is the correlation indicative of which of these two portfolios ended up riskier? Is the market beta indicative?

Q 8.25 Compute the expected rates of return and the portfolio betas for many possible portfolio combinations (i.e., different weights) of C and D from Table 8.1 on page 202. (Your weight in D is 1 minus your weight in C.) Plot the two against one another. What does your plot look like?

Q 8.26 The following represents the probability distribution for the rates of return for next month:

Probability	Pfio P	Market M
1/6	−20%	−5%
2/6	−5%	+5%
2/6	+10%	0%
1/6	+50%	+10%

Compute by hand (and show your work) for all the following questions.
(a) What are the risks and rewards of P and M?
(b) What is the correlation of M and P?
(c) What is the market beta of P?
(d) If you were to hold 1/3 of your portfolio in the risk-free asset, and 2/3 in portfolio P, what would its market beta be?

Q 8.27 Download the historical prices for the S&P 500 index (~spx or ~gspc) and for VPACX (the *Vanguard Pacific Stock Index* mutual fund) from Yahoo! *Finance*, beginning January 1, 2004, and ending December 31 of last year. Load them into a spreadsheet and position them next to one another. Compute the historical rates of return. Compute the risk and reward. Compute VPACX's market beta with respect to the S&P 500 index. How do your estimates compare to the Fund Risk as noted by Yahoo! *Finance*?

Q 8.28 Download 5 years of historical monthly (dividend-adjusted) prices for Coca-Cola (KO) and the S&P 500 from Yahoo! *Finance*.
(a) Compute the monthly rates of return.
(b) Compute the average rate of return and risk of portfolios that combine KO and the S&P 500 in the following proportions: (0.0, 1.0), (0.2, 0.8), (0.4, 0.6), (0.6, 0.4), (0.8, 0.2), (1.0, 0.0). Then plot them against one another. What does the plot look like?
(c) Compute the market beta of Coca-Cola.

Q 8.29 Are historical covariances or means more trustworthy as estimators of the future?

Q 8.30 Why do some statistical packages estimate covariances differently (and different from those we computed in this chapter)? Does the same problem also apply to expected rates of return (means) and betas?

CHAPTER 8 APPENDIX

Trade-Off between Risk and Return

8.6 AN INVESTOR'S SPECIFIC TRADE-OFF BETWEEN RISK AND REWARD

What is the optimal portfolio of assets? (A portfolio is a complete set of weights on all possible assets.)

This appendix develops the trade-off between risk and return. Although this is not central to the subject of corporate finance, it is central to the subject of investments. So, where are we and where are we going?

- You already know that diversification reduces risk.
- Therefore, you know that you like diversification.
- You know that assets that covary negatively with the rest of your portfolio are particularly desirable from a diversification perspective.
- The beta of an asset with respect to a portfolio is its measure of "toxicity" in the context of the portfolio.

The question that you cannot yet answer is

- Exactly how much of each asset should you purchase?

For example, is it better to purchase 25% in A and 75% in B, or 50% in each? How do you determine good investment weights? What is your optimal investment portfolio?

Table 8.4 computes different combinations of two assets to get various portfolio risk-reward characteristics.

Let's make up two new base assets, H and I. (If you wish, you can think of these assets as themselves being portfolios containing many different stocks.) How do you find the best combination portfolio of H and I? Table 8.4 shows some of the portfolios you could put together. Let's confirm the numbers for at least one of these. Portfolio K invests $w_H = 1/3$ in H and $w_I = 2/3$ in I, which means it has the following possible outcomes:

$$\begin{aligned}
&\text{In Scenario S1 } \clubsuit \quad \tilde{r}_K = 1/3 \cdot (-6\%) + 2/3 \cdot (-12\%) = -10\% \\
&\text{In Scenario S2 } \diamondsuit \quad \tilde{r}_K = 1/3 \cdot (+12\%) + 2/3 \cdot (+18\%) = +16\% \\
&\text{In Scenario S3 } \heartsuit \quad \tilde{r}_K = 1/3 \cdot (0\%) + 2/3 \cdot (+24\%) = +16\% \\
&\text{In Scenario S4 } \spadesuit \quad \tilde{r}_K = 1/3 \cdot (+18\%) + 2/3 \cdot (+6\%) = +10\%
\end{aligned}$$

$$\tilde{r}_K = w_H \cdot (\tilde{r}_H) + w_I \cdot (\tilde{r}_I)$$

The expected rate of return of this portfolio, given all possible future scenarios, is then

$$\mathcal{E}(\tilde{r}_K) = 1/4 \cdot (-10\%) + 1/4 \cdot (+16\%) + 1/4 \cdot (+16\%) + 1/4 \cdot (+10\%) = 8\%$$

$$\mathcal{E}(\tilde{r}) = \text{Sum over All Scenarios } S\text{: } \mathcal{P}rob(\text{Scenario } S) \cdot \text{Outcome in Scenario } S$$

To compute the variance of K, you follow the procedure laid out in Section 6.1B: First, take out the mean from the rates of return:

➤ Standard deviation, Section 6.1B, p. 141

$$\begin{aligned}
&\text{In Scenario S1 } \clubsuit \quad -10\% - 8\% = -18\% \\
&\text{In Scenario S2 } \diamondsuit \quad +16\% - 8\% = +8\% \\
&\text{In Scenario S3 } \heartsuit \quad +16\% - 8\% = +8\% \\
&\text{In Scenario S4 } \spadesuit \quad +10\% - 8\% = +2\% \\
&\qquad\qquad \tilde{r}_K - \mathcal{E}(\tilde{r}_K)
\end{aligned}$$

Second, square them and compute the average:

$$\mathcal{V}ar(\tilde{r}_K) = \frac{(-18\%)^2 + (+8\%)^2 + (+8\%)^2 + (+2\%)^2}{4} = 114\%\% \qquad (8.8)$$

The risk is therefore $\mathcal{S}dv(\tilde{r}_J) = \sqrt{\mathcal{V}ar(\tilde{r}_K)} = \sqrt{114\%\%} \approx 10.68\%$. You have now confirmed the three statistics for portfolio K in Table 8.4: the 8% expected rate of return (reward), 114%% variance, and 10.68% standard deviation (risk).

Do you care about your portfolio's beta or your portfolio's standard deviation? Make sure you understand the answer to this question.

IMPORTANT:

- As an investor, you usually care only about your portfolio's standard deviation (risk). (You rarely ever care about the overall market beta of your asset holdings.)
- If you are the CFO of a firm that wants to get into the market portfolio, so that investors willingly buy your shares, then you do care about your single firm's market beta. You should not care primarily about your firm's own standard deviation (idiosyncratic risk), because your investors do not care about it. They can diversify away your firm's idiosyncratic risk.

TABLE 8.4 PORTFOLIOS USED TO ILLUSTRATE MEAN-VARIANCE COMBINATIONS

	Base Assets		Combination Portfolios				
Pfio Name	100% in Pfio H H	100% in Pfio I I	1/4 in H 3/4 in I J	1/3 in H 2/3 in I K	1/2 in H 1/2 in I L	2/3 in H 1/3 in I M	3/4 in H 1/4 in I N
In Scenario S1 ♣	−6.0%	−12.0%	−10.50%	−10.00%	−9.00%	−8.00%	−7.50%
In Scenario S2 ♦	+12.0%	+18.0%	+16.50%	+16.00%	+15.00%	+14.00%	+13.50%
In Scenario S3 ♥	0.0%	+24.0%	+18.00%	+16.00%	+12.00%	+8.00%	+6.00%
In Scenario S4 ♠	+18.0%	+6.0%	+9.00%	+10.00%	+12.00%	+14.00%	+15.00%
"Reward" ($\mathcal{E}(\tilde{r})$)	6.00%	9.00%	8.25%	8.00%	7.50%	7.00%	6.75%
"Variance" ($\mathcal{V}ar(\tilde{r})$)	90.0%%	189.0%%	128.8%%	114.0%%	92.2%%	81.0%%	79.3%%
"Risk" ($\mathcal{S}dv(\tilde{r})$)	9.49%	13.75%	11.35%	10.68%	9.60%	9.00%	8.91%

These are the two base assets (and their combinations) used to illustrate the mean-variance efficient frontier in Section 8.8.

SOLVE NOW!

Q 8.31 Confirm the portfolio variance and standard deviation if you invest in portfolio M ($w_H = 2/3$) in Table 8.4.

Q 8.32 Confirm the portfolio variance and standard deviation if you invest in portfolio N ($w_H = 3/4$) in Table 8.4.

8.7 A SHORTCUT FORMULA FOR THE RISK OF A PORTFOLIO

We want to write the portfolio variance as a function of the component investment weights. This is a common shortcut formula.

There is a shortcut formula that can make portfolio variance computations faster. This shortcut allows you to compute the variance of a portfolio as a function of the weights in each constituent asset. To use it, you need to know the covariances between all assets. The formula also avoids having to first work out the rate of return of the combination portfolio in each and every scenario—not a big deal when there are four scenarios, but a very big deal if you have a thousand daily observations, each of which can count as a scenario, and you want to consider many portfolios with various weights.

Example: We still need the covariance between H and I.

➤ Covariance computation, Section 8.3B, p. 214

For our two assets, you need only one extra number for the new variance shortcut formula: You have to compute the covariance between your two base portfolios, here H and I. You have already worked with the covariance in Section 8.3B. It is defined as the average product of the two net-of-mean returns. Subtract the mean (6% for H and 9% for I) from each scenario's realization:

	Portfolio H		Portfolio I	
In Scenario S1 ♣	$\tilde{r}_H - \mathcal{E}(\tilde{r}_H)$	$= -12\%$	$\tilde{r}_I - \mathcal{E}(\tilde{r}_I)$	$= -21\%$
In Scenario S2 ♦	$\tilde{r}_H - \mathcal{E}(\tilde{r}_H)$	$= +6\%$	$\tilde{r}_I - \mathcal{E}(\tilde{r}_I)$	$= +9\%$
In Scenario S3 ♥	$\tilde{r}_H - \mathcal{E}(\tilde{r}_H)$	$= -6\%$	$\tilde{r}_I - \mathcal{E}(\tilde{r}_I)$	$= +15\%$
In Scenario S4 ♠	$\tilde{r}_H - \mathcal{E}(\tilde{r}_H)$	$= +12\%$	$\tilde{r}_I - \mathcal{E}(\tilde{r}_I)$	$= -3\%$

Therefore,

$$\mathcal{C}ov(\tilde{r}_H, \tilde{r}_I) = \frac{(-12\%)\cdot(-21\%) + (+6\%)\cdot(+9\%) + (-6\%)\cdot(+15\%) + (+12\%)\cdot(-3\%)}{4} = +45\%\%$$

$$\mathcal{C}ov(\tilde{r}_H, \tilde{r}_I) = \frac{\text{Sum over All Scenarios (or Observations) } S\text{: } [\tilde{r}_{H,S} - \mathcal{E}(\tilde{r}_H)]\cdot[\tilde{r}_{I,S} - \mathcal{E}(\tilde{r}_I)]}{N} \tag{8.9}$$

H and I are positively correlated—these investments tend to move together. Intuitively, this means, for example, that if the rate of return on portfolio H exceeds its 6% mean, portfolio I will also tend to exceed its own 9% mean.

Without further ado, the box that follows gives the shortcut formula for two assets.

IMPORTANT: The variance of a portfolio P that consists only of A and B, that is, with returns of $\tilde{r}_P = w_A \cdot \tilde{r}_A + w_B \cdot \tilde{r}_B$, where w_A is the portfolio weight in component A, and w_B is the portfolio weight in component B, is

$$\mathcal{V}ar(\tilde{r}_P) = w_A^2 \cdot \mathcal{V}ar(\tilde{r}_A) + w_B^2 \cdot \mathcal{V}ar(\tilde{r}_B) + 2 \cdot w_A \cdot w_B \cdot \mathcal{C}ov(\tilde{r}_A, \tilde{r}_B) \quad (8.10)$$

Check whether this is correct. Try it out on portfolio K, which invests 1/3 in H and 2/3 in I:

$$\begin{aligned}\mathcal{V}ar(\tilde{r}_K) &= (1/3)^2 \cdot \mathcal{V}ar(\tilde{r}_H) + (2/3)^2 \cdot \mathcal{V}ar(\tilde{r}_I) + 2 \cdot (1/3) \cdot (2/3) \cdot \mathcal{C}ov(\tilde{r}_H, \tilde{r}_I) \\ &= (1/3)^2 \cdot 90\%\% + (2/3)^2 \cdot 189\%\% + 2 \cdot (1/3) \cdot (2/3) \cdot (+45\%\%) \\ &= 114\%\%\end{aligned}$$

This is the same result as we computed in Formula 8.8, so the shortcut indeed gives the correct answer.

The general formula comes with a good memorization aid.

One way to remember this formula—and the more general version with more than two securities—is to create a matrix of all your assets. It's simple. Write all your assets' names on both edges, their weights next to them, and write into each cell what is on the edges as well as a covariance between what's on the edges:

		A	B	C	⋯
		w_A	w_B	w_C	
A	w_A	$w_A \cdot w_A \cdot \mathcal{C}ov(\tilde{r}_A, \tilde{r}_A)$	$w_A \cdot w_B \cdot \mathcal{C}ov(\tilde{r}_A, \tilde{r}_B)$	$w_A \cdot w_C \cdot \mathcal{C}ov(\tilde{r}_A, \tilde{r}_C)$	
B	w_B	$w_B \cdot w_A \cdot \mathcal{C}ov(\tilde{r}_B, \tilde{r}_A)$	$w_B \cdot w_B \cdot \mathcal{C}ov(\tilde{r}_B, \tilde{r}_B)$	$w_B \cdot w_C \cdot \mathcal{C}ov(\tilde{r}_B, \tilde{r}_C)$	
C	w_C	$w_C \cdot w_A \cdot \mathcal{C}ov(\tilde{r}_C, \tilde{r}_A)$	$w_C \cdot w_B \cdot \mathcal{C}ov(\tilde{r}_C, \tilde{r}_B)$	$w_C \cdot w_C \cdot \mathcal{C}ov(\tilde{r}_C, \tilde{r}_C)$	
⋮					⋱

That's it. By the way, did you notice that if you have m securities, there are only m variance terms in this matrix (on the diagonal), but $m^2 - m$ covariance terms? For 500 assets, you have 500 variance cells and 249,500 covariance cells. Adding the next security to the portfolio would add 1 variance term and 500 covariance terms. It should suggest to you that it need not be far-fetched to believe that the covariance of assets—how they fit together—can be more important than their own variances.

Apply the formula to compute the variance of K again.

Now substitute our specific investment weights for portfolio K, which are $w_H = 1/3$, $w_I = 2/3$. Let me also show you that investments that you do not own (call a sample one J) just drop out of the formula:

		H	I	J	⋯
		1/3	2/3	0	
H	1/3	$1/3 \cdot 1/3 \cdot \mathcal{C}ov(\tilde{r}_H, \tilde{r}_H)$	$1/3 \cdot 2/3 \cdot \mathcal{C}ov(\tilde{r}_H, \tilde{r}_I)$	$1/3 \cdot 0 \cdot \mathcal{C}ov(\tilde{r}_H, \tilde{r}_J)$	
I	2/3	$2/3 \cdot 1/3 \cdot \mathcal{C}ov(\tilde{r}_I, \tilde{r}_H)$	$2/3 \cdot 2/3 \cdot \mathcal{C}ov(\tilde{r}_I, \tilde{r}_I)$	$2/3 \cdot 0 \cdot \mathcal{C}ov(\tilde{r}_I, \tilde{r}_J)$	
J	0	$0 \cdot 1/3 \cdot \mathcal{C}ov(\tilde{r}_J, \tilde{r}_H)$	$0 \cdot 2/3 \cdot \mathcal{C}ov(\tilde{r}_J, \tilde{r}_I)$	$2/3 \cdot 0 \cdot \mathcal{C}ov(\tilde{r}_J, \tilde{r}_J)$	
⋮					⋱

All cells with J just multiply everything with a zero, so they can be omitted. Next, use the fact that, by definition, the covariance of something with itself is its variance. So, the matrix is

		H	I
		1/3	2/3
H	1/3	$1/3 \cdot 1/3 \cdot 90\%\%$	$1/3 \cdot 2/3 \cdot 45\%\%$
I	2/3	$2/3 \cdot 1/3 \cdot 45\%\%$	$2/3 \cdot 2/3 \cdot 189\%\%$

Add up all the cells, and you have the variance of portfolio K.

$$\mathcal{V}ar(\tilde{r}_K) = 1/3 \cdot 1/3 \cdot 90\%\% + 1/3 \cdot 2/3 \cdot 45\%\%$$
$$2/3 \cdot 1/3 \cdot 45\%\% + 2/3 \cdot 2/3 \cdot 189\%\% = 114\%\%$$

Again, this is the correct answer that you already knew.

For H and I, this formula is not any more convenient than computing the scenario or historical time series of portfolio returns first and then computing the variance of this one series. However, the formula is a lot more convenient if you have to compute the portfolio variance of thousands of different combinations of H and I and there are hundreds of scenarios. And it is precisely this process—recomputing the overall portfolio variance many times—that is at the heart of determining the best portfolio: You want to know how different portfolio weights change your portfolio risk. Your alternative to the shortcut would be to recompute the returns for each of the hundreds of possible portfolio weight combinations—which would quickly become very painful.

This formula is useful if you want to try thousands of different portfolios (investment weights).

SOLVE NOW!

Q 8.33 Show that the shortcut Formula 8.10 works for portfolio M, in which H is 2/3. That is, does it give the same 81.0%% noted in Table 8.4 on page 232?

Q 8.34 Show that the shortcut Formula 8.10 works for portfolio N, in which H is 3/4. That is, does it give the same 79.3%% noted in Table 8.4?

Q 8.35 (This question is very important. Please do not pass over it.) Let's consider a stock market index, such as the S&P 500. It had a historical average rate of return of about 12% per annum, and a historical standard deviation of about 20% per annum. Assume for the moment:

(a) **Known statistical distributions:** You know the expected reward and risk. In our example, we assume that they are the historical averages and risks. This is convenient.

(b) **Independent stock returns:** Stock returns are (mostly) uncorrelated over time periods. This is reasonable because if this were not so, you could earn money purchasing stocks based on their prior performance in a perfect market. (This will be the subject of Chapter 11.)

(c) **No compounding:** The rate of return over X years is the simple sum of X annual rates of return. (That is, we ignore the cross-product terms that are rates of return on rates of return.) This is problematic over decades, but not over just a few months or even years.

Our goal is to work out how asset risk grows with time under these assumptions. The variance shortcut formula will help us.

(a) Write down the formula for the total rate of return over 2 years.
(b) What is the expected total rate of return over 2 years?
(c) Write down the formula for the variance over 2 years.
(d) What is the specific risk here (variance and standard deviation) over 2 years?
(e) The **Sharpe ratio** is a common (though flawed) measure of portfolio performance. It is usually computed as the expected rate of return above the risk-free rate, then divided by the standard deviation. Assume that the risk-free rate is 6%. Thus, the 1-year Sharpe ratio is $(12\% - 6\%)/20\% \approx 0.3$. What is the 2-year Sharpe ratio?
(f) What are the expected rate of return and risk (variance and standard deviation) over 4 years? What is the 4-year Sharpe ratio?
(g) What are the expected rate of return and risk (variance and standard deviation) over 16 years? What is the 16-year Sharpe ratio?
(h) What are the expected rate of return and risk (variance and standard deviation) over T years? What is the T-year Sharpe ratio?
(i) What are the expected rate of return and risk (variance and standard deviation) over 1 month? What is the 1-month Sharpe ratio?
(j) What are the expected rate of return and risk (variance and standard deviation) over 1 trading day? What is the 1-day Sharpe ratio? Assume 250 trading days per year.

8.8 GRAPHING THE MEAN-VARIANCE EFFICIENT FRONTIER

Graphing the trade-off between risk and reward . . .

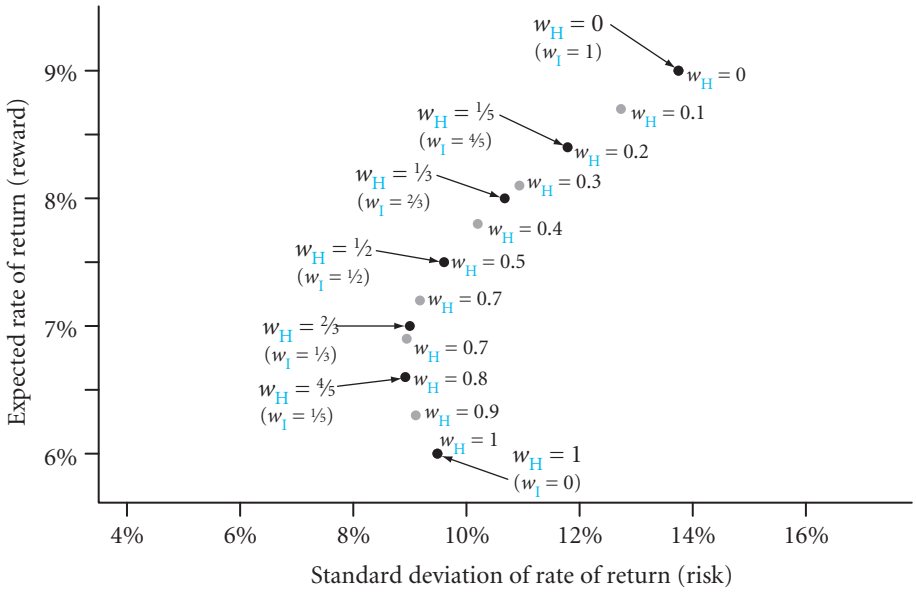
Let's now graph the portfolio risk on the x-axis and the portfolio reward on the y-axis for each portfolio from Table 8.4 on page 232. Figure 8.5 does it for you. Can you see a pattern? To make it easier, I have taken the liberty of adding a few more portfolios. (You can confirm that I have computed the risk and reward of one of these portfolios in Q 8.36.)

. . . is called the mean-variance efficient frontier.

If you picked many more portfolios with portfolio weights on H between 0 and 100%, you would eventually end up with Figure 8.6. The curve is called the **mean-variance efficient frontier** (**MVE frontier**), and it is the region where the best risk-reward portfolios lie. There must not be any portfolios to the northwest of this frontier—they would have a higher expected rate of return for a given risk, or lower risk for a given expected rate of return. If these existed, they would themselves be the MVE frontier. (The shape of the mean-variance efficient frontier is a so-called hyperbola when the x-axis is the standard deviation.)

The minimum-variance portfolio is the west-most portfolio.

The west-most portfolio on the efficient frontier is called the **minimum-variance portfolio** because you cannot create a portfolio with lower risk. You need a lot of algebra to find it, so I have worked this out for you. In our example, the minimum-variance portfolio has a weight of 76.191% on H and 23.809% on I, and it achieves



These are the portfolios from Table 8.4, and then some more in gray that I computed—a hobby.

FIGURE 8.5 The Risk-Reward Trade-Off between H and I: More Portfolios

as low a risk as 8.9%. Although the graph's scale is too small for you to check this graphically, you can compute the risk of this minimum-variance portfolio that I gave you and compare it to the risk of two portfolios that invest either a little more or a little less into H.

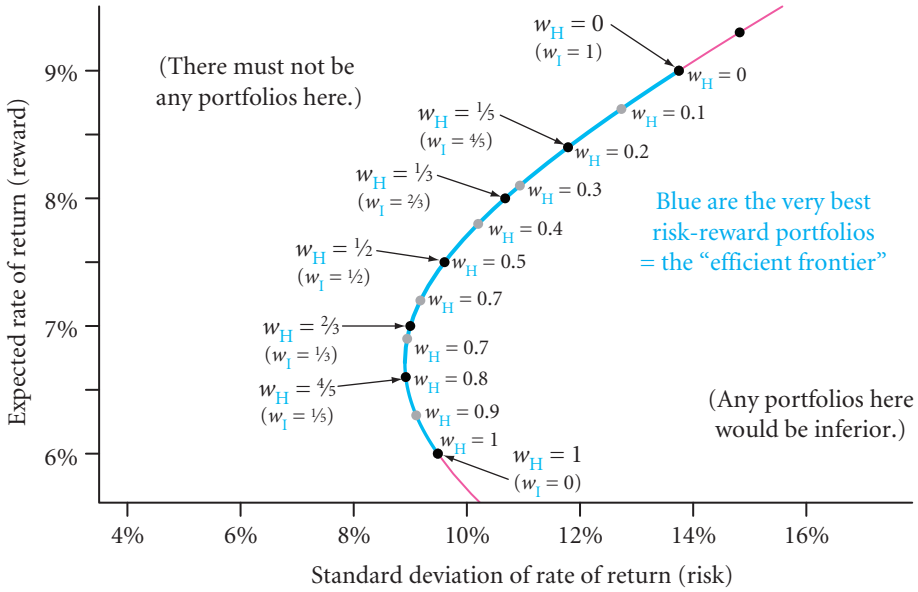
$$\begin{aligned} w_H = 76.0\% &: \mathcal{S}dv(\tilde{r}_P) \approx 8.9042911\% \\ w_H = 76.2\% &: \mathcal{S}dv(\tilde{r}_P) \approx 8.9042526\% \quad \longleftarrow \text{I claimed lowest risk} \\ w_H = 76.4\% &: \mathcal{S}dv(\tilde{r}_P) \approx 8.9042992\% \end{aligned} \tag{8.11}$$

$$\mathcal{S}dv(\tilde{r}_P) = \sqrt{\mathcal{V}ar(\tilde{r}_P)} = \mathcal{V}ar[w_H \cdot \tilde{r}_H + (1 - w_H) \cdot \tilde{r}_I] \tag{8.12}$$

If there are assets that can be combined to construct a risk-free asset, then the minimum-variance portfolio will touch the y-axis at 0. If there are only two assets, this means their correlation would have to be -1. More commonly, the minimum-variance portfolio does not touch the y-axis and still has positive risk.

There is one feature of a more general mean-variance graph that this particular graph cannot illustrate. If you had started with more than two base portfolios H and I, you could have found many combination portfolios that would have been outright inferior. They would have been a cloud of points inside and southeast of the efficient frontier. However, the efficient frontier itself would still look very similar to what is in Figure 8.6—a hyperbola on the upper northwest frontier.

One feature is not visible in this figure, because there are only two portfolios: With more assets, many portfolio combinations lie in the interior—a cloud of points.



This connects the points on the efficient frontier to Figure 8.5. Additionally, it completes the efficient frontier beyond interior portfolios, that is, allowing for portfolios that short one or the other portfolio (in magenta).

FIGURE 8.6 The Risk-Reward Trade-Off between H and I: Sets

Allowing Shorted Positions

Extending the MVEF to allow for short positions extends the graph.

Each point on the mean-variance frontier represents one set of investment weights. Interestingly, the relevant formulas work just as well with negative weights as they do with positive weights. For example, if $w_H = (-0.1)$ and $w_I = 1.1$, then the sum of your individual investments is still 100%, and

$$\mathcal{E}(\tilde{r}_P) = (-0.1) \cdot 6\% + (1.1) \cdot 9\% = 9.3\%$$

$$\mathcal{E}(\tilde{r}_P) = w_H \cdot \mathcal{E}(\tilde{r}_H) + w_I \cdot \mathcal{E}(\tilde{r}_I)$$

and

$$\mathcal{S}dv(\tilde{r}_P) = \sqrt{(-0.1)^2 \cdot 90\%\% + (1.1)^2 \cdot 189\%\% + 2 \cdot (-0.1) \cdot (1.1) \cdot 45\%\%} \approx 14.82\%$$

$$\mathcal{S}dv(\tilde{r}_P) = \sqrt{w_H^2 \cdot \mathcal{V}ar(\tilde{r}_H) + w_I^2 \cdot \mathcal{V}ar(\tilde{r}_I) + 2 \cdot w_H \cdot w_I \cdot \mathcal{C}ov(\tilde{r}_I, \tilde{r}_H)}$$

(If you wish, you can first confirm this: This portfolio would return −12.6% (♣), 18.6% (♦), 26.4% (♥), or 4.8% (♠). Therefore, the expected rate of return is 9.3%, and the standard deviation is 14.82%.) This portfolio is marked at the top in Figure 8.6. It is on the continuation of the hyperbola. Actually, I have done more, drawing the rest of the hyperbola in magenta. These are portfolios that contain shorted assets.

The economic meaning of shorting reexplained.

➤ Shorting stocks, Section 7.2A, p. 191

But what is the meaning of an investment with negative weight? It was explained in Section 7.2A: It is shorting a stock. In brief, perfect shorting works as follows: If you short a security, you promise to provide the appropriate returns, rather than earn them. For example, say you want to go short $200 in H and I want to go long $200 in H. I would purchase H from you. This would work as follows:

- I must give you $200 today. (If you want, you can invest this to earn interest.)
- Next year, you must give me exactly what I would get if I had purchased H, not from you, but from someone else who really would have given me the security. That is, if ♣ comes about, you must pay me $188; if ♦ comes about, you must pay me $224; if ♥ comes about, you must pay me $200; and if ♠ comes about, you must pay me $236.

In other words, I won't notice whether you sold me the security or someone else (who had it) sold me the security. This is simple ownership—a 100% investment ownership. Your own rate of return is the exact opposite of my return. For example, if I earn −6%, you would gain +6%. After all, you received $200 from me (at time 0) and are only returning $188 to me (at time 1). What would your return be if you sold $200 of H to me, thereby going short, and then used the $200 to purchase H from someone else in the market? It would always be zero—going long and short by the same amount cancels out perfectly. In a perfect market, you would not earn any money or lose any money.

SOLVE NOW!

Q 8.36 Compute the risk and reward of the portfolio $w_H = 0.1$, $w_I = 0.9$, as in Table 8.4 on page 232. Confirm that this portfolio is drawn correctly in Figure 8.5.

Q 8.37 If there are two risky portfolios that have a correlation of −1 with positive investment weights, what would the expected rate of return on this portfolio be?

Q 8.38 If H and I were more correlated, what would the efficient frontier between them look like? If H and I were less (or more negatively) correlated, what would the efficient frontier between them look like? (Hint: Think about the variance of the combination portfolio that invests half in each.)

Q 8.39 Draw the efficient frontier for the following two base assets, H and Z:

	In Scenario			
Base Portfolio	S1 ♣	S2 ♦	S3 ♥	S4 ♠
H	−6%	+12%	0%	+18%
Z	−12%	+18%	+15%	+15%

Also, compute the covariance between H and Z. Is it higher or lower than what you computed in the text for H and I? How does the efficient frontier compare to what you have drawn in this chapter?

8.9 ADDING A RISK-FREE ASSET

In the real world, you usually have access to a risk-free Treasury. It turns out that the presence of a risk-free asset plays an important role, not only in the model of the next chapter (the CAPM), but also in these mean-variance graphs. So let us now add a risk-free rate ("F") of 4%. Start with the following three basis portfolios:

A special case: The risk and reward of combinations of portfolios with the risk-free asset are both simple linear functions.

Future	H	I	F
In Scenario S1 ♣	−6.0%	−12.0%	4.00%
In Scenario S2 ♦	+12.0%	+18.0%	4.00%
In Scenario S3 ♥	0.0%	+24.0%	4.00%
In Scenario S4 ♠	+18.0%	+6.0%	4.00%
"Reward" ($\mathcal{E}(R)$)	6.00%	9.00%	4.00%
"Variance" ($\mathcal{V}ar(R)$)	90.0%%	189.0%%	0.0%%
"Risk" ($\mathcal{S}dv(R)$)	9.49%	13.75%	0.00%

Begin by determining the risk and reward of a portfolio S that invests 1/2 in H and 1/2 in F: Its rate of return is defined as $\tilde{r}_S = w_H \cdot \tilde{r}_H + w_F \cdot \tilde{r}_F = 1/2 \cdot \tilde{r}_H + 1/2 \cdot 4\%$. The expected reward of this portfolio is

$$\mathcal{E}(\tilde{r}_S) = 1/2 \cdot 6\% + (1 - 1/2) \cdot 4\% = 5\% \qquad (8.13)$$
$$\mathcal{E}(\tilde{r}_S) = w_H \cdot \mathcal{E}(\tilde{r}_H) + (1 - w_H) \cdot r_F$$

➤ Portfolio variance, Formula 8.10, p. 234

For the risk component, use Formula 8.10. A risk-free rate, such as the 4% Treasury rate, has neither a variance nor a covariance with anything else. (Makes sense that a fixed constant number that is always the same has no variance, doesn't it?) For portfolio S, use $(1 - w_H) = w_F$ and you get

$$\mathcal{V}ar(\tilde{r}_S) = (1/2)^2 \cdot 90\%\% + (1 - 1/2)^2 \cdot 0\% + 2 \cdot 1/2 \cdot (1 - 1/2) \cdot 0\% = 1/4 \cdot 90\%\%$$
$$\mathcal{V}ar(\tilde{r}_S) = w_H^2 \cdot \mathcal{V}ar(\tilde{r}_H) + w_F^2 \cdot \mathcal{V}ar(r_F) + 2 \cdot w_H \cdot w_F \cdot \mathcal{C}ov(\tilde{r}_H, r_F) = (w_H)^2 \cdot \mathcal{V}ar(\tilde{r}_H)$$

This formula is a lot simpler than the typical variance formula, with its second variance term and its covariance term. It also means that we can compute the standard deviation more easily:

$$\mathcal{S}dv(\tilde{r}_S) = \sqrt{(1/2)^2 \cdot 90\%\%} = 1/2 \cdot \sqrt{90\%\%} \approx 1/2 \cdot 9.49\% \approx 4.74\% \qquad (8.14)$$
$$\mathcal{S}dv(\tilde{r}_S) = \sqrt{(w_H)^2 \cdot \mathcal{V}ar(\tilde{r}_H)} = w_H \cdot \sqrt{\mathcal{V}ar(\tilde{r}_H)} = w_H \cdot \mathcal{S}dv(\tilde{r}_H)$$

This states that the risk of your overall portfolio is proportional to the risk of your investment in asset H, with your investment weight being the proportionality factor.

Combining different weights of any risky portfolio with the risk-free asset yields a straight line.

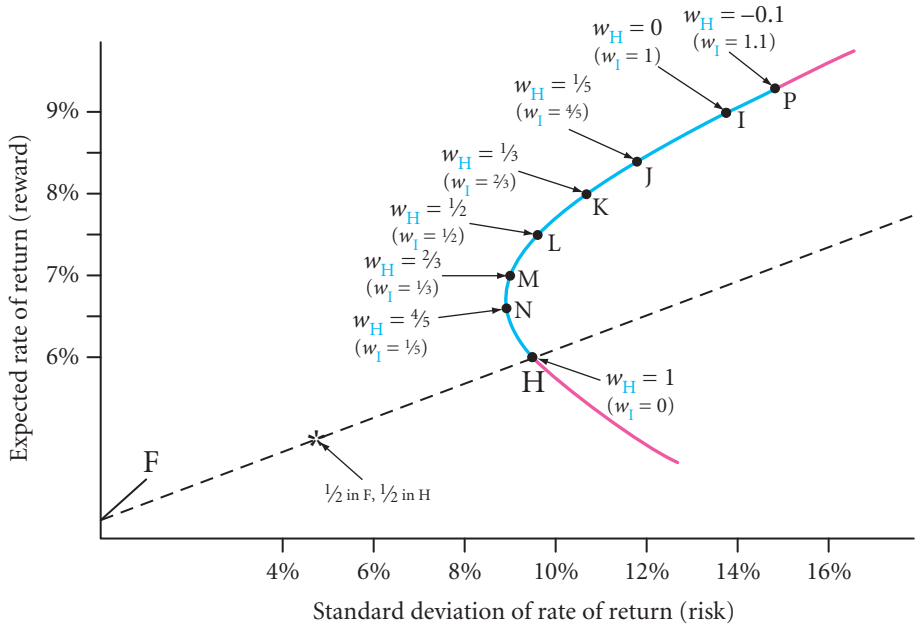
You can repeat this for many different portfolio weights:

Weight w_H	0.0	0.2	0.4	0.6	0.8	1.0
Expected Return	4.0%	4.4%	4.8%	5.2%	5.6%	6.0%
Standard Deviation	0.000%	1.898%	3.796%	5.694%	7.592%	9.490%

The algebra that shows that the relation between the risk and reward of a risky portfolio and the risk-free asset is a line.

If you plot these points into the figure, you will immediately notice that the relationship between risk and reward is now a line. Figure 8.7 does it for you.

You can also show this algebraically. Rearrange Formula 8.14 into $w_H = \mathcal{S}dv(\tilde{r}_S)/\mathcal{S}dv(\tilde{r}_H) = \mathcal{S}dv(\tilde{r}_S)/9.49\%$. Then use this to substitute out w_H in Formula 8.13:



This adds a risk-free rate of 4% to Figure 8.5. The line represents risks and rewards for portfolios that combine portfolio H and the risk-free rate F. Please note that this line is *not* the security markets line (the CAPM). Here, the *x*-axis is the standard deviation (of the overall portfolio rate of return). In the security market line (SML) explained in chapter 9, the *x*-axis is the market beta (of individual assets).

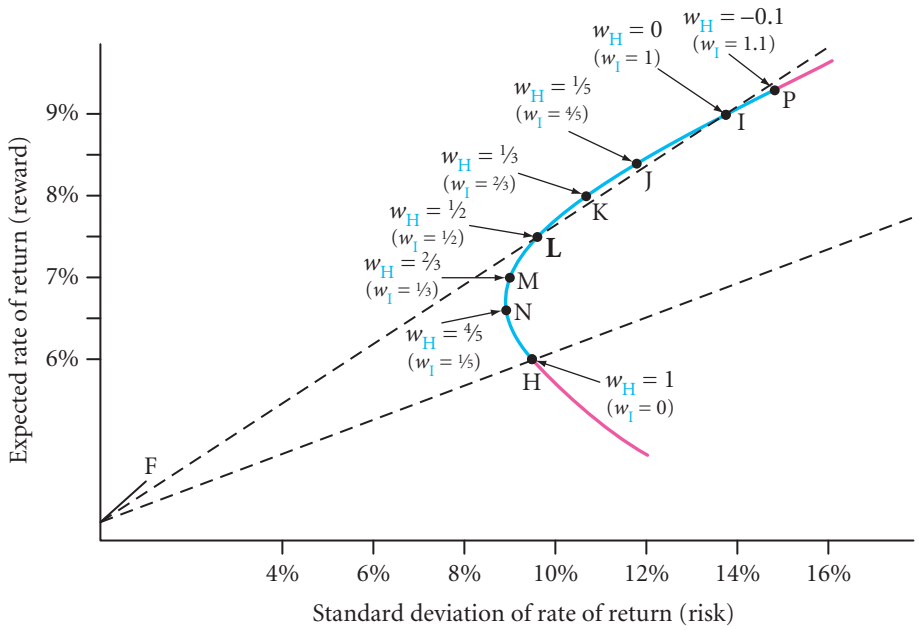
FIGURE 8.7 The Risk-Reward Trade-Off between H and F

$$\begin{aligned}
\mathcal{E}(\tilde{r}_S) &= w_H \cdot 6\% + (1 - w_H) \cdot 4\% &&= w_H \cdot (6\% - 4\%) + 4\% \\
&= \left[\frac{\mathcal{S}dv(\tilde{r}_S)}{9.49\%}\right] \cdot (6\% - 4\%) + 4\% = &&4\% + 0.21 \cdot \mathcal{S}dv(\tilde{r}_S) \\
\mathcal{E}(\tilde{r}_S) &= w_H \cdot \mathcal{E}(\tilde{r}_H) + (1 - w_H) \cdot r_F &&= w_H \cdot (\mathcal{E}(\tilde{r}_H) - r_F) + r_F \\
&= \left[\frac{\mathcal{S}dv(\tilde{r}_S)}{\mathcal{S}dv(\tilde{r}_H)}\right] \cdot [\mathcal{E}(\tilde{r}_H) - r_F] + r_F = r_F + \left[\frac{\mathcal{E}(\tilde{r}_H) - r_F}{\mathcal{S}dv(\tilde{r}_H)}\right] \cdot \mathcal{S}dv(\tilde{r}_S)
\end{aligned}$$

This is the formula for a line: r_F is the intercept and $\left[(\mathcal{E}(\tilde{r}_H) - r_F)/(\mathcal{S}dv(\tilde{r}_H))\right]$ is the slope.

IMPORTANT: When you plot the portfolio mean versus the portfolio standard deviation for combination portfolios of a risk-free asset F with any risky portfolio P, they lie on the straight line between F and P.

But would you really want to purchase such a combination of H and F? Could you purchase a different portfolio in combination with F that would do better? Would the combination of L and F not perform better?



Adding to Figure 8.7, the new line represents risks and rewards for portfolios that combine portfolio L and the risk-free asset F.

FIGURE 8.8 The Risk-Reward Trade-Off between L and F

Figure 8.8 draws combinations of the risk-free asset and portfolio L. This combination of F and L indeed does a lot better—but you can do even better yet. Can you guess what portfolio you would purchase?

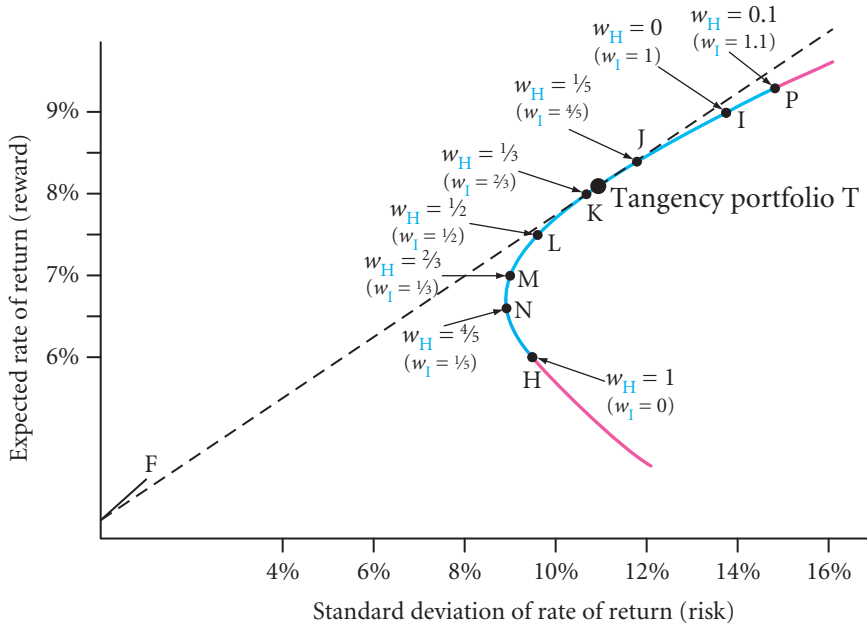
With a risk-free asset, the best portfolio is the line that is tangent to the efficient frontier of risky assets.

The answer is drawn in Figure 8.9—you would purchase a combination portfolio of the risk-free asset and whatever portfolio on the previous efficient frontier would be *tangent*—you tilt the line up until it just touches the mean-variance frontier among the risky assets. This line is called the **capital market line**. Here, the exact investment proportions in the risky assets are difficult to see, but if you could blow up the figure, you would see that this is the portfolio that invests about 30% in H and 70% in I. Let's call it T, for tangency portfolio.

Do all smart investors make the same portfolio decision in the presence of a risk-free asset? Yes and no.

Who would want to purchase a portfolio combination that invests more or less than 30% in H and 70% in I? Nobody! Each and every smart investor would purchase only a combination of F and T, regardless of risk aversion. (This is called the **two-fund separation theorem**.) Different risk tolerances would lead them to allocate different sums to the tangency portfolio and the risk-free asset, but no investor would purchase a risky portfolio with investment weights different from those in the tangency portfolio T.

IMPORTANT: In the presence of a risk-free asset, all smart investors purchase combinations of the tangency portfolio and the risk-free asset.



The capital market line represents risks and rewards for portfolios that combine the tangency portfolio T and the risk-free rate F. It represents the best opportunities available.

FIGURE 8.9 The Risk-Reward Trade-Off between T and F

CAPM Preview

Chapter 9 explains the most common model of security pricing, the CAPM. In brief, it states that the market portfolio is mean-variance efficient—and nothing else. How can this happen? Well, if *every* investor is smart and all the various CAPM assumptions and conditions are satisfied (explained soon), then each investor holds only a combination of T and the risk-free asset. Math dictates that this means that the value-weighted market portfolio of all investors' holdings is therefore also a combination of T and the risk-free asset. Therefore, it is also mean-variance efficient.

In equilibrium, if all investors buy combinations of the risk-free asset and the tangency portfolio, the market portfolio is on the efficient frontier.

IMPORTANT: In the CAPM, the market portfolio of risky claims is the tangency portfolio.

(Of course, conversely, if some investors do not hold the market tangency portfolio, then the overall market portfolio [could but] need not be the tangency portfolio.)

If the CAPM holds, that is, if T is the market portfolio, then portfolio optimization is beautifully easy for any investor—just purchase a combination of the market portfolio and the risk-free asset. You never even need to compute an efficient frontier. Of course, in the real world, the market portfolio may not be the tangency portfolio—but then, this is the same as stating that the CAPM does not hold. *In fact, the CAPM is nothing more and nothing less than the statement that the market portfolio is the tangency portfolio.*

The CAPM can make investing really easy—no computer program necessary!

SOLVE NOW!

Q 8.40 What kind of portfolios are the points to the right of H on the line itself in Figure 8.7?

Q 8.41 Compute the covariance of H and F.

Q 8.42 Formula 8.11 noted that the minimum-variance portfolio without a risk-free asset invests about 76.2% in H and about 24.8% in I. (Work with the rounded numbers to make your life easier.) With the risk-free asset offering 4%, what portfolio would you purchase that has the same risk, and what would its improvement in reward be? First think about how to solve this. However, this is a difficult question, so we will go through it step by step.

(a) Copy down the risk of this minimum-variance portfolio when there is no risk-free asset.

(b) What is the reward of this minimum-variance portfolio?

(c) With a risk-free rate of 4%, it turns out that the tangency portfolio invests 30% in H and 70% in I. What are its returns in each of the four scenarios?

(d) What is its reward? (Check this visually in the graph!)

(e) What is its risk? (Check this visually in the graph!)

(f) Using the analog of Formula 8.14, what investment weight w_T in T would give you the same risk as the minimum-variance portfolio? (If you had \$100, how much would you put into T, and how much would you put into a risk-free savings account?)

(g) Given this weight w_T, what is the reward of this combination portfolio? How much better is this than the situation where no risk-free asset was available?

Q 8.43 Would the tangency portfolio invest in more or less H if the risk-free rate were 3% instead of 4%? (Hint: Think visually.)

KEY TERMS

capital market line, 242
on margin, 246
mean-variance efficient frontier, 236
minimum-variance portfolio, 236
MVE frontier, 236
sharpe ratio, 236
two-fund separation theorem, 242

SOLVE NOW! SOLUTIONS

Q 8.31 The rates of return of portfolio M in Table 8.4 are −8% (♣), +14% (♦), 8% (♥), and 14% (♠). The deviations from the mean are −15%, 7%, 1%, and 7%. When squared, they are 225%%, 49%%, 1%%, and 49%%. The sum is 324%%; the average is 81%%. Thus, the standard deviation is indeed 9%.

Q 8.32 The portfolio variance of portfolio N in Table 8.4 is

$$\begin{aligned}\mathcal{S}dv(\tilde{r}_{\mathrm{H}}) = \sqrt{\mathcal{V}ar(\tilde{r}_{\mathrm{H}})} &= \sqrt{\frac{(-7.5\% - 6.75\%)^2 + (13.5\% - 6.75\%)^2 + (6\% - 6.75\%)^2 + (15\% - 6.75\%)^2}{4}} \\ &= \sqrt{\frac{203.0625\%\% + 45.5625\%\% + 0.5625\%\% + 68.0625\%\%}{4}} \\ &\approx \sqrt{79.31\%\%} \approx 8.91\%\end{aligned}$$

Q 8.33 For M, the covariance between H and I was computed as 45%% in Formula 8.9. The variance of H is 90%% (from Table 8.4 on page 232), the variance of I is 189%% (from the same figure). Therefore, using the shortcut Formula 8.10, $\mathcal{V}ar(\tilde{r}_{\mathrm{M}}) = (2/3)^2 \cdot 90\%\% + (1/3)^2 \cdot 189\%\% + 2 \cdot (2/3) \cdot (1/3) \cdot 45\%\% = 81\%\%$.

Q 8.34 The covariance between H and I is 45%% (Formula 8.9). The variance of H is 90%%, the variance of I is 189%% (Table 8.4). Therefore, the shortcut Formula 8.10 gives

$$\mathcal{V}ar(\tilde{r}_{\mathrm{M}}) = (3/4)^2 \cdot 90\%\% + (1/4)^2 \cdot 189\%\% + 2 \cdot (3/4) \cdot (1/4) \cdot 45\%\% = 79.3125\%\%$$

Q 8.35 This is an important question. In fact, you should memorize Formula 8.15 that describes how risk grows over time. The assumption that there is no compounding (that you can ignore the cross-product) and that risk is roughly constant per period is reasonable over periods that are not more than a few years long.

(a) If we can ignore the cross-products, then we are using a simple weighted-average formula with weights of 1 on each term: $\tilde{r}_{0,2} \approx 1 \cdot \tilde{r}_{0,1} + 1 \cdot \tilde{r}_{1,2}$. (The exact formula would have been $\tilde{r}_{0,2} = \tilde{r}_{0,1} + \tilde{r}_{1,2} + \tilde{r}_{0,1} \cdot \tilde{r}_{1,2}$.)

(b) The expected rate of return over 2 years is $\mathcal{E}(\tilde{r}_{0,2}) \approx \mathcal{E}(\tilde{r}_{0,1}) + \mathcal{E}(\tilde{r}_{1,2}) = 12\% + 12\% = 24\%$.

(c) The variance of the rate of return over 2 years is $\mathcal{V}ar(\tilde{r}_{0,2}) \approx 1 \cdot \mathcal{V}ar(\tilde{r}_{0,1}) + 1 \cdot \mathcal{V}ar(\tilde{r}_{1,2}) + 2 \cdot 1 \cdot 1 \cdot \mathcal{C}ov(\tilde{r}_{0,1}, \tilde{r}_{0,2})$. In a perfect market, the last term should be approximately zero.

(d) The variance over 2 years for our specific example is

$$\begin{aligned}\mathcal{V}ar(\tilde{r}_{0,2}) &\approx 1 \cdot \mathcal{V}ar(\tilde{r}_{0,1}) + 1 \cdot \mathcal{V}ar(\tilde{r}_{1,2}) + \quad 0 \\ &= \quad (20\%)^2 \quad + \quad (20\%)^2 \quad = 2 \cdot (20\%)^2 = 800\%\%\end{aligned}$$

Therefore, the standard deviation is $\sqrt{2} \cdot 20\% \approx 28\%$.

(e) The Sharpe ratio is $2 \cdot (12\% - 6\%)/28\% \approx 0.43$.

(f) The variance is $4 \cdot (20\%)^2 = 1600\%\%$. The standard deviation is $20\% \cdot \sqrt{4} = 40\%$. The Sharpe ratio is $(6\% \cdot 4)/(20\% \cdot \sqrt{4}) = 0.3 \cdot \sqrt{4} = 0.6$.

(g) The variance is $16 \cdot (20\%)^2 = 6400\%$. The standard deviation is $20\% \cdot \sqrt{16} = 80\%$. The Sharpe ratio is $0.3 \cdot \sqrt{16} \approx 1.2$.

(h) The variance is $T \cdot (20\%)^2$. The standard deviation is $20\% \cdot \sqrt{T}$. In other words, the standard deviation grows with the square root of the number of time periods:

IMPORTANT: How asset risk grows with time: $\mathcal{S}dv(\tilde{r}_{0,T}) \approx \sqrt{T} \cdot \mathcal{S}dv(\tilde{r}_{0,1})$ (8.15)

If the rates of return on an asset are approximately uncorrelated over time (a perfect market consequence), if the risk in different time periods remains constant, and ignoring all cross-product terms. The Sharpe ratio is $0.3 \cdot \sqrt{T}$.

(i) The formulas also work with fractions. The variance is therefore $1/12 \cdot (20\%)^2 \approx 33.3\%\%$. The standard deviation is therefore $\sqrt{1/12} \cdot 20\% \approx 5.8\%$. The monthly Sharpe ratio is $\sqrt{1/12} \cdot 30\% \approx 0.09$.

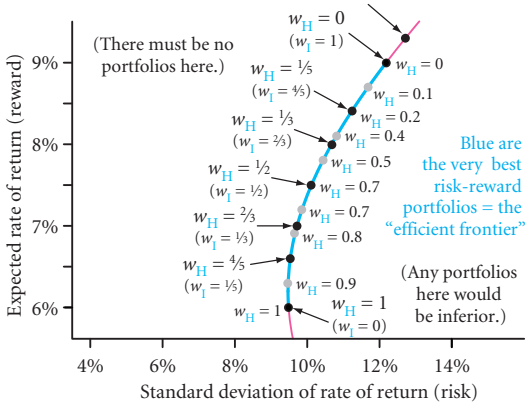
(j) The variance is $1/250 \cdot (20\%)^2 = 1.6\%\%$. The standard deviation is $\sqrt{1/250} \cdot 20\% \approx 1.3\%$. The daily Sharpe ratio is about 0.019.

Q 8.36 The mean rate of return for portfolio ($w_H = 0.1$, $w_I = 0.9$) is $0.1 \cdot 6\% + 0.9 \cdot 9\% = 8.7\%$. You can also compute this from the rates of return in the 4 states -11.4%, 17.4%, 21.6%, and 7.2%. Demeaned, these returns are -20.1%, 8.7%, 12.9%, and -1.5%. The variance is therefore $(404.01\%\% + 75.69\%\% + 166.41\%\% + 2.25\%\%)/4 = 162.09\%\% = 0.016209$. Therefore, the standard deviation (risk) is $\sqrt{162.09\%\%} \approx 12.7\%$.

Q 8.37 Two risky portfolios with a correlation of -1 can be combined into an asset that has no risk. Thus, its expected rate of return has to be the same as that on the risk-free asset—or you could get rich in a perfect market.

Q 8.38 If the correlation was higher, diversification would help less, so the risk would be higher. Therefore, the efficient frontier would not bend as far toward the west (a risk of 0). An easy way to check this is to rearrange the returns so that they correlate more positively, as you will do in the next question. If the correlation was lower, diversification would help more, so the risk would be lower. Therefore, the efficient frontier would bend closer toward the west (a risk of 0).

Q 8.39 The covariance between H and Z is $85.5\%\%$, which is much higher than the $45\%\%$ covariance between H and I from Formula 8.9 on page 233. This means that the correlation between H and Z shoots up to 74% (from 35% for the correlation between H and I). This means that the efficient frontier is less dented toward the west. Put differently, the minimum-variance portfolio moves toward the east.



Q 8.40 Portfolios to the right of H on the line have a negative weight in F and a weight above 1 in H. (The portfolio weights must add to 100%!) This means that they would borrow money at a 4% annual interest rate to purchase more of portfolio H. (Purchasing stocks with money borrowed at an interest rate is called **on margin**.)

Q 8.41 Because the net-of-mean F is always 0, so is its coproduct with anything else. This means that the covariance of the risk-free asset with any risky asset is zero, too.

Q 8.42 This question asks you to show how much better off you are with this particular risk-free asset for a particular risk choice.

(a) In Formula 8.12 on page 237, we showed that this no-risk-free minimum-variance portfolio with an investment weight of 76.2% in H and 24.8% in I has a risk of about 8.90%.

(b) The reward of this no-risk-free-asset-available, minimum-variance portfolio is $\mathcal{E}(\tilde{r}) = 76.2\% \cdot 6\% + 24.8\% \cdot 9\% \approx 6.8\%$.

(c) With a weight of 30% in H and 70% in I, the rates of return in the four scenarios for the tangency portfolio T are as follows:

$$\begin{aligned}
\text{In Scenario } \clubsuit\text{:} \quad & 0.3 \cdot (-6\%) + 0.7 \cdot (-12\%) = -10.2\% \\
\text{In Scenario } \diamondsuit\text{:} \quad & 0.3 \cdot (12\%) + 0.7 \cdot (18\%) = +16.2\% \\
\text{In Scenario } \heartsuit\text{:} \quad & 0.3 \cdot (0\%) + 0.7 \cdot (24\%) = +16.8\% \\
\text{In Scenario } \spadesuit\text{:} \quad & 0.3 \cdot (18\%) + 0.7 \cdot (6\%) = +9.6\%
\end{aligned}$$

(These calculations will reappear later in Table 9.2 on page 290.)

(d) The reward of the tangency portfolio is $\mathcal{E}(\tilde{r}_T) = (-10.2\% + 16.2\% + 16.8\% + 9.6\%)/4 = 8.1\%$.
(e) Its risk is $\mathcal{S}dv(\tilde{r}_T) = \sqrt{[(-18.3\%)^2 + (8.1\%)^2 + (8.7\%)^2 + (1.5\%)^2]/4} \approx 10.94\%$.
(f) You want the expected rate of return of a portfolio that uses the risk-free asset and that has a risk of 10.94% (i.e., the same that the no-risk minimum-variance portfolio had). Solve

$$8.9\% = w_T \cdot 10.94\%$$
$$\mathcal{S}dv(\tilde{r}) = w_T \cdot \mathcal{S}dv(\tilde{r}_T)$$

Therefore, $w_T \approx 81.35\%$. In words, a portfolio of 81.35% in the tangency portfolio T and 18.65% in the risk-free asset F has the same risk of 10.94%.
(g) You now want to know the expected rate of return on the portfolio $(w_T, w_F) = (81.35\%, 18.65\%)$:

$$\mathcal{E}(\tilde{r}) \approx 81.35\% \cdot 8.1\% + 18.65\% \cdot 4\% \approx 7.33\%$$
$$\mathcal{E}(\tilde{r}) = w_T \cdot \mathcal{E}(\tilde{r}_T) + w_F \cdot r_F$$

You therefore would expect to receive a $7.33\% - 6.71\% \approx 62$ basis points higher expected rate of return if you have access to this risk-free rate.

Q 8.43 If the risk-free rate were lower, then the tangency line would become steeper. The tangency portfolio would shift from around K to around L. Therefore, it would involve more H.

PROBLEMS

The [ ] *indicates problems available in* myfinancelab

Q 8.44 Recompute the portfolio variance if you invest in a portfolio O with $w_H = 90\%$ and $w_I = 10\%$ in Table 8.4.
(a) Compute the rates of return on the portfolio in each scenario, and then treat the resulting portfolio as one asset. What is portfolio O's risk and reward?
(b) Compute the same variance with the shortcut Formula 8.10 on page 234.

Q 8.45 An asset has an annual mean of 12% and standard deviation of 30% per year. What would you expect its monthly mean and standard deviation to be?

Q 8.46 Mathematically and based on Figure 8.6 on page 238, the risk and reward of the portfolio $w_H = -0.2$, $w_I = -1.2$.

Q 8.47 In the absence of a risk-free asset, would anyone buy the portfolio $w_H = 110\%$, $w_I = -10\%$?

Q 8.48 The Vanguard European stock fund, Pacific stock fund, and Exxon Mobil reported the following historical dividend-adjusted prices:

Year	1991	1992	1993	1994	1995	1996
VEURX	6.53	7.15	6.91	9.34	9.03	11.17
VPACX	7.18	7.41	6.30	9.52	9.08	9.97
XOM	9.57	10.07	10.88	10.97	15.29	19.18

Year	1997	1998	1999	2000	2001
VEURX	13.50	17.45	21.42	23.38	23.13
VPACX	8.39	7.17	7.01	10.41	8.10
XOM	24.63	30.14	33.94	37.42	34.57

Year	2002	2003	2004	2005	2006
VEURX	17.50	14.42	21.22	24.87	29.53
VPACX	5.64	5.42	7.94	9.08	11.93
XOM	31.50	38.01	48.67	54.41	75.67

(a) Compute the means and covariances of the rates of return on these three assets.
(b) Draw the efficient frontier if you can only invest in VEURX and VPACX.

(c) Now add Exxon Mobil. Use Excel to draw 1,000 random numbers in two columns, called wE and wP. (Create one formula, and copy it into all of the cells.) Each of these 2,000 cells should use the formula 'rand()*3-1'. Create a new column that is 1.0 minus wE and wP, and call it wX. Now consider these random numbers as investment weights in VEURX, VPACX, and XOM. Compute the risk and reward for each of these portfolios (one portfolio is three numbers: one wE, one wP, and one wX), using the standard deviation and expected rate of return formulas. Finally, create an x-y plot that shows, for each of your wE, wP, and wX portfolios, the risk-reward combinations. What does the plot look like?

(d) If the risk-free rate stood at 5% per annum, what would be the tangency portfolio?

Q 8.49 Return to the example with a risk-free asset in Formula 8.14 on page 240. What are the risk and reward of a portfolio that invests $w_H = 150\%$? (This means that if you have \$100, you would borrow \$50 at the 4% annual interest rate to purchase \$150 of H—more than your portfolio wealth itself.)

Economist Intelligence Unit

The Economist

Regulatory risk

Trends and strategies for the CRO



A report from the Economist Intelligence Unit

Sponsored by:
ACE, Cisco Systems, Deutsche Bank, KPMG and IBM



Preface

Regulatory risk: Trends and strategies for the CRO is the second in a series of reports from the Economist Intelligence Unit's Global Risk Briefing, a research programme targeted at senior executives responsible for managing corporate risk. Alasdair Ross was the author of the report, and Gareth Lofthouse was the editor. The Global Risk Briefing is sponsored by ACE, Cisco Systems, Deutsche Bank, KPMG and IBM.

The research for this paper is based on a survey of 230 senior risk managers, as well as interviews with chief risk officers (CROs). The Economist Intelligence Unit bears sole responsibility for the content of this report.

Our thanks are due to everyone who shared their time and insights in the survey and interviews. We are also grateful to the Professional Risk Managers International Association (PRMIA) for their help in reaching CROs and other risk managers to support this research.

July 2005

© The Economist Intelligence Unit 2005



Executive summary

Regulatory risk, a term describing the problems arising from new or existing regulations, is now one of the greatest threats to business, according to a global survey of 230 senior risk managers by the Economist Intelligence Unit. In the eyes of many corporate leaders, regulatory risk is now a greater source of concern than country risk, market and credit risk, IT and people risks, or terrorism and natural disasters.

How did regulation, much of which is designed to reduce business risks, become a major source of risk in its own right? Most companies accept the need for rules to govern business, and are used to working within regulatory constraints. But a spate of new regulations in recent years has had major—and some would argue unforeseen—consequences for business. For companies with international operations, in particular, the cost and complexity of ensuring compliance have risen sharply. So have the penalties, direct and indirect, of non-compliance.

Anticipating how a patchwork of domestic and international regulations may be interpreted, changed and augmented has become a key challenge for senior risk managers. This report, written by the Economist Intelligence Unit and sponsored by ACE, Cisco Systems, Deutsche Bank, KPMG and IBM, focuses on the challenge of managing these regulatory risks. Drawing on a global survey of executives from a cross-section of industries, as well as interviews with several chief risk officers (CROs), this research sheds light on why regulation has become a major source of risk, and on what professional risk managers are doing about it. The report includes the following main findings:

- **The cost of regulation is rising.** Nearly 90% of companies believe that the cost of regulation will rise over the next three years; 36% of them believe it will rise substantially. This is bad news for companies, but also for their customers who, according to almost two-thirds of executives in the survey, will ultimately foot the bill for compliance. Regulation has other, indirect consequences for business. Thirty-six percent of executives say that regulation has stifled innovation in their companies. Just under one-third fear their firms will become uncompetitive with companies operating in less regulated countries. Executives acknowledge the need for regulation, but the majority believe the benefits of recent regulations are outweighed by the problems.

- **Regulatory risks increase as firms expand overseas.** Increased complexity in the international regulatory environment is one of the biggest challenges facing companies, with regulations in one country potentially having an increased impact on firms' global operations. Only 17% of companies are very confident that they are compliant with regulation in their overseas markets, compared with 40% that are very confident of compliance in their home markets. Similarly, only one in five companies have high confidence that they are compliant with international regulations such as Basel II and International Accounting Standards.

- **Regulation is a significant deterrent to overseas investment.** Fully 57% of executives in the survey say that their perception of a country's regulatory burden has an important or very important impact on decisions to invest, compared with only 4% that believe it is of low importance. Over one-third of companies have been deterred from investing in a new market because of regulatory issues. The results have interesting implications for the US, which is seen by executives as the country where regulatory burdens are increasing most significantly.

© The Economist Intelligence Unit 2005



● **Companies must seek to stay one step ahead of the regulators**. Executives believe the two best strategies for pre-empting regulatory change are to adopt best practice before it is mandated, and to maintain regular communications with domestic and international regulators. A proactive approach can help avert the need for regulation in the first place, and minimise the disruption caused by new regulation where it arises. With companies struggling to maintain compliance with regulations in multiple jurisdictions, recruiting expertise in different countries' regulatory environments is also seen as an important strategy. Direct lobbying of governments is deemed far less effective as a strategy for managing regulatory risk.

After numerous high-profile failures in corporate governance in recent years, few would argue with the need for strong and effective regulation. But many business leaders are concerned that recent regulations have brought unintended consequences for business. As one CRO interviewed for this report puts it, "The main regulatory risk is that, failing to understand our business, the regulator does something that damages us." That makes it all the more imperative that companies work closely with regulators at home and abroad, helping them to set standards for good governance without the need for excessively onerous regulation.

New face of an old challenge

The tension between private business and the public interest is as old as commerce. Back in the days, a century ago, when the Rockefellers, the Carnegies and the Schwabs dominated the US economy from the helm of virtual cartels known as business trusts, the state began to take direct action by introducing "anti-trust" mechanisms. Trust-busting became the model for a generation of legislation that curbed companies' ability to exploit an advantageous position in the marketplace. As the 20th century began, any scion of America's great entrepreneurial families would have immediately grasped the meaning of the term regulatory risk.

Of course, regulation remains a hot topic for companies today. In the past decade, a succession of market failures put policymakers on the spot. Financial crises from Brazil to Thailand demanded more reliable mechanisms for steering international financial flows. Power shortages in California demonstrated the shortfalls in existing regulatory frameworks. Governance failures such as those at Enron and WorldCom focused attention on the complexities of the modern conglomerate, and the inequities in the share-out of risks and rewards between corporate officers and stakeholders.

The response of regulators to these developments has been energetic. Regulatory rules now cover every aspect of a business and its relationship with society, from business competition to financial management, from health and safety to conservation of the natural environment.

There is nothing new in companies being compelled to operate within regulatory constraints, and most executives recognise the need for rules to govern business. Recently, however, many executives have taken issue with the manner in which regulation is designed and implemented, and particularly with the trend in recent years towards a more intrusive, more comprehensive and more prescriptive approach. In a world of global brands and pervasive media, market abuses have the potential to elicit powerful public reactions, focusing pressure on governments and state agencies to respond with equally emphatic counter-measures.

The most powerful example is the US's Sarbanes-Oxley act of 2002, compelling company directors to provide evidence of probity on a range of issues on pain of imprisonment. Few question that legislation was needed, and the earliest versions of the bill were broadly supported by business, but what eventually emerged took regulation of corporate behaviour to new levels—

© The Economist Intelligence Unit 2005



perhaps unintentionally. Michael Oxley, one of the sponsors of the legislation, describes the post-WorldCom atmosphere in which the legislation was drafted as "a hothouse", and now acknowledges that the legislation that bears his name is imperfect as a result.

Erwin Martens, CRO at TIAA-CREF, a New York-based financial institution, concurs. "Sarbanes-Oxley has hit both the bottom line and the business structure. Businesses are asking 'Is this what you meant to do?' and regulators are saying 'Well no, but this is what happened'."

There is a feeling among many executives that, in their haste to prevent another WorldCom or Enron, the regulators have rushed in far-reaching regulations without fully comprehending their impact on business. "The main regulatory risk is that, failing to understand our business, the regulator does something that damages us," says Paul de Hoest, CRO for Egg Bank Plc in London. "So it is all about identifying and sharing our objectives."

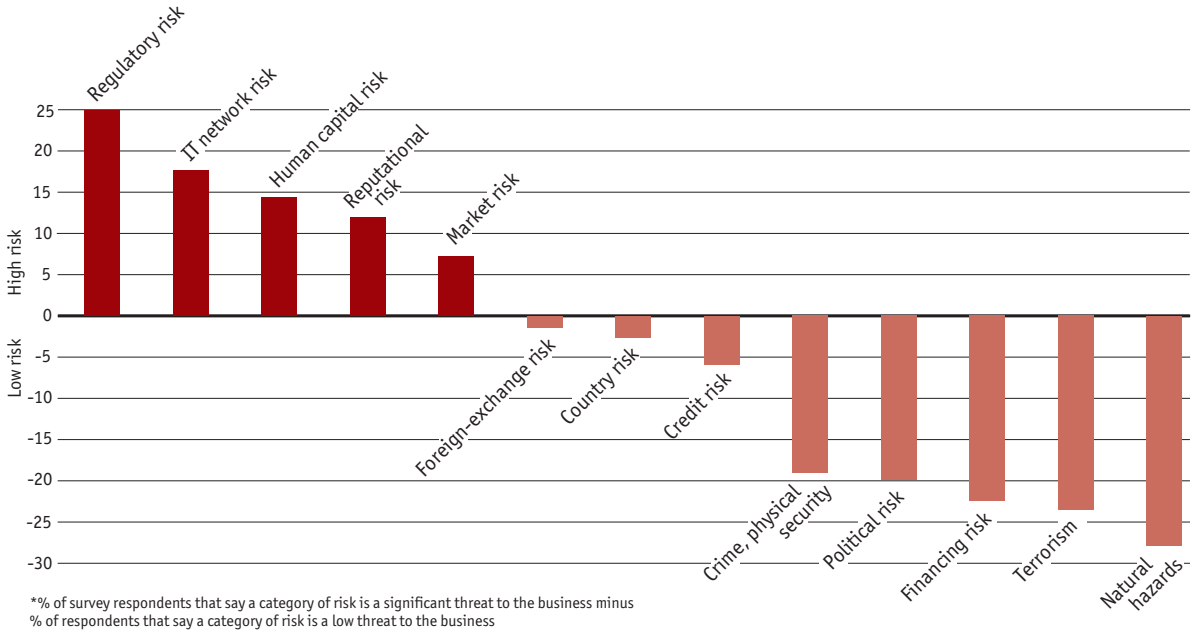
Mr Hoest is not alone in seeing regulation as a significant source of risk in its own right. The Economist Intelligence Unit's Risk Barometer (an index that tracks major business threats on a quarterly basis) now shows that regulatory risk is seen by executives as the most significant threat to business, ahead of country risk, market and credit risk, IT and people risks, or terrorism and natural disasters.

As to whether the regulators are achieving their stated aims, many CROs are sceptical. "Risk is like a balloon, you squeeze in one place and it just comes up somewhere else," says Mr de Hoest. Certainly the majority of global executives surveyed for this report believe that the problems arising from recent regulations outweigh any benefits.

Complexity equals cost

As we have seen, companies have had to come to terms with a raft of new regulations in recent years. But as much as the absolute level of regulation, it is the pace of change that causes managers headaches. "What's

Category risk barometer
Average % score for each category*



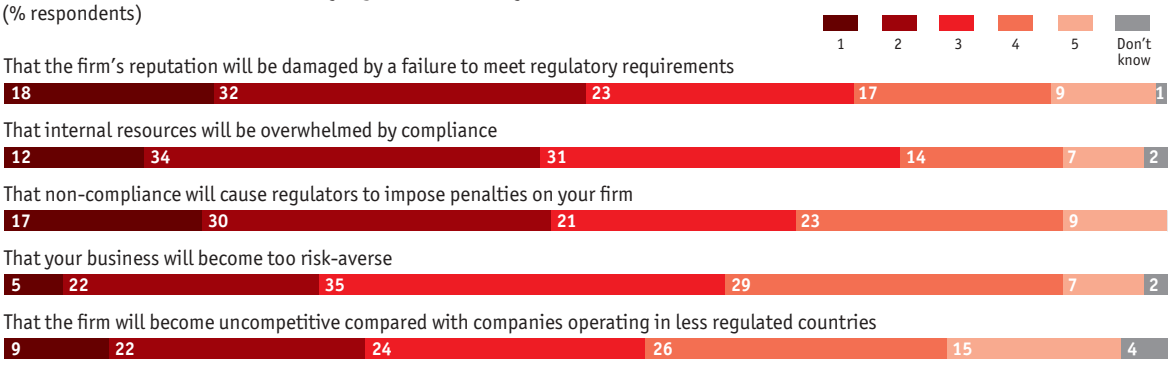
*% of survey respondents that say a category of risk is a significant threat to the business minus % of respondents that say a category of risk is a low threat to the business

Source: Economist Intelligence Unit survey, June 2005

© The Economist Intelligence Unit 2005



How significant are the following regulatory risks to your business?
Rate on a scale of 1 to 5, where 1=Very high risk and 5=Very low risk.
(% respondents)



Source: Economist Intelligence Unit survey, June 2005

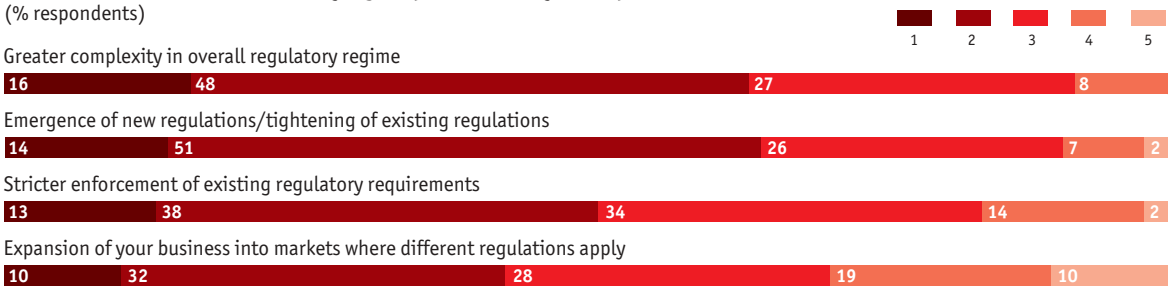
unique at this time is the level of flux," observes Mr Martens.

Many large companies have ended up juggling several compliance projects, and find it increasingly difficult just to keep pace with a fast-changing and unpredictable regulatory environment. Executives in the survey see rising complexity in the regulatory regime as the biggest source of regulatory risk, followed closely by the emergence of new regulations and the tightening of existing ones. Half the survey respondents also see stricter enforcement as another trend that will increase the burden of compliance over the coming years, while 42% see expansion into new markets as another likely source of compliance costs.

The survey indicates that regulatory compliance is by no means assured, particularly for companies with global business operations. Only 17% of companies are very confident that they are compliant with regulation in their overseas markets, compared with 40% that are very confident that they are compliant with regulations in their home markets. Moreover, international regulations, such as International Accounting Standards and Basel II, add an extra layer of complexity. Although most companies feel broadly on track to meet the requirements of international regulation, only one in five companies are very confident in this regard.

Executives are keenly aware that a breach of regulatory rules could bring severe consequences that stretch far beyond any direct penalty imposed by the enforcement agencies. More than anything else, executives fear the impact that a failure to meet

How significant an impact do you think each of the following will have on your business over the next three years?
Rate on a scale of 1 to 5, where 1=Very high impact and 5=Very low impact.
(% respondents)



Source: Economist Intelligence Unit survey, June 2005

© The Economist Intelligence Unit 2005



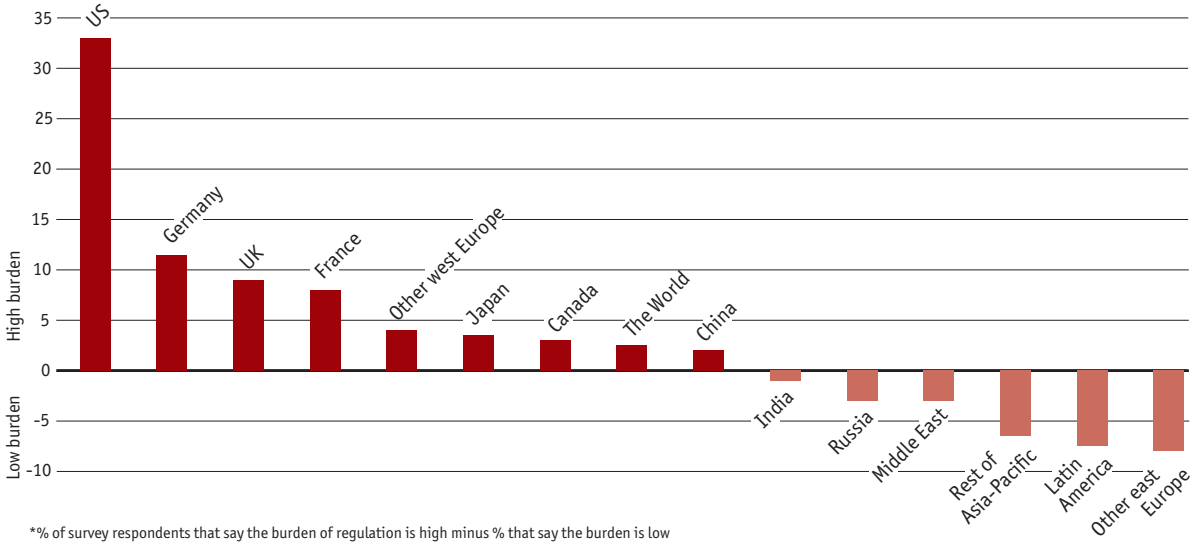
Regulatory risks: a global perspective

In addition to shedding light on the costs and consequences of regulation, the survey also reveals the countries where regulatory tightening is having the greatest impact on business. Topping this list by a huge margin, the US stands out as the country creating the greatest source of regulatory risk for the majority of companies in the survey.

Given that historically business regulation has been relatively light in the US, this result probably signifies a strong reaction to Sarbanes-Oxley, as well as the importance of the US market to companies in the survey. Sarbanes-Oxley has been in force for domestic corporations for a year; foreign-based firms will need to comply with the legislation by the end of 2006.

There is then a big gap between the US and the second-placed country, Germany, which around one-third of respondents believe imposes a high regulatory burden. The UK and France follow closely. Other west European markets, Japan, Canada and China raise some concern and, at the bottom of the ranking, east European countries (excluding Russia) and Latin American ones are perceived as relatively light regulators.

Do you view the current regulatory environment as placing a high or low burden on business in the following countries or regions?
Average % score by region*



*% of survey respondents that say the burden of regulation is high minus % that say the burden is low
Source: Economist Intelligence Unit survey, June 2005

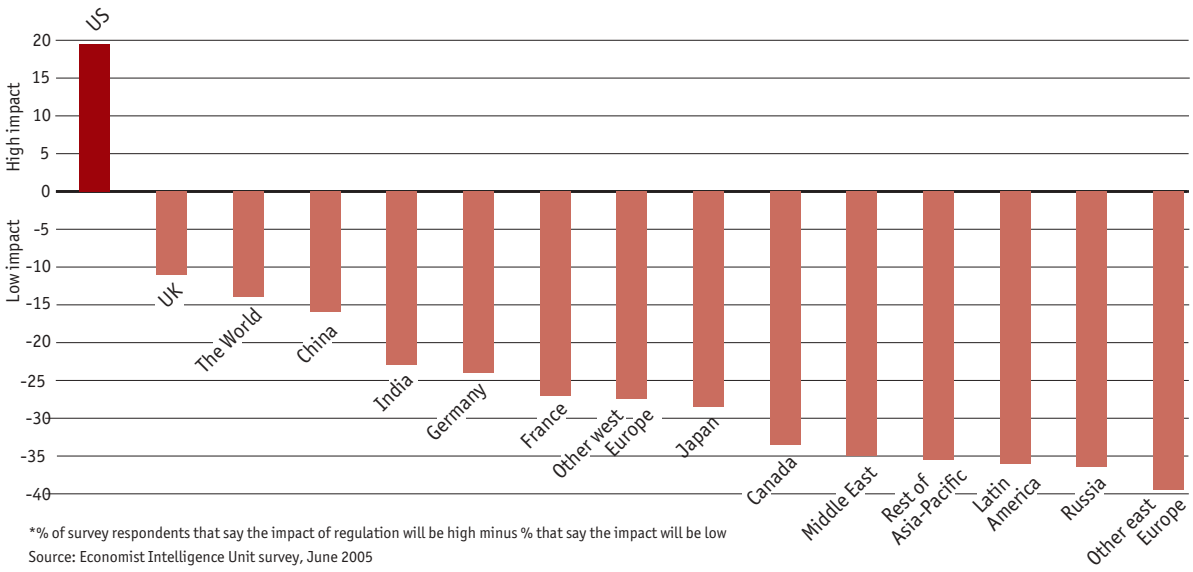
© The Economist Intelligence Unit 2005



The US is also the only market where most respondents expect a high impact from regulatory change over the next three years. In all others, a majority expects the impact to be low. However, the breakdown by region and industry shows that the strongest reaction comes from those firms based in the US and those in the financial sector. Firms outside the US and from other sectors, although generally in agreement that the changing regulatory scene in the US will have a high impact, are less pronounced in this view.

These perceptions are of more than academic interest. For companies operating in the US, it is imperative to establish good lines of communication with the country's increasingly energetic regulators. But there are wider implications too. Over half of executives in the survey say that their perception of a country's regulatory burden has an important or very important impact on decisions to invest, and over one-third say they have been deterred from investing in a new market because of regulatory issues. Policymakers in the US would do well to take these points on board when planning the next round of regulations.

How significant an impact do you think changes in regulation in these countries or regions will have on your business over the next three years?
Average % score by region*

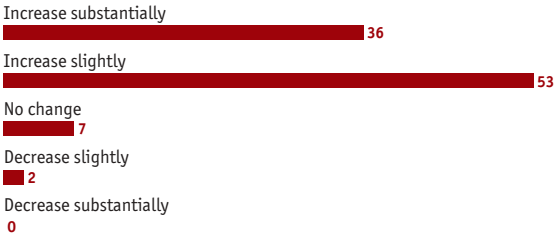


*% of survey respondents that say the impact of regulation will be high minus % that say the impact will be low
Source: Economist Intelligence Unit survey, June 2005

© The Economist Intelligence Unit 2005



Do you expect the overall cost of regulation on your business to increase or decrease over the next three years?
(% respondents)



Source: Economist Intelligence Unit survey, June 2005

regulatory requirements would have on their companies' reputation. A large proportion of companies also worry that their internal resources will be overwhelmed by the burden of compliance. Some executives go further: a significant minority believe there is a danger that regulation will make their firms too risk-averse, or make them uncompetitive when compared with rivals operating in less regulated countries.

Although executives in all industries register similar fears in the survey, the level of concern varies between industries. For companies outside the financial services industry, the risk of resources being overwhelmed by regulatory compliance was a greater concern even than the risk of reputational damage. In general, executives in non-financial firms register lower levels of concern than those in financial services companies.

Few executives in the survey expect these regulatory burdens to diminish, at least in the short term. More than two-thirds believe that the trend towards tighter regulation will increase over the next three years. This is felt even more strongly in the financial services industry, where three-quarters of companies predict further regulatory tightening. An overwhelming 90% also expect the cost of regulatory compliance to rise over the same time period.

Executives are also concerned about the indirect costs of regulation, however, as resources are diverted from value-creating activities into complying with red tape. For example, 36% of respondents are in broad agreement with the statement that regulation has stifled innovation, and a similar proportion say that regulatory issues have deterred their firms from investing in new markets. A large majority of firms (63%) expect the costs of regulation to be passed on to customers through higher prices. If these views are borne out, the recent raft of regulations will bring undesirable consequences for the very people they were meant to protect—shareholders and customers.

Staying one step ahead

If regulation has become a major source of risk in its own right, what should risk managers be doing about it? For most executives in the survey, the key challenge is to predict and prepare for changes in the regulatory environment. But given the complexity and rapid pace of change in regulation, this is easier said than done.

In response to this challenge, most companies in the survey advocate two main strategies. The first involves adopting best practice whether required to by current regulation or not. This appears designed both to pre-empt regulation before it arises, and to deny the regulator a reason to introduce it in the first place.

Even where regulators feel compelled to act, early adoption of best practice can save money. "Sarbanes-Oxley was a huge issue for us, and we embraced it very early on," says John Wengler, CRO at Entergy Corp, a Texas-based energy producer and distributor with operations in seven south eastern states (see box). In Entergy's case, the new legislation was in tune with the existing management culture, and this lessened the burden of implementation.

"We had a good control environment to start with and we were able to leverage it," says Mr Wengler. "Adopting Sarbanes-Oxley was synergistic, not diversionary."

Other firms have found it more difficult to stay one

© The Economist Intelligence Unit 2005



Case study: Entergy

Energy provision in the US is perhaps the most heavily regulated business anywhere. The country's generating and distributing companies are subject not only to the general regulatory environment that all corporations face, but also to a tight framework of sectoral rules designed to ensure a smooth supply of power in return for a predictable and attractive return on investment. Even this strictly sectoral set of rules varies from region to region, and from product to product.

Entergy Corp, is a Texas-based energy producer and distributor with operations in seven south eastern states. It also owns the country's second-largest fleet of nuclear generation facilities.

Much of Entergy's conventional electricity generation and distribution business falls under state-level regulatory regimes that set the price at which power is sold. This part of the business is subject to the regulators of seven different states, each with its own approach to the industry. "Each state is truly different," says John Wengler, the company's CRO. "For instance, Louisiana's public service commissioners are elected, but in Texas they're appointed. There's a more public atmosphere in Louisiana."

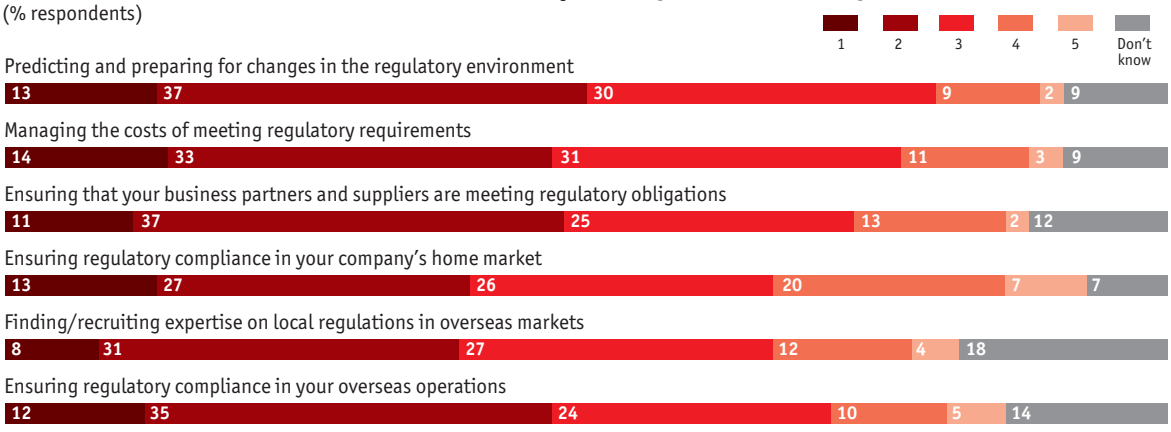
Regulatory uncertainty is substantial in all areas of the business. For instance, Entergy's conventional operations in Texas are currently subject to regulated rates, but the state is talking about whether to allow utilities to compete on the open market. In the nuclear business, the company is monitoring developments on the proposed Yucca Mountain nuclear waste disposal site in Nevada. If this is rejected, the company will need to build the costs of securing alternative storage facilities into its business plans.

Cutting across this is the overriding issue in US energy regulation, the fate of the 1935 Federal Public Utility Holding Company Act (PUHCA), which places broad restrictions on where and how energy companies can invest. Proponents of reform say it would stimulate investment and improve efficiency. But a powerful array of critics believe it would encourage speculation and market abuse, as well as place the environment at risk.

"Regulatory risk is fundamentally different," reflects Mr Wengler. "You can't calculate value at risk (VAR) from regulatory sources. But there is a common denominator: dollars. More precisely, uncertainty attached to dollars."

step ahead of the regulators, however. In the financial services sector, for example, the introduction of the Basel II framework requires a fundamental re-think of risk assessment and control. This has been costly even for some institutions that had adopted its philosophy at an early stage. "Lots of companies considered capital needs from the risk perspective, but not in the precise numerical terms of Basel II," says Mr de Hoest of Egg Bank. "Some had to re-invent systems they had already invested heavily in." Although Egg Bank has

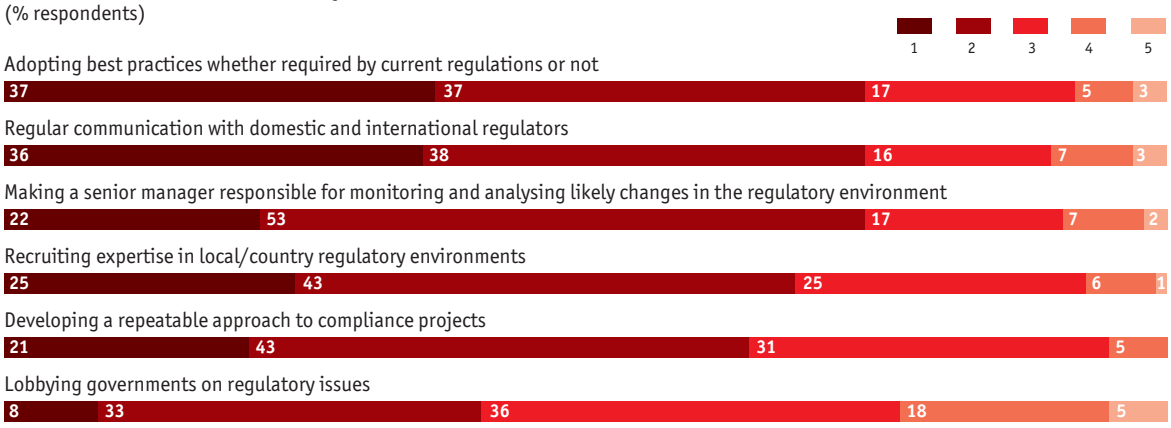
How significant for your company are the following challenges associated with managing regulatory risk on an international basis? Rate on a scale of 1 to 5, where 1=Major challenge and 5=Not a challenge.
(% respondents)



Source: Economist Intelligence Unit survey, June 2005



How effective do you believe the following strategies are for predicting and managing regulatory risk? Rate on a scale of 1 to 5, where 1=Very effective and 5=Ineffective.
(% respondents)



Source: Economist Intelligence Unit survey, June 2005

been able to integrate the Basel II requirements into its existing processes, Mr de Hoest says it won't get back its investment pound-for-pound.

This leads us to the second strategy, which entails building as close a relationship as possible with domestic and international regulators. This is a two-way process. It allows firms to keep abreast of the way regulators are thinking, and the issues that are giving them concern. It also gives firms the opportunity to ensure that regulators are familiar with the realities of the business world, and to reassure them that management is on top of issues that might otherwise require regulation.

Managing regulatory risk "is based on early warning," says Mr de Hoest. This means keeping abreast of "what's coming out of regulators' consultation papers and so on, what's coming down the track."

For firms outside the US, and for those in the financial sector, communication with the regulator is seen as more important than adopting best practice. For firms outside the financial sector, best practice comes top, but other procedural issues such as developing a repeatable approach to compliance projects and recruiting expertise in the local regulatory environment are considered more important than maintaining close relations with the regulator.

Although all respondents gave a high priority to staying in touch with the regulatory agencies, direct lobbying of governments on regulatory issues was identified as an effective strategy by relatively few firms. Indeed, in the financial sector responses were equally divided between those who saw this as an effective strategy and those who considered it ineffective.

© The Economist Intelligence Unit 2005



Conclusion

The tension between corporate self-discipline and state regulation is as active as ever. In a globalised economy awash with financial flows and dominated by multinational conglomerates, the stakes are also extremely high. For risk managers, navigating through this period of change is a formidable challenge, and one that must be achieved without detracting from the task of monitoring and mitigating a multitude of other risks that fall under the remit of the CRO.

In this regard, an integrated approach to risk management is especially crucial. "The trick is getting everybody working together, because it's the cracks in the system that catch you out," says Mr Martens of TIAA-CREF. In a world where risks are increasingly inter-dependent, it's a lesson that is as relevant to policymakers and regulators as it is to corporations.

Survey and ranking methodology

The findings in this press release are based on a survey of 230 senior executives. Forty-four percent of the participants were based in the US, 32% were based in Europe, and 15% were based in Asia Pacific. All survey findings in this report and press release are drawn from a survey conducted in June 2005.

Thirty-nine percent of the companies participating in the survey were from the financial services sector. Respondents from 17 other industries participated in the survey, including professional services, energy sector companies, IT and telecoms firms, and pharmaceutical and healthcare companies.

The survey questions asked respondents to rate options from strongly positive to strongly negative, with neutral in the middle. Results were weighted (strong positives and negatives were doubled, moderate positives and negatives were unchanged, and neutrals were discarded) and combined to produce an average % score, and ranked accordingly.

© The Economist Intelligence Unit 2005

Whilst every effort has been taken to verify the accuracy of this information, neither The Economist Intelligence Unit Ltd. nor the sponsor of this report can accept any responsibility or liability for reliance by any person on this white paper or any of the information, opinions or conclusions set out in the white paper.

LONDON
15 Regent Street
London
SW1Y 4LR
United Kingdom
Tel: (44.20) 7830 1000
Fax: (44.20) 7499 9767
E-mail: london@eiu.com

NEW YORK
111 West 57th Street
New York
NY 10019
United States
Tel: (1.212) 554 0600
Fax: (1.212) 586 1181/2
E-mail: newyork@eiu.com

HONG KONG
60/F, Central Plaza
18 Harbour Road
Wanchai
Hong Kong
Tel: (852) 2585 3888
Fax: (852) 2802 7638
E-mail: hongkong@eiu.com

Investment Opportunities as Real Options: Getting Started on the Numbers

by Timothy A. Luehrman



Harvard Business Review

Reprint 98404

Harvard Business Review

JULY–AUGUST 1998

Reprint Number

C.K. PRAHALAD AND KENNETH LIEBERTHAL	**THE END OF CORPORATE IMPERIALISM**	98408
LINDA HILL AND SUZY WETLAUFER	**LEADERSHIP WHEN THERE IS NO ONE TO ASK: AN INTERVIEW WITH ENI'S FRANCO BERNABE**	98402
B. JOSEPH PINE II AND JAMES H. GILMORE	**WELCOME TO THE EXPERIENCE ECONOMY**	98407
ROBIN COOPER AND ROBERT S. KAPLAN	**THE PROMISE–AND PERIL–OF INTEGRATED COST SYSTEMS**	98403
THOMAS H. DAVENPORT	**PUTTING THE ENTERPRISE INTO THE ENTERPRISE SYSTEM**	98401
NIGEL NICHOLSON	**HOW HARDWIRED IS HUMAN BEHAVIOR?**	98406
REGINA FAZIO MARUCA	HBR CASE STUDY **HOW DO YOU MANAGE AN OFF-SITE TEAM?**	98405
DURWARD K. SOBEK II, JEFFREY K. LIKER, AND ALLEN C. WARD	IDEAS AT WORK **ANOTHER LOOK AT HOW TOYOTA INTEGRATES PRODUCT DEVELOPMENT**	98409
TIMOTHY A. LUEHRMAN	MANAGER'S TOOL KIT **INVESTMENT OPPORTUNITIES AS REAL OPTIONS: GETTING STARTED ON THE NUMBERS**	98404
INTRODUCTION BY RICHARD L. NOLAN	PERSPECTIVES **CONNECTIVITY AND CONTROL IN THE YEAR 2000 AND BEYOND**	98411
DAVID WARSH	BOOKS IN REVIEW **WHAT DRIVES THE WEALTH OF NATIONS?**	98410

Here's a way to apply option pricing to strategic decisions without hiring an army of Ph.D.'s.

Investment Opportunities as Real Options:

Getting Started on the Numbers

by Timothy A. Luehrman

THE ANALOGY between financial options and corporate investments that create future opportunities is both intuitively appealing and increasingly well accepted. Executives readily see why investing today in R&D, or in a new marketing program, or even in certain capital expenditures (a phased plant expansion, say) can generate the possibility of new products or new markets tomorrow. But for many nonfinance managers, the journey from insight to action, from the puts and calls of financial options to actual investment decisions, is difficult and deeply frustrating.

Experts do a good job of explaining what option pricing captures that conventional discounted-cash-flow (DCF) and net-present-value (NPV) analyses do not. Moreover, simple option pricing for exchange-traded puts and calls is fairly straightforward, and many books present the basics lucidly. But at that point, most executives get stuck. Their interest piqued, they want to know How can I use option pricing on *my* project? and How can I use this with real numbers rather than with sterilized examples? Unfortunately, how-to advice is scarce on this subject and mostly aimed at specialists, preferably with Ph.D.'s. As a result, corporate analyses that generate real numbers have been rare, expensive, and hard to understand.

The framework presented here bridges the gap between the practicalities of real-world capital projects and the higher mathematics associated with formal option-pricing theory. It produces quantitative output, can be used repeatedly on many projects, and is compatible with the ubiquitous DCF spreadsheets that are at the heart of most corporate capital-budgeting systems. What this framework cannot supply is absolute precision: when a very precise number is required, managers will still have to call on technical experts with specialized financial tools. But for many projects in many companies, a "good enough" number is not only good enough but considerably better than the number a plain DCF analysis would generate. In such cases, forgoing some precision in exchange for simplicity, versatility, and explicability is a worthwhile trade.

We'll begin by examining a generic investment opportunity – a capital budgeting project – to see what makes it similar to a call option. Then we'll compare DCF with the option-pricing approach to evaluating the project. Instead of looking only at the differences between the two approaches, we will also look for points of commonality. Recognizing the differences adds extra insight to the analysis, but exploiting the commonalities is the key to making the framework understandable and compatible with familiar techniques. In fact, most of the data the framework uses come from the DCF spreadsheets that managers routinely prepare to evaluate investment proposals. And for option values, the framework uses the Black-Scholes option-pricing table instead of complex equations. Finally, once we've built the framework, we'll apply it to a typical capital-investment decision.

Mapping a Project Onto an Option

A corporate investment opportunity is like a call option because the corporation has the right, but not the obligation, to acquire something – let us say, the operating assets of a

Timothy A. Luehrman *is a professor of finance at Thunderbird, the American Graduate School of International Management, in Glendale, Arizona. He is the author of "What's It Worth: A General Manager's Guide to Valuation," HBR May–June 1997.*

 Copyright © 1998 by the President and Fellows of Harvard College. All rights reserved.

new business. If we could find a call option sufficiently similar to the investment opportunity, the value of the option would tell us something about the value of the opportunity. Unfortunately, most business opportunities are unique, so the likelihood of finding a similar option is low. The only reliable way to find a similar option is to construct one.

To do so, we need to establish a correspondence between the project's characteristics and the five variables that determine the value of a simple call option on a share of stock. By mapping the characteristics of the business opportunity onto the template of a call option, we can obtain a model of the project that combines its characteristics with the structure of a call option. The option we will use is a *European* call, which is the simplest of all options because it can be exercised on only one date, its expiration date. The option we synthesize in this way is not a perfect substitute for the real opportunity, but because we've designed it to be similar, it is indeed informative. The diagram "Mapping an Investment Opportunity onto a Call Option" shows the correspondences making up the fundamental mapping.

Many projects involve spending money to buy or build a productive asset. Spending money to exploit such a business opportunity is analogous to exercising an option on, for example, a share of stock. The amount of money expended corresponds to the option's *exercise price* (denoted for simplicity as X). The present value of the asset built or acquired corresponds to the *stock price* (S). The length of time the company can defer the investment decision without losing the opportunity corresponds to the option's *time to expiration* (t). The uncertainty about the future value of the project's cash flows (that is, the riskiness of the project) corresponds to the *standard deviation of returns* on the stock (σ). Finally, the time value of money is given in both cases by the *risk-free rate of return* (r_f). By pricing an option using values for these variables generated from our project, we learn more about the value of the project than a simple discounted-cash-flow analysis would tell us.

Mapping an Investment Opportunity onto a Call Option

Investment Opportunity	Variable	Call Option
Present value of a project's operating assets to be acquired	S	Stock price
Expenditure required to acquire the project assets	X	Exercise price
Length of time the decision may be deferred	t	Time to expiration
Time value of money	r_f	Risk-free rate of return
Riskiness of the project assets	σ^2	Variance of returns on stock

When Are Conventional NPV and Option Value Identical?

Conventional NPV and option value are identical when the investment decision can no longer be deferred.

Conventional NPV	Option Value
NPV = (value of project assets) − (expenditure required)	When $t = 0$, σ^2 and r_f do not affect call option value. Only S and X matter.
This is S. This is X.	
So: NPV = $S - X$	At expiration, call option value is $S - X$ or 0, whichever is greater.
Here, we must decide "go" or "no go."	Here, it's "exercise" or "not."

Linking NPV And Option Value

Traditional DCF methods would assess this opportunity by computing its net present value. NPV is the difference between how much the operating assets are worth (their present value) and how much they cost:

NPV = present value of assets − required capital expenditure.

When NPV is positive, the corporation will increase its own value by making the investment. When NPV is negative, the corporation is better off not making the investment.

When are the project's option value and NPV the same? When a final decision on the project can no longer be deferred; that is, when the company's "option" has reached its expiration date. At that time, either

the option value = $S - X$

or

the option value = 0

whichever is greater. But note that

NPV = $S - X$

as well, because we know from our map that S corresponds to the present value of the project assets and X to the required capital expenditure. To reconcile the two completely, we need only observe that when NPV is negative, the corporation will not invest, so the project value is effectively zero (just like the option value) rather than negative. In short, both approaches boil down to the same number and the same decision. (See the diagram "When Are Conventional NPV and Option Value Identical?")

This common ground between NPV and option value has great practical significance. It means that corporate spreadsheets set up to compute conventional NPV are highly relevant for option pricing. Any spreadsheet that computes NPV already contains the information necessary to compute S and X, which are two of the five option-pricing variables. Accordingly, executives who want to begin using option pricing need not discard their current DCF-based systems.

When do NPV and option pricing diverge? When the investment decision may be deferred. The possibility of deferral gives rise to two additional sources of value. First, we would always rather pay later than sooner, all else being equal, because we can earn the time value of money on the deferred expenditure. Second, while we're waiting, the world can change. Specifically, the value of the operating assets we intend to acquire may change. If their value goes up, we haven't missed out; we still can acquire them simply by making the investment (exercising our option). If their value goes down, we might decide not to acquire them. That also is fine (very good, in fact) because, by waiting, we avoid making what would have turned out to be a poor investment. We have preserved the ability to participate in good outcomes and insulated ourselves from some bad ones.

For both of these reasons, being able to defer the investment decision is valuable. Traditional NPV misses the extra value associated with deferral because it assumes the decision cannot be put off. In contrast, option pricing *presumes* the ability to defer and provides a way to quantify the value of deferring. So to value the investment, we need to develop two new metrics that capture these extra sources of value.

Quantifying Extra Value: NPVq. The first source of value is the interest you can earn on the required capital expenditure by investing later rather than sooner. A good way to capture that value is to suppose you put just enough money in the bank now so that when it's time to invest, that money plus the interest it has earned is sufficient to fund the required expenditure. How much money is that? It is the discounted present value of the capital expenditure. In option notation, it's the present value of the exercise price, or PV(X). To compute PV(X), we discount X for the requisite number of periods (t) at the risk-free rate of return (r_f):

$$PV(X) = X \div (1 + r_f)^t.$$

The extra value is the interest rate (r_f) times X, compounded over however many time periods (t) are involved. Alternatively, it is the difference between X and PV(X).

We know that conventional NPV is missing that extra value, so let's put it in. We have seen that NPV can be expressed in option notation as:

$$NPV = S - X.$$

Let's rewrite it using PV(X) instead of X. Thus:

$$\text{"modified" NPV} = S - PV(X).$$

Note that our modified NPV will be greater than or equal to regular NPV because it explicitly includes interest to be earned while we wait. It picks up one of the sources of value we are interested in.

Modified NPV, then, is the difference between S (value) and PV(X) (cost adjusted for the time value of money). Modified NPV can be positive, negative, or zero. However, it will make our calculations a lot easier if we express the relationship between cost and value in such a way that the number can never be negative or zero.

So instead of expressing modified NPV as the *difference* between S and PV(X), let's create a new metric: S *divided* by PV(X). By converting the difference to a ratio, all we are doing, essentially, is converting negative values to decimals between zero and one.[1] We'll call this new metric NPVq, where "q" reminds us that we are expressing the relationship between cost and value as a quotient:

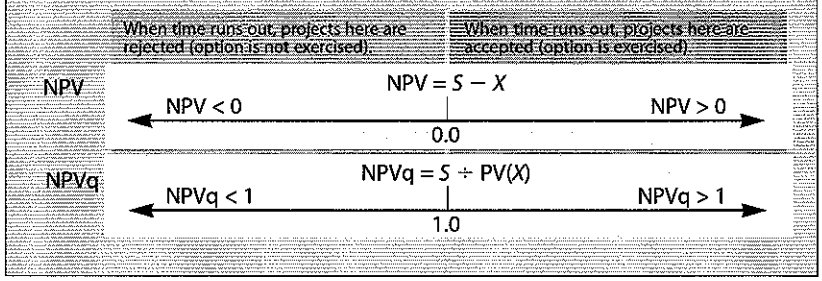
$$NPVq = S \div PV(X).$$

Modified NPV and NPVq are not equivalent; that is, they don't yield the same numeric answer. For example, if $S = 5$ and PV(X) = 7, NPV = −2 but NPVq = 0.714. But the difference in the figures is unimportant because we haven't lost any information about the project by substituting one metric for another. When modified NPV is positive, NPVq will be greater than one; when NPV is negative, NPVq will be less than one. Anytime modified NPV is zero, NPVq will be one. There is a perfect correspondence between them, as the diagram "Substituting NPVq for NPV" shows.

Quantifying Extra Value: Cumulative Volatility. Now let's move on to the second source of additional value, namely that while we're waiting, asset value may change and affect our investment decision for the better. That possibility is very important, but naturally it is more difficult to quantify because we are not actually sure that asset values will change or, if they do, what the future values will be. Fortunately, rather than measuring added value directly,

Substituting NPVq for NPV

We can rank projects on a continuum according to values for NPVq, just as we would for NPV. When a decision can no longer be deferred, NPV and NPVq give identical investment decisions, but NPVq has some mathematical advantages.



we can measure *uncertainty* instead and let an option-pricing model quantify the value associated with a given amount of uncertainty. Once again, we'll go through two steps. First, we'll identify a sensible way to measure uncertainty. Then we'll express the metric in a mathematical form that will be easier for us to use but will not cause us to lose any practical content.

The only way to measure uncertainty is by assessing probabilities. Imagine that the project's future value is to be drawn from an urn containing all possible future values, weighted according to their likelihood of occurring. That is, if a value of $100 were twice as likely as $75 or $125, there would be twice as many $100 balls in the urn as $75 balls or $125 balls.

How can we quantify this uncertainty? Perhaps the most obvious measure is simply the range of all possible values: the difference between the lowest and the highest possibilities. But we can do better than that by taking into account the relative likelihood of values between those extremes. If, for example, very high and very low values are less likely than "medium" or "average" values, our measure of uncertainty should reflect that. The most common probability-weighted measure of dispersion is *variance*, often denoted as *sigma squared* (σ^2). Variance is a summary measure of the likelihood of drawing a value far away from the average value in the urn. The higher the variance, the more likely it is that the values drawn will be either much higher or much lower than average. In other words, we might say that high-variance assets are riskier than low-variance assets.

Variance is an appealing measure of uncertainty, but it's incomplete. We have to worry about a time dimension as well: how much things can change while we wait depends on how long we can afford to wait. For business projects, things can change a lot more if we wait two years than if we wait only two months. So in option valuation, we speak in terms of variance *per period*. Then our measure of the total amount of uncertainty is variance per period times the number of periods, or $\sigma^2 t$.

This sometimes is called *cumulative variance*. An option expiring in two years has twice the cumulative variance as an otherwise identical option expiring in one year, given the same variance per period. Alternatively, it may help to think of cumulative variance as the amount of variance in the urn times the number of draws you are allowed, which again is $\sigma^2 t$.

Cumulative variance is a good way to measure the uncertainty associated with business investments. Now we'll make two modifications, again for mathematical convenience, that won't affect the ability of the variable to tell us what we want to know about uncertainty. First, instead of using the variance of project *values*, we'll use the variance of project *returns*. In other words, rather than working with the actual dollar value of the project, we'll work with the percentage gained (or lost) per year. There is no loss of content because a project's return is completely determined by the project's value:

$$\text{return} = \frac{(\text{future value} - \text{present value})}{\text{present value}}.$$

The probability distribution of possible values is usually quite asymmetric; value can increase greatly but cannot drop below zero. Returns, in contrast, can be positive or negative, sometimes symmetrically positive or negative, which makes their probability distribution easier to work with.

Second, it helps to express uncertainty in terms of *standard deviation* rather than variance. Standard deviation is simply the square root of variance and is denoted by σ. It tells us just as much about uncertainty as variance does, but it has the advantage of being denominated in the same units as the thing being measured. In our business example, future asset values are denominated in units of currency – say, dollars – and returns are denominated in percentage points. Standard deviation, then, is likewise denominated in dollars or percentage points, whereas variance is denominated in *squared* dollars or *squared* percentage points, which are not intuitive. Since we are going to work with returns instead of values, our units will be percentage points instead of dollars.

To make these refinements to our measure of total uncertainty, we do the following:

First, stipulate that σ^2 denotes the variance of returns per unit of time on our project.

Second, multiply variance per period by the number of periods (t) to get cumulative variance ($\sigma^2 t$).

Third, take the square root of cumulative variance to change units, expressing the metric as standard deviation rather than variance. Let's call this last quantity *cumulative volatility* ($\sigma\sqrt{t}$) to distinguish it from cumulative variance.

Valuing the Option. Together, our two new call-option metrics, NPVq and $\sigma\sqrt{t}$, contain all the information needed to value our project as a European call option using the Black-Scholes model. They capture the extra sources of value associated with opportunities. And they are composed of the five fundamental option-pricing variables onto which we mapped our business opportunity. NPVq is actually a combination of four of the five variables: S, X, r_f, and t. Cumulative volatility combines the fifth, σ, with t. (See the diagram "Linking Our Metrics to the Black-Scholes Model.") By combining variables in this way, we get to work with two metrics instead of five. Not only is that easier for most of us to grasp, it also allows us to plot two-dimensional pictures, which can be helpful substitutes for equations in managers' discussions and presentations. Finally, each of the metrics has a natural business interpretation, which makes option-based analysis less opaque to nonfinance executives.

The graph "Locating the Option Value in Two-Dimensional Space" shows how to use NPVq and $\sigma\sqrt{t}$ to obtain a value for the option. NPVq is on the horizontal axis, increasing from left to right. As NPVq rises, so does the value of the call option.

What causes higher values of NPVq? Higher project values (S) or lower capital expenditures (X). Note further that NPVq also is higher whenever the present value of X is lower. Higher interest rates (r_f) or longer time to expiration (t) both lead to lower present values of X. Any of these changes (lower X or higher S, r_f, or t) increases the value of a European call.

Cumulative volatility is on the vertical axis of the graph, increasing from top to bottom. As $\sigma\sqrt{t}$ increases, so does call value. What causes higher values of $\sigma\sqrt{t}$? Greater uncertainty about a project's future value and the ability to defer a decision longer. Either of these changes (higher σ or t) likewise increases the value of a European call.

Plotting projects in this two-dimensional space creates a visual representation of their relative option values. No matter where you start in the graph, call value increases

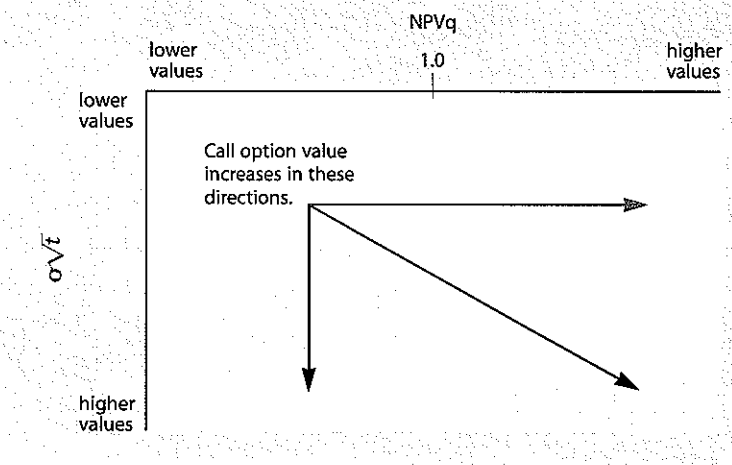
Linking Our Metrics to the Black-Scholes Model

Our two new metrics together contain all five variables in the Black-Scholes model. Combining five variables into two lets us locate opportunities in two-dimensional space.

Investment Opportunity	Call Option	Variable	Option Value Metrics
Present value of a project's operating assets to be acquired	Stock price	S	
Expenditure required to acquire the project assets	Exercise price	X	NPVq
Length of time the decision may be deferred	Time to expiration	t	
Time value of money	Risk-free rate of return	r_f	
Riskiness of the project assets	Variance of returns on stock	σ^2	$\sigma\sqrt{t}$

Locating the Option Value in Two-Dimensional Space

We can locate investment opportunities in this two-dimensional space.



when you move down, to the right, or in both directions at once. Projects in the lower-right corner of the graph are high on both NPVq and $\sigma\sqrt{t}$ metrics and their option value is high compared with projects in the upper-left corner.

Locating various projects in the space reveals their value relative to one another. How do we get absolute values? That is, how can we get a *number*? Having gotten this far, we find that getting a number is easy. Because NPVq and $\sigma\sqrt{t}$ contain all five Black-Scholes variables, we can fill in a table with Black-Scholes call values that correspond to *every pair* of NPVq and $\sigma\sqrt{t}$ coordinates. I call this "pricing the space," and the table does it for us.

The exhibit "Using the Black-Scholes Option-Pricing Model to 'Price the Space'" shows part of the filled-in Black-Scholes table. Each number expresses the value of a specific call option as a percentage of the underlying project's (or asset's) value. For example, for a project whose NPVq equals 1.0 and $\sigma\sqrt{t}$ equals 0.5, the value given in the table is 19.7%. Any European call option for which NPVq is 1.0 and $\sigma\sqrt{t}$ is 0.5 will have a value equal to 0.197 times S. If the assets associated with a particular project have a value (S) of $100, then the project viewed as a call option has a value of $19.70. If S were $10, the call option value would be $1.97, and so forth. Option values in the table are expressed in relative terms, as percentages of S, rather than in absolute dollars, to enable us to use the same table for both big and small projects. It's also convenient not to have to manipulate the Black-Scholes equation every time we want to value a project. The Black-Scholes model is used once, to generate the table itself.[2] After that, we need only locate our project in the table and multiply by a factor of S.

Why is the option value $19.70 less than the asset value of $100? We've been analyzing sources of extra value associated with being able to defer an investment. The key is to remember that *extra* refers to a comparison between option value and *net* present value (NPV), not to a comparison between option value and present value (S). In this example, we are not expecting the option value to be greater than S; we are expecting it to be greater than NPV, which is S minus capital expenditures (X). S equals $100 here, but we didn't say what X was. Since NPVq equals 1.0 in this example, X must in fact be greater than $100; otherwise NPVq would be greater than 1.0. For concreteness, suppose that this is a

Using the Black-Scholes Option-Pricing Model to "Price the Space"

Each number in the table gives the value of a European call for specified values of NPVq and $\sigma\sqrt{t}$, as a percentage of S, the value of project assets.

Example:

Suppose S = $100
X = $105
t = 1 year
r_f = 5%
σ = 50% per year
then NPVq = 1.0
and $\sigma\sqrt{t}$ = 0.50.

The table gives a value of 19.7%.

Interpretation:

Viewed as a call option, the project has a value of:

Call value = 0.197 × $100 = $19.70.

Compare this to its conventional NPV:

NPV = $S - X$
= $100 − $105
= $−5.

Black-Scholes value of a European call option, expressed as a percentage of underlying asset value.

$\sigma\sqrt{t}$ \ NPVq	0.80	0.82	0.84	0.86	0.88	0.90	0.92	0.94	0.96	0.98	**1.00**	1.02	1.04	1.06	1.08
0.05	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.3	0.6	1.2	**2.0**	3.1	4.5	6.0	7.5
0.10	0.0	0.1	0.2	0.3	0.5	0.8	1.2	1.7	2.3	3.1	**4.0**	5.0	6.1	7.3	8.6
0.15	0.5	0.7	1.0	1.3	1.7	2.2	2.8	3.5	4.2	5.1	**6.0**	7.0	8.0	9.1	10.2
0.20	1.5	1.9	2.3	2.8	3.4	4.0	4.7	5.4	6.2	7.1	**8.0**	8.9	9.9	10.9	11.9
0.25	2.8	3.3	3.9	4.5	5.2	5.9	6.6	7.4	8.2	9.1	**9.9**	10.9	11.8	12.8	13.7
0.30	4.4	5.0	5.7	6.3	7.0	7.8	8.6	9.4	10.2	11.1	**11.9**	12.8	13.7	14.6	15.6
0.35	6.2	6.8	7.5	8.2	9.0	9.8	10.6	11.4	12.2	13.0	**13.9**	14.8	15.6	16.5	17.4
0.40	8.0	8.7	9.4	10.2	11.0	11.7	12.5	13.4	14.2	15.0	**15.9**	16.7	17.5	18.4	19.2
0.45	9.9	10.6	11.4	12.2	12.9	13.7	14.5	15.3	16.2	17.0	**17.8**	18.6	19.4	20.3	21.1
0.50	**11.8**	**12.6**	**13.4**	**14.2**	**14.9**	**15.7**	**16.5**	**17.3**	**18.1**	**18.9**	**19.7**	20.5	21.3	22.1	22.9
0.55	13.8	14.6	15.4	16.1	16.9	17.7	18.5	19.3	20.1	20.9	21.7	22.4	23.2	24.0	24.8
0.60	15.8	16.6	17.4	18.1	18.9	19.7	20.5	21.3	22.0	22.8	23.6	24.3	25.1	25.8	26.6
0.65	17.8	18.6	19.3	20.1	20.9	21.7	22.5	23.2	24.0	24.7	25.3	26.2	27.0	27.7	28.4
0.70	19.8	20.6	21.3	22.1	22.9	23.6	24.4	25.2	25.9	26.6	27.4	28.1	28.8	29.5	30.2
0.75	21.8	22.5	23.3	24.1	24.8	25.6	26.3	27.1	27.8	28.5	29.2	29.9	30.6	31.3	32.0
0.80	23.7	24.5	25.3	26.0	26.8	27.5	28.3	29.0	29.7	30.4	31.1	31.8	32.4	33.1	33.8
0.85	25.7	26.5	27.2	28.0	28.7	29.4	30.2	30.9	31.6	32.2	32.9	33.6	34.2	34.9	35.5
0.90	27.7	28.4	29.2	29.9	30.6	31.3	32.0	32.7	33.4	34.1	34.7	35.4	36.0	36.6	37.3

one-year option – that is, suppose we can defer the decision for one year – and that the risk-free rate of return (r_f) is 5%. Then for NPVq to equal 1.0, X must be $105 Recall that:

$$\begin{aligned} NPVq &= S \div PV(X) \\ &= \$100 \div (\$105 \div 1.05). \end{aligned}$$

Thus the conventional NPV is actually negative:

$$\begin{aligned} NPV &= S - X \\ &= \$100 - \$105 \\ &= -\$5. \end{aligned}$$

And an option value of $19.70 really is substantially greater than conventional NPV.

Using the Framework: An Example in Seven Steps

To illustrate how to apply the framework, consider this example of a hypothetical, but representative, capital investment. Division managers at a company we'll call Franklin Chemical are proposing a phased expansion of their manufacturing facilities. They plan to build a new, commercial-scale plant immediately to exploit innovations in process technology. And then they anticipate further investments, three years out, to expand the plant's capacity and to enter two new markets. The initial investment is obviously strategic because it creates the opportunity for subsequent growth. Yet executives responsible for the company's capital budget are unimpressed by the project because its NPV is essentially zero. (Cash flow projections and NPV calculations for the projected investment are shown in the table "Franklin Chemical's Initial Calculations for a Proposed Expansion.") In fact, in the annual jockeying for funds, this program may not beat out competing alternatives. Its champions are frustrated and feel sure that the company's conventional NPV approach is missing something.

They are right.

This project has considerable option value because the initial expenditure of $125 million buys the right to expand (or not) three years later. This is important because the expenditures in the third year are large – three times the initial investment.

Step 1 is to recognize the option and describe it. It takes practice to recognize the options that may be buried in conventional projects. However, there are at least two easy ways to see the option in our example. The first is simply to look beyond the numbers and examine the project's description. It surely says something about the two-phased

Franklin Chemical's Initial Calculations for a Proposed Expansion

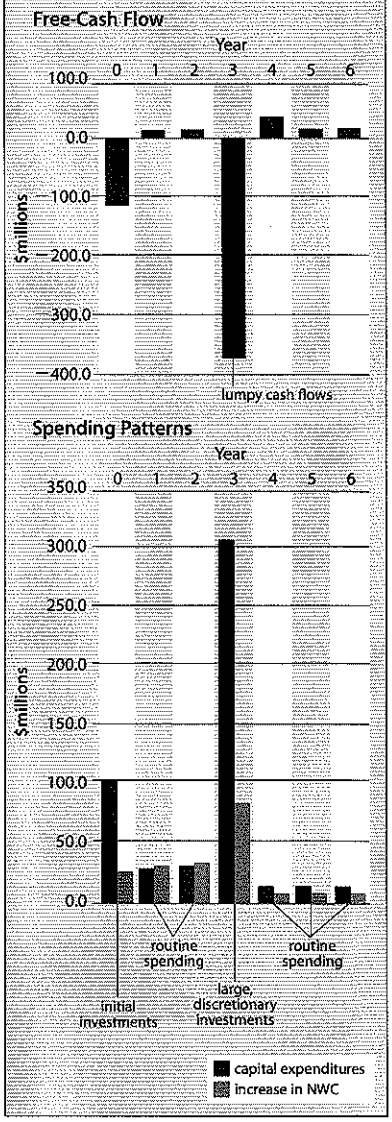
Discounted-cash-flow valuation: conventional NPV is our starting point.

Year	0	1	2	3	4	5	6
Operating projections							
revenues		455.0	551.0	800.0	1080.0	1195.0	1255.0
– cost of good sold		341.3	414.9	596.0	811.1	893.9	941.3
= gross profit		113.8	136.1	204.0	268.9	301.1	313.8
– SG&A expense		110.4	130.0	219.2	251.6	280.3	287.4
= operating profit		3.3	6.1	-15.2	17.3	20.8	26.3
Cash flow calculation							
EBIT (1-tax rate)		2.2	4.0	-10.0	11.5	13.7	17.4
+ depreciation		19.0	21.0	21.0	46.3	48.1	50.0
– capital expenditures	100.0	8.1	9.5	307.0	16.0	16.3	17.0
– increase NWC	25.0	4.1	5.5	75.0	7.1	8.0	9.7
= free cash flow, assets	-125.0	9.0	10.0	-371.0	34.7	37.5	40.7
+ terminal value, assets (perpetuity value with 5% per year growth)							610.3
Discount to present value							
× discount factor (12%)	1.000	0.893	0.797	0.712	0.636	0.567	0.507
= PV (by year)	-125.0	8.0	8.0	-264.1	22.0	21.3	329.8
NPV (sum of all years)	0.1						

Figures are in $millions and have been rounded.

Recognizing the Option

Patterns in the total cash flows and/or in expenditure items such as capital expenditures and investments in net working capital provide clues about the structure of the option.



nature of the program by way of justifying the large outlays in year 3. The other is to examine the pattern of the project cash flows over time. The cash flows in the chart are very uneven: two figures are an order of magnitude larger than the other five, and both of these are negative. A graph of the capital-expenditures line would clearly show the spike in spending in year 3. Such a large sum is almost surely *discretionary*. That is, the company can choose not to make the investment, based on how things look when the time comes. This is a classic expansion option, sometimes called a *growth option*. (See the graphs "Recognizing the Option.")

Franklin's project has two major parts. The first part is to spend $125 million now to acquire some operating assets. The second part is an option to spend an additional sum, more than $300 million, three years from now to acquire the additional capacity and enter the new markets. The option here is a call option, owned by the company, with three years to expiration, that can be exercised by investing certain amounts in net working capital (NWC) and fixed assets. Viewing the project in this way, we want to evaluate the following: NPV (entire proposal) = NPV (phase 1 assets) + call value (phase 2 assets).

Phase 1 refers to the initial investment and the associated cash flows. It can be valued using NPV as usual. *Phase 2* refers to the opportunity to expand, which may or may not be exploited in year 3. To value phase 2, we will use the framework outlined above to synthesize a comparable call option and then value it.

Step 2 is to map the project's characteristics onto call option variables. This mapping will create the synthetic option we need and indicate where in the DCF spreadsheet we need to go to obtain values for the variables. The value of the underlying assets (S) will be the present value of the assets acquired when and if the company exercises the option. The exercise price (X) will be the expenditures required to acquire the phase 2 assets. The time to expiration (t) is three years, according to

How to Estimate Cumulative Volatility

The variable in our option-pricing model that managers are least accustomed to estimating is variance (σ^2), or standard deviation (σ), which we used to get our metric for cumulative volatility ($\sigma\sqrt{t}$). For a real option, σ cannot be found in a newspaper or in a financial statement, and most people do not have highly developed intuition about, say, the annualized standard deviation of returns on assets associated with entering a new market. In the example given in the text, we assume σ is 40% per year. Is that reasonable? Here are several sound approaches to creating or judging estimates of σ:

Take a(n educated) guess.
Assets to which you would assign higher hurdle rates because of their higher-than-average systematic risk are also likely to have higher values of σ. How high is "high" for standard deviation? Returns on broad-based U.S. stock indexes had a standard deviation of approximately 20% per year for most of the past 15 years; exceptions (upward spikes) were associated with events like the 1987 stock-market crash and the 1990–1991 Persian Gulf crisis. Individual stocks generally have a higher standard deviation than the market as a whole; returns on General Motors' stock, for example, have a σ of about 25% per year. Individual projects within companies can be expected to have a still higher σ. When I work with manufacturing assets and have no specific information at all about σ, I begin by examining a range: from 30% to 60% per year.

Gather some data.
For some businesses, we can estimate volatility using historical data on investment returns in the same or related industries. Alternatively, we might compute what is called *implied volatility* using current prices of options traded on organized exchanges. The idea is to observe a market price for an option whose parameters are all known except for σ. We then use a model like Black-Scholes to figure out what σ must be given all the other variables. Today we can get implied volatility for shares of a very large number of companies in many industries. Often it is possible to use implied volatilities from options on stocks to infer σ for assets in corresponding industries. The quality and availability of such data have improved enormously in the past ten years.

Simulate σ.
Spreadsheet-based projections of a project's future cash flows, together with Monte Carlo simulation techniques, can be used to synthesize a probability distribution for project returns. Once you have the synthesized distribution, the computer can quickly calculate the corresponding standard deviation. Simulation software packages for desktop computers are commercially available and work with the same popular spreadsheet applications that generate your company's DCF models. These tools also have become far more widely available and much easier to use in recent years.

the projections given in the DCF analysis, although we might want to quiz the managers involved to determine whether the decision actually could be made sooner or later. The three-year risk-free rate of interest (r_f) is 5.5% (which is the market rate of interest on a three-year U.S. government bond). Note for comparison that the risk-adjusted discount rate being applied in the spreadsheet is 12%. Finally, the standard deviation of returns on these operating assets (σ) is not given anywhere in the spreadsheet. For now, we'll assume that figure is 40% per year, a value that is neither particularly high nor low. The insert "How to Estimate Cumulative Volatility" explains ways to obtain values for σ in cases like these.

Step 3 is to rearrange the DCF projections for two purposes: to separate phase 1 from phase 2 and to isolate values for *S* and *X*. I generally find it easier to work on *S* and *X* first. That requires making a judgment about what spending is discretionary versus nondiscretionary or what spending is routine versus extraordinary. It also requires making a similar judgment about which cash inflows are associated with phase 1 as opposed to phase 2.

In this project, expenditures on net working capital and fixed assets obviously are lumpy. The very large sums in year 3 clearly are discretionary and form part of the exercise price (X). The smaller sums in other years are plausibly routine and may be netted against phase 2 cash inflows, ultimately to be discounted and form part of S, the value of the phase 2 assets.

Sometimes it is easy to separate phase 1 cash flows from phase 2 cash flows because whoever prepared the DCF analysis built it up from detailed, phase-specific operating projections. When that's the case, as it is in our example, all we have to do is use the disaggregated detail underlying the summary DCF analysis, and the table "Franklin's Projections Rearranged" presents this information for our example. At other times, we have to allocate cash flows to each phase. A common expedient is simply to break out the phase 1 cash inflows and terminal value. Then, phase 2 cash inflows and terminal value are whatever is left over. Note that when we discount cash flows for the two phases separately, we obtain the same NPV as before.

Step 4 is to establish a benchmark for phase 2's option value based on the rearranged DCF analysis. Having separated phases 1 and 2, we can get a conventional discounted-cash-flow NPV for each, which can be seen in the table showing Franklin Chemical's rearranged calculations. This table shows that phase 1 alone has a positive NPV of $16.3 million while phase 2's NPV is −$16.2 million. The sum of the two is the same NPV, $0.1 million, that we obtained originally. Already, we have a quantitative option-related insight. The value of the whole proposal must be *at least* $16.3 million because the option value of phase 2, whatever it turns out to be, cannot be less than zero. In fact, if the option value of phase 2 is significant, the project as a whole will have a much higher value than $16.3 million, to say nothing of the $0.1 million we started with. This insight is available only when we separate the project's two phases and realize that we will have a choice about whether or not to undertake phase 2.

Our first DCF benchmark for phase 2 is −$16.2 million. Actually, though, phase 2's conventional NPV is worse than that, and it's worth a digression to see why. The DCF valuation contains a common mistake. It discounted the discretionary spending in year 3 at the same 12% risk-adjusted rate that had already

Franklin's Projections Rearranged

The DCF projections have been rearranged to separate phases 1 and 2 and isolate *S* and *X*.

Year	0	1	2	3	4	5	6
Phase 1							
cash flow	0.0	9.0	10.0	11.0	11.6	12.1	12.7
+ terminal value							191.0
− investment	-125						
× discount factor (12%)	1.000	0.893	0.797	0.712	0.636	0.567	0.507
= PV (each year)	-125.0	8.0	8.0	7.8	7.3	6.9	103.2

NPV (sum of years) **16.3** This is the conventional NPV of phase 1.

Phase 2							
cash flow				0.0	23.1	25.4	28.0
+ terminal value							419.3
− investment				-382			
× discount factor (12%)				0.712	0.636	0.567	0.507
= PV (each year)				-271.9	14.7	14.4	226.6

NPV (sum of years) **-16.2** This is the conventional NPV of phase 2.

Phases 1 and 2							
cash flow	0.0	9.0	10.0	11.0	34.7	37.5	40.7
+ terminal value							610.3
− investment	-125			-382			
× discount factor (12%)	1.000	0.893	0.797	0.712	0.636	0.567	0.507
= PV (each year)	-125.0	8.0	8.0	-264.1	22.0	21.3	329.8

NPV (sum of years) **0.1** This is the same conventional NPV we got previously.

Figures are in $millions and have been rounded.

been applied to the project's cash flows. That rate is almost certainly too high because such expenditures are rarely subject to the same operating and product-market forces that make the project's cash flows risky. Construction costs, for example, may be uncertain, but they are usually much more dependent on engineering factors, weather conditions, and contractors' performance than on customers' tastes, competitive conditions, industry capacity utilization, and such. Overdiscounting future discretionary spending leads to an optimistically biased estimate of NPV. To see the magnitude of this effect, discount the year 3 expenditures of $382 million at 5.5% instead of 12% (again, it's as if we were putting investment funds into treasury bonds between now and year 3). Then, phase 2 has a conventional DCF value of −$69.6 million, not −$16.2 million, and the NPV for the whole project goes from $0.1 million to −$53.4 million, a very substantial difference (See the table "Getting the Right Benchmark.")

Step 5 is to attach values to the option-pricing variables. Having reformulated the DCF spreadsheet, we can now pull values for S and X from it. X is the amount the company will have to invest in net working capital and fixed assets (capital expenditures) in year 3 if it wants to proceed with the expansion: $382 million. S is the present value of the new phase 2 operating assets. In other words, it's the DCF value now (at time zero) of the cash flows those assets are expected to produce in the fourth year and beyond. The same table shows that to be $255.7 million. The other option-pricing variables have already been mentioned: t is 3 years, r_f is 5.5%, and σ is 40% per year.

Step 6 is to combine the five option-pricing variables into our two option-value metrics: NPVq and $\sigma\sqrt{t}$. In this case:

$$\text{NPVq} = S \div \text{PV}(X) = \frac{\$255.7}{382 \div (1.055)^3} = 0.786$$

And

$$\sigma\sqrt{t} = 0.4 \times \sqrt{3} = 0.693.$$

The exhibit "Deriving the Option-Value Metrics for Franklin's Project" shows how all five variables were derived and how they combine to form our two metrics.

Getting the Right Benchmark

Most companies overdiscount future discretionary spending (in steps 2 and 3, we used 12%). If we discount phase 2 spending at 5.5% instead, we get a new (lower) benchmark.

Year	0	1	2	3	4	5	6
Phase 1							
cash flow	0.0	9.0	10.0	11.0	11.6	12.1	12.7
+ terminal value							191.0
− investment	-125						
× discount factor (12%)	1.000	0.893	0.797	0.712	0.636	0.567	0.507
= PV (each year)	-125.0	8.0	8.0	7.8	7.3	6.9	103.2
NPV (sum of years)	**16.3**	This is the conventional NPV of phase 1.					
Phase 2							
cash flow				0.0	23.1	25.4	28.0
+ terminal value							419.3
− investment				-382			
× discount factor (12%)				0.712	0.636	0.567	0.507
× discount factor (5.5%)				0.852			
= PV (cash flows 12%)				0.0	14.7	14.4	226.6
= PV (investment 5.5%)				-325.3			
NPV (sum of years)	**-69.6**	This is a revised DCF benchmark for phase 2.					
Phases 1 and 2							
cash flow	0.0	9.0	10.0	11.0	34.7	37.5	40.7
+ terminal value							610.3
− investment	-125			-382			
× discount factor (12%)	1.000	0.893	0.797	0.712	0.636	0.567	0.507
= PV (each year)	-125.0	8.0	8.0	7.8	22.0	21.3	329.8
= PV (investment 5.5%)				-325.3			
NPV (sum of years)	**-53.4**	This is a revised DCF benchmark for the project as a whole.					

Figures are in $millions and have been rounded.

This is r_f. This is t.

Year	0	1	2	3	4	5	6
Phase 2							
cash flow				0.0	23.1	25.4	28.0
terminal value							419.3
investment				-382			
discount factor (12%)				0.712	0.636	0.567	0.507
discount factor (5.5%)				0.852			
PV (cash flows 12%)				0.0	14.7	14.4	226.6
PV (investment 5.5%)				-325.3			
NPV	**69.6**	This is a revised DCF benchmark for phase 2.					

This is X.

The sum of these present values is S.

Step 7 is to look up call value as a percentage of asset value in our Black-Scholes option-pricing table. The table does not show values that correspond exactly to our computed value for NPVq and $\sigma\sqrt{t}$, but by interpolating we can see that the value of our synthesized call option is about 19% of the value of the underlying assets (S). Accordingly, the dollar value of the option is 0.19 times $255.7 million, which equals $48.6 million. Recall the value of the entire proposal is given by: NPV (entire proposal) = NPV (phase 1 assets) + call value (phase 2 assets). Filling in the figures gives: NPV (entire proposal) = $16.3 million + $48.6 million = $64.9 million.

Our final estimate of $64.9 million is a long way from the original figure for NPV of $0.1 million and even further from −$53.4 million. Yet the option-pricing analysis uses the same inputs from the same spreadsheet as the conventional NPV. What looks like a marginal-to-terrible project through a DCF lens is in fact a very attractive one. This seems especially so when we compare $64.9 million with the required initial investment of $125 million; the value associated with the option to expand the plant in year 3 is fully half again as much as the initial investment. Few projects look so good.

What should you do next? All the things you would usually do when evaluating a capital project. Perform sensitivity analyses. Check and update assumptions. Examine particularly interesting or threatening scenarios. Compare and interpret the analysis in light of other historical or contemporary investments and transactions. In addition, now that you've begun pricing synthesized real options, you may want to consider adding a few other items to your list. Some are discussed further in the insert "How Far Can You Extend the Framework?" They include checking for project features that would make it more suitable to use an American rather than a European option. They also include checking for clear disadvantages associated with deferring investment, such as competitive preemption, which would offset some or all of the sources of value associated with waiting. Some of these concerns can be handled by straightforward modifications to the framework. Others require more sophisticated modeling than either this framework alone or conventional NPV generally can provide. Even in those cases, though, a naïvely formulated option value will augment whatever insight may be drawn from a DCF treatment alone. Remember that simply by recognizing the structure of the problem we gleaned an important insight about the value of the project in our example (that it had to be at least $16.3 million) before we actually priced the option.

Does the framework really work? Yes. Even though we have taken some liberties, we know more about our project after using it than we did before. And if it seemed worthwhile, we could further refine our initial estimate of option value. But the key to getting useful insight from option

Deriving the Option-Value Metrics for Franklin's Project

Variable	Project Characteristic	Source in DCF Spreadsheet	Value
S	The present value of the project assets associated with phase 2.	Separate phase 1 from phase 2 cash flows; compute the DCF value of phase 2 alone.	$255.7
X	Spending required in year 3 to obtain the phase 2 assets.	Identify large discretionary portions of capital expenditures and net-working-capital spending associated with phase 2 only, in year 3.	$382
t	Length of time phase 2 spending may be deferred.	Appears in spreadsheet to be three years; check with managers.	3 years
r_f	Time value of money at same horizon (t) as option.	Obtain market rate of interest on three-year U.S. government bond, assumed here to be 5.5%.	5.5%
σ	Standard deviation per year on phase 2 assets.	Not in DCF spreadsheet; obtain from similar traded assets, from implied volatilities on traded options, or try a range. Assume 40% per year to start.	40% per year

Metric

$$NPVq = \frac{\$255.7}{\$382 \div (1.055)^3} = 0.786$$

$$\sigma\sqrt{t} = 0.4 \times \sqrt{3} = 0.693$$

How Far Can You Extend the Framework?

Real corporate projects will present immediate challenges to some of the simplifications underlying this framework. Can the framework be souped up to handle more complex problems? Or does its very simplicity present insurmountable limitations? When might it generate seriously misleading information?

Bells and whistles can be added to the framework fairly easily, although they require extra data, and the details are beyond the scope of this article. For example, the framework assumes that the amount and timing of a project's capital expenditures (*X*) are certain. But what if they are not? That's usually the case with business opportunities. The framework can be adapted to handle those circumstances, but the adaptation helps only if we can describe the uncertainty. Specifically, we need to know the probability distribution of *X* and the joint probability distribution of *S* and *X*. That is, it matters whether *X* tends to be high when *S* tends to be high, whether the opposite is true (that is, *X* tends to be high when *S* is low and vice versa), or whether they are both not only uncertain but also unrelated.

As another example, suppose the uncertainty (or variance) associated with a project changes over time. That's fairly common, and it makes sense that it should affect our estimate of cumulative volatility. Once again, if we know *how* the variance changes over time, or if we can make plausible guesses, the framework can be adapted without much trouble. The insert "Where to Find Additional Help" cites some readings that address how such real-world problems affect option values.

Both of these examples describe situations in which the real limitation is not the framework but rather the data or our knowledge of the project's parameters. Even when we know that we lack necessary data, the framework can help by showing us what the effect on value would be if the data were one thing or another. We might conclude that it's worthwhile to gather or create better data.

Some other real-world complications are thornier. The framework does a good job of capturing the extra value associated with deferring an investment decision. But what if there are particular costs associated with deferral? For example, companies trying to be first to market with the next generation of a hot product will incur large costs if deferral allows a competitor to preempt them. Anytime there are predictable costs to deferring, the option to defer an investment is less valuable, and we would be foolish to ignore those costs. If such additional costs were the only issue, we could easily handle them in our framework as long as we knew when the decision to invest or not to invest finally would be made. But in the real world, we often don't know. Companies may not be compelled to invest at a certain moment, but rather may have discretion to time their investments. So the problem is to decide not only whether to invest but also when.

In effect, many real options are *American* rather than European. American options can be exercised at any time prior to expiration; European options may be exercised only at expiration. The option-pricing table embedded in this framework prices European options. American options are more valuable than European options whenever the costs associated with deferral are predictable. When that's the case, I recommend using the framework to set the lower boundary of the option value and then supplementing that figure with another table or a spreadsheet-based algorithm to convert European option values to the corresponding American option values. That is not quite as easy as merely rewriting the pricing table because it would take a three-dimensional or possibly four-dimensional table to accommodate the extra variables needed. But it can be set up in a spreadsheet. In short, the combination of costs associated with deferral and a company's ability to time investments calls for some added analysis. But the framework needs to be augmented, not scrapped.

Finally, the Black-Scholes option-pricing model that generated the numbers in the table makes some simplifying assumptions of its own. They include assumptions about the form of the probability distribution that characterizes project returns. They also include assumptions about the *tradability* of the underlying project assets, that is, about whether those assets are regularly bought and sold. And they include assumptions about the ability of investors to continually adjust their investment portfolios. When the Black-Scholes assumptions fail to hold, this framework still yields qualitative insights but the numbers become less reliable. Consequently, it may be worthwhile to consult an expert about alternative models to improve the quantitative estimates of option value.

Where to Find Additional Help

Graduate-level corporate-finance textbooks cover the basics of option pricing, beginning from first principles:

Zvi Bodie and Robert C. Merton, *Finance* (Upper Saddle River, N.J.: Prentice-Hall, 1998).

Richard A. Brealey and Stewart Myers, *Principles of Corporate Finance*, fifth edition (New York: McGraw-Hill, 1996).

Stephen A. Ross, Randolph W. Westerfield, and Jeffrey Jaffe, *Corporate Finance*, fourth edition (Chicago: Richard D. Irwin, 1996).

Other books go beyond the basics to treat specialized problems and present more advanced models for option pricing. Well-known titles include:

John C. Hull, *Options, Futures, and Other Derivatives*, third edition (Upper Saddle River, N.J.: Prentice-Hall, 1997).

David G. Luenberger, *Investment Science* (New York: Oxford University Press, 1998).

Paul Wilmott, Jeff Dewynne, and Sam Howison, *Option Pricing* (Oxford, England: Oxford Financial Press, 1993).

A few books focus on real options in particular. Each takes a somewhat different approach to modeling corporate opportunities:

Martha Amram and Nalin Kulatilaka, *Real Options: Managing Strategic Investment in an Uncertain World* (Boston, Mass.: Harvard Business School Press, forthcoming 1998).

Avinash K. Dixit and Robert S. Pindyck, *Investment Under Uncertainty* (Princeton, N.J.: Princeton University Press, 1994).

Lenos Trigeorgis, *Real Options: Managerial Flexibility and Strategy in Resource Allocation* (Cambridge, Mass.: MIT Press, 1996).

Shorter readings on selected related topics include:

Avinash K. Dixit and Robert S. Pindyck, "The Options Approach to Capital Investment," HBR May–June 1995.

W. Carl Kester, "Today's Options for Tomorrow's Growth," HBR March–April 1984.

Timothy A. Luehrman, "What's It Worth? A General Manager's Guide to Valuation," HBR May–June 1997.

Lenos Trigeorgis, ed., *Real Options in Capital Investment: Models, Strategies, and Applications* (Westport, Conn.: Praeger, 1995).

Finally, expanded versions of the option-pricing table published in this article are available in the Brealey and Myers text cited above and in Luehrman, "Capital Projects as Real Options: An Introduction," HBS case no. 295-074.

pricing sooner rather than later is to build on, rather than abandon, the DCF-based NPV analysis your company already uses. Had we set out to value the option from scratch, it would have been more difficult and taken longer. It also would have been hard to tell how well we had done and when to stop working on it. Option pricing should be a complement to existing capital-budgeting systems, not a substitute for them. The framework presented here is a way to start where you are and get somewhere better.

1. There are other mathematical advantages, beyond the scope of this article, associated with using the quotient instead of the difference. Also, students of economics will recognize the similarity between NPVq and the famous "Tobin's q," which measures the ratio of the value of an asset to its replacement cost.

2. This form of option-pricing table is nearly as old as the Black-Scholes model itself. I first ran into it as an M.B.A. student using Richard Brealey and Stewart Myers's text, *Principles of Corporate Finance* (New York: McGraw-Hill, 1981).

In the September–October 1998 issue of HBR, Timothy Luehrman extends the framework presented here. In "Strategy as a Portfolio of Real Options," he explores how managers can use option pricing to improve the way they make decisions about the sequence and timing of strategic investments.

Reprint 98404
To place an order, call 1-800-988-0886.

CASE STUDIES AND *HARVARD BUSINESS REVIEW* ARTICLE REPRINTS

Many readers have asked for an easy way to order case studies and article reprints or to obtain permission to copy. In response, we have established a Customer Service Team to grant permission, send rush copies in paper form, deliver files in Acrobat (PDF) format electronically (*Harvard Business Review* articles only), or customize collections.

Please contact the Customer Service Team:

Phone: 617-496-1449
United States and Canada: 800-668-6780
(8 A.M. – 6 P.M. weekdays, voice mail after hours)
Fax: 617-496-1029 (24 hours, 7 days a week)
E-mail: custserv@hbsp.harvard.edu
(24 hours, 7 days a week)
Web Site: http://www.hbsp.harvard.edu

Prices (minimum order, $10):

Harvard Business Review Reprints
(Discounts apply to multiple copies of the same article.)

1–9 copies	$5 each
10–99	$4
100–499	$3.50
Electronic	$3.50 each

Harvard Business School Case Studies
$5 each

For quantity estimates or quotes on customized products, call Frank Tamoshunas at 617-495-6198.
Fax: 617-496-8866

PERMISSIONS

For information on permission to quote or translate Harvard Business School Publishing material, contact:

Customer Service Department
Harvard Business School
Publishing Corporation
60 Harvard Way
Boston, MA 02163

Phone: 617-496-1449
United States and Canada: 800-668-6780
Fax: 617-495-6985
E-mail: custserv@hbsp.harvard.edu

HARVARD BUSINESS REVIEW SUBSCRIPTION SERVICE

United States and Canada
Phone: 800-274-3214
Rates per year: United States, $85; Canada, U.S.$95

International and Mexico
Phone: 44-1858-435324
Fax: 44-1858-468969
Rates per year: international, U.S.$145; Mexico, U.S.$95
Orders, inquiries, and address changes:
Harvard Business Review
Tower House, Sovereign Park
Lathkill Street, Market Harborough
Leicestershire LE16 9EF
England

International customer service E-mail address: harvard@subscription.co.uk

Payments accepted: Visa, MasterCard, American Express; checks at current exchange rate payable to *Harvard Business Review*.
Bills and other receipts may be issued.

CATALOGS

Harvard Business School Publishing Media Catalog
This 32-page, full-color catalog features more than 40 management development video and interactive CD-ROM programs.

Harvard Business School Press
This latest full-color catalog features books for the fast-paced business world where you live and work.

Harvard Business School Publishing Catalog of Best-Selling Teaching Materials
This collection of teaching materials contains those items most requested by our customers.

Harvard Business School Publishing Catalog of New Teaching Materials
Designed for individuals looking for the latest materials in case method teaching.

2009 ROE Survey

Austerity Savings

Volatile economic conditions push regulators in new directions.

By Phillip S. Cross

A review of the rate cases decided over the past year indicates that the economy remains at the forefront in the news, and on the minds of regulators in rate-case proceedings. The issue has taken a new twist, however, as regulators are now placed in the unenviable position of determining an allowance for return on equity (ROE) that's fair to consumers and investors in a volatile economy. When *Fortnightly* presented this feature last year, we reported that regulators were seeking to determine the effect the dip in the stock market, falling interest rates and tightening credit might have on financial modeling, as well as subjective views of the return necessary to attract investors. This time, the cases that stand out are those in which regulators are exploring the limits of their discretion under the regulatory compact to balance the interests of consumers and shareholders in the face of a severe economic downturn.

The task of setting the return or profit component of regulated rates for utility service is one that begins with a review of mathematically derived estimates of the return expected by investors in the future. Regulators also are called on to use their informed judgement to produce a result that's fair to consumers and investors alike. The final answer often is expressed as a range of "reasonable" results that would at either end provide a fair return to investors and reasonable rates for consumers. This gives regulators some wiggle room when determining a final ROE figure or when seeking other ways to hold down rates for consumers—or to keep rates high enough to make sure a utility has access to capital.

A recent case decided in Michigan shows how the financial crisis might redound to the benefit of shareholders in a rate-case setting. In that case, the state public service commission (PSC) ruled that Detroit Edison's ROE should remain at 11 percent, even though its staff recommended a rate of 10.5 percent and other parties presented evidence supporting lower figures. The utility had asked for an allowance of 11.25 percent, a rate only slightly above the approved rate set in 2006. The PSC concluded that maintaining the status quo on the company's ROE in light of Michigan's economic circumstances and the U.S. credit crisis was the most prudent course of action. The commission said the worldwide crisis and ensuing breakdown in confidence among financial institutions led to rising long-term borrowing rates. It also noted that the credit-system freeze causes concern for the utility's continued ability to provide financing for infrastructure investment needs, and then to continue to provide safe, reliable and abundant power at reasonable rates. The PSC concluded that "a cautious approach in changing the company's ROE is necessary to ensure investor confidence and company access to capital markets" [*Re The Detroit Edison Co., Case No. U-15244, 270 PUR4th 134 (Mich.P.S.C.2008)*].



Discretionary Cuts

Regulators in Connecticut looked at the crisis another way. While setting rates for United Illuminating (UI), the Connecticut Department of Public Utility Control (DPUC) lowered the company's ROE from a level of 9.7 percent set in 1996 to 8.75 percent in a rate case heard this year. It rejected a claim by the electric utility that financial models relied on in the past should be adjusted to account for a change in investor behavior as a result of the crisis, including a shift away from looking at dividend payments as a measure of long-term growth and instead focusing more on earnings per share as a guide to investment decisions. The company claimed that dividend growth has remained stagnant due to heightened financial concerns in the utility industry. Expressing a keener interest in the macroeconomic issues at play, the DPUC concluded that although the overall outlook for the economy as a whole is weak, investors likely will continue looking to the utility sector as a safe haven amidst a volatile market

Phillip S. Cross is *Fortnightly's* legal editor. Email him at pcross@pur.com.

environment. The DPUC said that even though the company was embarking on a high volume of capital spending and infrastructure improvements, that would be offset by UI's strong financial position, limited risk profile, visible forward-earnings stream, high dividend yield, strong balance sheet and strong cash position. Despite higher spreads and yields, utilities still outperform most sectors of the bond market. As such, the cost of equity for the electric industry is among the lowest of all industries in the United States. All these indicators suggested a substantial decline in the overall equity cost rate, in the view of the DPUC [*Re The United Illuminating Co., Docket No. 08-07-04, Feb. 4, 2009 (Conn.D.P.U.C.)*].

Focusing directly on the plight of consumers during the current economic crisis, the DPUC in a second case reduced rates for a natural gas local distribution company (LDC) by $16.2 million, reflecting an allowed ROE of 9.31 percent. The department rejected claims by the utility that a rate increase was required due to current economic conditions that had resulted in nearly 15,000 residential service terminations due to non-payment of bills. Rather than hike rates to cover past-due bills, the current economic conditions required the LDC to share in the economic difficulties of Connecticut citizens by aggressively managing its operational expenses and capital investments, the department said. Driving home this point, the DPUC disallowed for rate-making purposes, costs incurred for non-qualified pension plans, finding that ratepayers shouldn't have to fund excessive pension benefits in difficult economic times [*Re Connecticut Natural Gas Corp.,274 PUR4th 345 (Conn.D.P.U.C. 2009)*].

In perhaps the most dramatic example of ratemaking meets an economy in crisis, the New York Public Service Commission (PSC) has in recent cases addressed consumer issues by imposing what it calls an "austerity savings" adjustment for energy utilities operating in the state. In those cases, the PSC actually increased the ROE in accordance with the results of financial models, but at the same time took away revenues by adjusting cost-of-service estimates to reflect the savings expected under mandated austerity savings programs. The PSC was careful to explain, however, that if the cost savings weren't found, the utility could petition for a deferral of the costs and possible recovery in a future rate period.

> When consumers are experiencing a harsh economic climate, regulators expect frugality from utilities.

For example, the PSC recently has approved a rate increase of $721 million for Consolidated Edison of New York. In that case, the PSC established an ROE of 10 percent for the utility, an increase from its earlier authorized ROE of 9.1 percent. The PSC reviewed several measures designed to reduce the level of the increase in the context of the current economic downturn. It determined that Con Edison should impose additional cost-cutting measures and directed the company to identify and implement an "austerity budget" that would reduce its revenue requirement by $60 million for the coming year.

The issue of the proper ROE remained separate from the austerity savings ruling, however. Through the trial briefing stage, the company supported an 11-percent equity return allowance but reflected only 10 percent in its May 2008 tariff filing. The PSC noted that it's unusual for a utility to support one equity return in testimony and to reflect a lower one in the revenue request set forth in its tariff filing, but accepted the filing and went on to examine the results of financial models presented in the case. The PSC did note that the revenue requirement difference between 10 percent and 11 percent was approximately $115 million a year. The company described its 10-percent request as part of its proposal to "ameliorate bill impacts on customers." The PSC went on to find that assigning a two-thirds weight to results under its own discounted cash flow analysis, and one-third weight to an average of the capital asset pricing model to the results presented by the parties to the case, showed that independent of the company's offer to settle for a 10-percent return based on its original request,

ROE Survey Methodology

Fortnightly's 2009 ROE survey covers cost of equity capital determinations by state public utility commissions during the period Sept. 1, 2008 through Aug. 31, 2009. (A few more recent cases outside the period are provided where available.) *Fortnightly's* methodology remains similar to its previous ROE surveys; requests for information on the results of recent rate proceedings were sent to both regulators and utility financial officials. In addition, direct examination of commission rate orders, when available, provided more information.

The traditional cost-of-service rate case remains the primary source of information on how utility regulators view the issue of shareholder earnings requirements. Nevertheless, performance-based rate plans and cases called to conduct periodic earnings reviews also contain findings about the appropriate ROE for utilities, and are reported in this survey. Explanatory notes accompany most entries, and citations are provided for orders published in *Public Utilities Reports, Fourth Series (PUR4th).*–**P.S.C.**

the same increase in the company's ROE proved to be the one that the evidence had shown to be adequate to compensate investors and attract capital in the near future.

Nevertheless, turning back to the issue of the macroeconomic concerns in the marketplace, the PSC stated that expenditures that are reasonable during average or good economic times aren't necessarily reasonable when economic conditions are extremely poor. When consumers are experiencing an extraordinarily harsh economic climate, a certain measure of frugality is properly expected from utilities and a reprioritizing of expenditures may be needed, the PSC said, citing such measures as freezing executive pay, restricting hiring, cutting travel costs and other so-called "discretionary" expenses. With this said, it ordered a downward adjustment to the company's revenue requirement amounting to $60 million, half of which will be subject to further review and potential deferral based on a review of the company's ability and best efforts to implement the required measures effectively. It pointed out that this amounts to approximately 3.6 percent of non-fuel operation and maintenance costs and emphasized that the company's management will be responsible for determining how best to achieve the $60 million revenue requirement reduction while maintaining reliability, service quality, and safety [*Re Consolidated Edison Co. of New York, Inc., Cases 08-E-0539, 08-M-0618, Apr. 24, 2009 (N.Y.P.S.C.); See also Central Hudson Gas & Electric Corp., 274 PUR4th 257 (N.Y.P.S.C. 2009) where the PSC also discussed macroeconomic conditions that may be used as a basis for requiring the so-called austerity adjustments to a company's revenue requirements*].

Company	Utility Type	Case, Docket or Decision	Application Date	Order Date	Test-year End Date	Increase (Decrease) Requested ($Million)	Increase (Decrease) Granted ($Million)	Rate of Return on Common Equity	
								Previously Authorized Rate (%)	Newly Authorized Rate (%)
Arkansas									
Oklahoma Gas & Electric	Electric	08-103-U	8/29/08	5/20/09	12/31/07	26.4	13.6	10	10.25
Arizona									
Southwest Gas	Gas	G-01551A-07-0504, 270 PUR4th 465	8/31/07	12/24/08	4/30/07	50.2	33.5	9.5	10
California									
Southwest Gas - Northern Jurisdiction	Gas	07-12-022	12/21/07	11/21/08	12/31/09	-0.1	-1.04	10.9	10.5
Southwest Gas - South Lake Tahoe Jurisdiction	Gas	07-12-022	12/21/07	11/21/08	12/31/09	2.1	1.82	NA	10.5
Southwest Gas - Southern Jurisdiction	Gas	07-12-022	12/21/07	11/21/08	12/31/09	7.1	2.44	10.9	10.5
Colorado									
Black Hills Energy	Gas	08S-2906	6/30/08	3/10/09	12/31/07	2.7	1.38	12	10.25
Public Service of Colorado	Electric	08S-520E, C09-0595, 275 PUR4th 149	11/25/08	6/9/09	12/31/09	174.4	12.0	10.5	10.5[1]
Connecticut									
Connecticut Natural Gas	Gas	08-12-06, 274 PUR4th 345	1/16/08	6/30/09	6/30/08	16.4	-16.2	10.1	9.41[2]
Southern Connecticut Gas	Gas	08-12-07	1/20/08	7/17/09	6/30/10	50.1	-12.46	10	9.26
United Illuminating	Electric	08-07-04	8/8/08	6/3/09	12/31/07	52.2	25.27[3]	9.75	8.75
Delaware									
Chesapeake Utilities	Gas	07-186, 268 PUR4th 370	7/6/07	9/2/08	3/31/07	1.9	0.329	10.86	10.25
Florida									
Peoples Gas System	Gas	080318-GU, 274 PUR4th 177	8/1/08	6/9/09	12/1/09	26.5	19.2	11.25	10.75
Tampa Electric	Electric	080317-EI	8/1/08	4/30/09	12/1/09	228.2	147.3	11.75	11.25
Idaho									
Atlanta Power	Electric	ATL-E-08-02, 271 PUR4th 134	5/1/08	12/19/08	12/31/06	0.109	0.106	NA	12[4]
Avista	Electric	AVU-E-08-01	4/3/08	9/30/08	12/31/07	32.3	23.20*	NA	10.2
Avista	Gas	AVU-G-08-01	4/3/08	9/30/08	12/31/07	4.7	3.90*	NA	10.2
Rocky Mountain Power	Electric	PAC-E-08-07	9/19/08	4/16/09	12/31/07	5.9	4.38	10.25	10.25[5]
Illinois									
Commonwealth Edison	Electric	07-0566, 268 PUR4th 1	10/17/07	9/10/08	12/31/06	362.3	273	10.045	10.3
Northern Illinois Gas	Gas	08-0363, 272 PUR4th 161	4/29/08	3/25/09	12/31/09	140.4	69	8.85	10.11

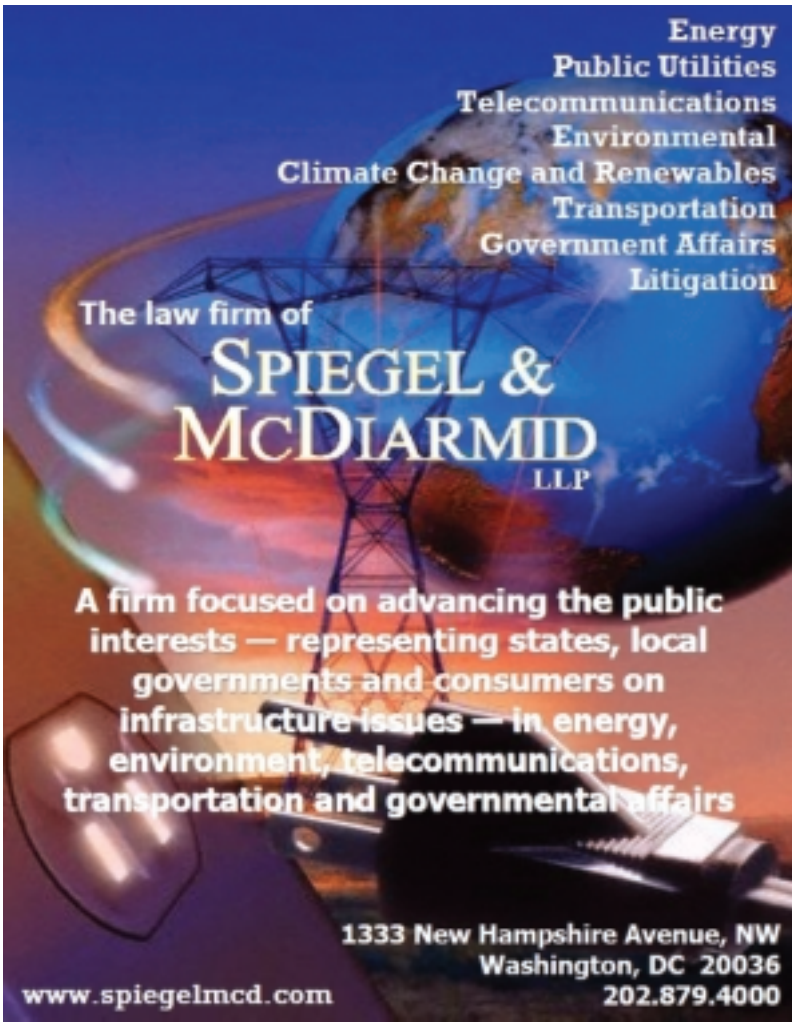
Company	Utility Type	Case, Docket or Decision	Application Date	Order Date	Test-year End Date	Increase (Decrease) Requested ($Million)	Increase (Decrease) Granted ($Million)	Rate of Return on Common Equity	
								Previously Authorized Rate (%)	Newly Authorized Rate (%)
Indiana									
Indiana Michigan Power	Electric	43306, 273 PUR4th 310	6/19/07	3/4/09	9/30/07	125.5	41.6	12	10.5
Indiana Natural Gas	Gas	43434	1/30/08	10/8/08	8/31/07	0.7	0.5	10.4	10.2
Indiana Utilities	Gas	43520	6/17/08	1/21/09	12/31/07	0.4	0.4	10.4	10.3
Iowa									
Black Hills Energy	Gas	RPU-08-3, 275 PUR4th 44	6/2/08	6/3/09	12/31/07	13.56	10.39*	NA*	10.1
Interstate Power & Light	Electric	RPU-08-1	3/31/08	2/13/09	NA	NA	NA	NA	10.1[6]
MidAmerican Energy	Electric	RPU-08-4	7/23/08	8/27/08	NA	NA	NA	NA	11.7[7]
Kansas									
Westar Energy	Electric	08-WSEE-1041-RTS	5/28/08	1/21/09	12/31/07	90	65	10	10.4[8]
Kentucky									
Kentucky Utilities	Electric	2007-00565	7/1/08	2/5/09	4/30/08	22.2	-8.9*	11	10.63
Louisiana									
Cleco Power LLC	Electric	U-21496-K	6/1/09	NA	9/30/08	NA[9]	NA	11.25	11.25
Cleco Power LLC	Electric	U-30689	7/14/08	NA	6/30/09[10]	250.1	173.3	11.25	NA[11]
Massachusetts									
New England Gas	Gas	08-35, 271 PUR4th 1	7/17/08	2/2/09	12/31/07	5.6	3.68	NA	10.05
Michigan									
Detroit Edison	Electric	U-15244	4/13/07	12/23/08	12/31/06	123	83.629	11	11
Michigan Gas Utilities	Gas	U-15549	5/16/08	1/13/09	12/31/09	13.9	6	11.4	10.45
Minnesota									
Minnesota Energy Resources	Gas	GR-08-835	7/31/08	6/29/09	12/31/08	22	15.4	11.71	10.21
Minnesota Power	Electric	E-015/GR-08-415	5/2/08	5/4/09	6/30/09	39.8	20.4	11.6[12]	10.74
Missouri									
AmerenUE	Electric	ER-2008-0318, 271 PUR4th 475	4/4/08	1/27/09	3/31/08	251	162	10.2	10.76
Kansas City Power & Light	Electric	ER-2009-0089	9/5/08	6/10/09	12/31/07	101.5	95	10.75	NA*
Kansas City Power & Light (Greater Missouri Operations[13])	Electric	ER-2009-0090	9/5/08	6/10/09	12/31/07	83.1	63	10.25	NA*
Mississippi									
CenterPoint Energy	Gas	Rider RRA[14]	1/2/08	11/17/08	9/30/07	2.5	0.9	9.86	9.67
Montana									
Montana-Dakota Utility	Electric	D2007.7.79, 264 PUR4th 516	7/12/07	4/23/08	12/31/06	7.77	4.1	12.3	10.25
NorthWestern Energy	Electric	D2007.7.82	7/31/07	7/7/08	12/31/06	31.37	10	10.75	NA*
NorthWestern Energy	Gas	D2007.7.82	7/3/07	7/7/08	12/31/06	10.5	5	10.75	NA*
NorthWestern	Electric	D2008.6.69, 269 PUR4th 277[15]	6/27/08	11/13/08	NA				10[16]
Nevada									
Nevada Power dba NV Energy	Electric	08-12002	12/1/08	6/24/09	6/30/08	324	224	10.7	10.5
New Jersey									
New Jersey Natural Gas	Gas	GR07110889	11/20/07	10/3/08	4/30/08	58.36	32.5	11.5	10.3
New Mexico									
Southwestern Public Service	Electric	08-00354-UT	12/18/08	7/14/09	6/30/08	24.6	14.2	10.5	NA*
Zia Natural Gas	Gas	08-00036-UT	1/31/08	11/25/08	8/31/08	3.2	2.5	NA	10.27[17]
New York									
Central Hudson Gas & Electric	Electric	08-E-0887	7/31/08	6/22/09	6/30/10	66.1	38[18]	9.6	10
Central Hudson Gas & Electric	Gas	08-G-0888, 274 PUR4th 257	7/31/08	6/22/09	6/30/10	20.2	13.6[18]	9.6	10
Consolidated Edison of New York	Electric[19]	08-E-0539	5/9/09	3/24/09	3/31/10	935.1	721.4	9.1	10
Corning Natural Gas	Gas	08-G-1137	9/24/08	8/20/09	6/30/10	1.7[20]	0.973[21]	10[22]	10.7[23]
North Carolina									
Piedmont Natural Gas	Gas	G-9, Sub 550, 269 PUR4th 320	2/29/08	10/24/08	12/31/08	40.51	15.68	NA	10.6[24]
North Dakota									
Northern States Power	Electric	PU-07-776, 271 PUR4th 333	12/7/07	1/14/09	12/31/08	17.9	10.9	12	10.75

Company	Utility Type	Case, Docket or Decision	Application Date	Order Date	Test-year End Date	Increase (Decrease) Requested ($Million)	Increase (Decrease) Granted ($Million)	Rate of Return on Common Equity	
								Previously Authorized Rate (%)	Newly Authorized Rate (%)
Ohio									
Cleveland Electric Illuminating	Electric	07-552-EL-UNC[25]	5/8/07	1/21/09	2/29/08	108.6	29.2	NA	10.5[26]
Columbia Gas of Ohio	Gas	08-72-GA-AIR	3/3/08	12/3/08	9/20/08	87.81	47.14	NA	10.39
Duke Energy Ohio	Electric	08-709-EL-AIR	7/25/08	7/8/09	12/31/08	85.6	55.3	10.29	10.63
East Ohio Gas dba Dominion East Ohio	Gas	07-829-GA-AIR	10/30/07	10/15/08	12/31/07	76.02	37.48	12.15	9.92
Ohio Edison	Electric	07-552-EL-UNC[25]	5/8/07	1/21/09	2/29/08	160.7	68.9	NA	10.5
Toledo Edison	Electric	07-552-EL-UNC[25]	5/8/2007	1/21/09	2/29/08	70.5	38.5	NA	10.5
Vectren Energy Delivery of Ohio	Gas	07-1081-GA-AIR	11/20/07	1/7/09	5/31/08	27.33	14.78	10.60	10.65
Oklahoma									
CenterPoint Energy	Gas	PUD-20900055	3/13/09	7/9/09	12/31/07	1.9	1.5	10.25	10.5
Oregon									
Portland General Electric	Electric	UE 197	2/27/08	1/1/09	12/31/09	146	120	10.1	10.1
Pennsylvania									
Columbia Gas of Pennsylvania	Gas	2008-2011621	1/28/08	10/28/08	9/30/08	58.9	41.5	NA*	NA*
Equitable Gas	Gas	R-2008-2024325	6/30/08	2/26/09	12/31/08	51.949	38.35[27]	NA*	NA*
Rhode Island									
Narragansett Electric	Gas	3943, 272 PUR4th 96	4/1/08	1/29/09	9/30/07	20.04	13.66	NA	10.5
South Dakota									
Otter Tail Power	Electric	EL08-030	10/31/08	6/30/09	12/31/07	3.9	3	11.75[28]	NA*
Texas									
Entergy Texas	Electric	34800	9/26/07	3/16/09	3/31/07	605	46.7	10.95	10
Oncor Electric Delivery	Electric	35717	6/27/08	8/31/09	12/31/07	253.5	115.1	11.25	10.25
Southwestern Public Service	Electric	35763	6/12/08	6/2/09	12/31/07	84.2	57.4	11.50	NA[29]
Texas-New Mexico Power	Electric	36025	8/20/08	8/21/09	3/31/08	8.7	6.8	11.25	10.25*
Utah									
Rocky Mountain Power, a division of PacifiCorp	Electric	08-035-38	4/7/08	4/21/09	12/31/07	160.6	45.0	10.25	10.61[24]
Vermont									
Central Vermont Public Service	Electric	7485	10/31/08	2/13/09	12/31/07	0.94	NA	10.21	9.77[30]
Virginia									
Appalachian Power[31]	Electric	PUE-2008-00045	5/30/08	10/15/08	12/31/07	66.5	60.6	10.2	10.2[24]
Appalachian Power	Electric	PUE-2008-00046	5/30/08	11/17/08	12/31/07	207.9	167.9	10	10.2[24]
Atmos Energy	Gas	PUE-2008-00007	3/6/08	9/30/08	9/30/07	0.9	0.9	10	10[24]
Roanoke Gas	Gas	PUE-2008-00088	9/16/08	6/10/09	6/30/08	1.2	1.2	10.1	10.1[24]
Washington									
Avista	Electric	UE-080416	3/4/08	12/29/08	12/31/07	36.6	32.5	10.2	10.2
Avista	Gas	UG-080417	3/4/08	12/29/08	12/31/07	6.6	4.8	10.2	10.2
Northwest Natural Gas	Gas	UG-080546	3/28/08	12/26/08	9/30/07	4.3	2.72	NA	10.1
PacifiCorp	Electric	UE-080220	2/6/08	10/8/08	6/30/07	34.9	20.4	10.2	10.2
Puget Sound Energy	Electric	UE-072300	12/3/07	10/8/08	9/30/07	174.8	130.2	10.4	10.15
Puget Sound Energy	Gas	UG-072301	12/3/07	10/8/08	9/30/07	56.8	49.2	10.4	10.15
Wisconsin									
Northern States Power	Electric	4220-UR-115, 264 PUR4th 236	6/1/07	1/8/08	12/31/08	67.4	39.4	11	10.75
Northern States Power	Gas	4220-UR-115, 264 PUR4th 236	6/1/07	1/8/08	12/31/08	5.3	5.3	11	10.75
Wisconsin Power & Light	Electric	6680-UR-116	2/22/08	12/30/08	12/31/09	93.3	0	10.8	10.8[32]
Wisconsin Power & Light	Gas	6680-UR-116	2/22/08	12/30/08	12/31/09	0.8	-3.9	10.8	10.8[32]
Wisconsin Public Service	Electric	6690-UR-119, 270 PUR4th 421	4/1/08	12/30/08	12/31/09	84.8	0	10.9	10.9[32]
Wisconsin Public Service	Gas	6690-UR-119, 270 PUR4th 421	4/1/08	12/30/08	12/31/09	15.7	-3	10.9	10.9[32]
Wyoming									
Rocky Mountain Power	Electric	20000-333-ER-08, 275 PUR4th 127	NA	5/20/09	NA	NA	18	10.25	10.25[33]
Wyoming Gas	Gas	30009-48-GR-08	10/17/08	7/1/09	5/31/08	1.61	1.31		10.87
Questar Gas	Gas	30010-94-GR-08	NA	7/1/09	12/31/08	.483	.378	NA	10.5

NOTES

* Settlement agreement. No ROE figure stated.

1. Parties to approved settlement agree to use currently authorized ROE of 10.5% to calculate new rates and for future regulatory filings through 12/31/09.
2. Authorized rate shown is reduced by 10-basis points in setting final revenue requirement for management imprudence concerning billing problems.
3. Revenue amount is award for final two years of company's four-year rate plan.
4. Commission states that 12% rate is equal to return allowed for other small utilities in the state.
5. No ROE stated in settlement agreement on commission order reflecting significant difference in party positions. Parties agree to use existing rate of 10.25% for other purposes such as avoided cost filings.
6. Result from an "advanced ratemaking principles" case involving a coal-fired generating facility.
7. Result from an "advanced ratemaking principles" case (settlement) involving wind-power projects.
8. Figure reflects specific finding by commission regarding appropriate ROE for future environmental cost recovery filings.
9. Proceeding to review level of earnings under established formula rate plan.
10. Test year utilized is actual year ending Sept. 30, 2007 with *pro formas* to June 30, 2009 and Rodemacher Unit No. 3 full year operations.
11. ROE range of 10.7% to 11.3%.
12. Figure shown contained in rate case order decided in 1994.
13. Formerly Aquila Electric Operations.
14. Rate Regulation Adjustment Rider. Formula rate mechanism features an annual recalculation of the allowed ROE and a graduated sharing of earnings above the authorized figure.
15. Order authorizing an electric utility to include in rate base, at a value of $407 million, its interest in a coal-fired generating plant.
16. Revenue requirement for the life of the plant is based on the ROE figure shown.
17. Order states that the ROE figure adopted only for the current rate case has no precedential value. Commission cites as a basis for this comment the utility's small size relative to other gas utilities in the state, as well as its status as a privately held entity.
18. Figures reflected downward "austerity adjustment" to revenue requirement of $2.4 million for electric service and $.6 million for natural gas. Adjustment reflects recognition of economic downturn. Adjustment will have no effect on ROE if utility defers or reduces expenses by an equal amount.
19. Delivery Service for full-service and retail-access customers.
20. Annual increase for year ending June 30, 2010. Joint proposal submitted supplants rate application and calls for a three-year rate plan with earnings sharing adjustment mechanism.
21. Rate period 9/1/09 through 8/31/10.
22. Settlement approved 5/22/06.
23. If utility elects to file for new rates prior to 10/1/2010, ROE resets to 10.5%.
24. Figure included in approved settlement agreement.
25. Rate filing for First Energy operating companies' electric distribution services.
26. Midpoint of range of 10% to 11% adopted by commission.
27. Total award includes a separate adjustment for low-income customer assistance program, which is reconciled annually.
28. Figure shown approved in 1987 rate order.
29. Settlement agreement. 10.25% ROE set for calculation of allowance for funds used during construction during rate-effective period.
30. Calculated using formula set forth in alternative regulation plan.
31. Proceeding concerning environmental and system reliability costs only.
32. Company reports that figure shown is implied as part of revenue settlement process. ROE not stated in commission order.
33. Approved revenue stipulation does not state authorized ROE. Commission found it appropriate to continue to use 10.25% ROE adopted in the company's last rate case for purposes of evaluating earnings on a prospective basis.





Utility Regulatory News, the weekly electronic newsletter dedicated to keeping you up to date on state PUC developments delivered every Friday afternoon. To order call 800-368-5001
E-mail: *pur@pur.com* or Online: *www.pur.com*

PUBLIC POLICY FOR THE
Privatesector
The World Bank Group September 1996

Note No. 87

Price Caps, Rate-of-Return Regulation, and the Cost of Capital

Ian Alexander and Timothy Irwin

This Note compares the effects of price cap and rate-of-return regulation on the risks borne by regulated utilities. It presents evidence that price cap regulation subjects firms to greater risks and therefore raises their cost of capital. This result has one clear implication: firms regulated by price caps must be permitted to earn higher returns. If they are not, they will be unable to attract new investment capital and the quality of their service will decline.

Price caps and rate-of-return regulation

There are two main approaches to preventing monopolistic infrastructure firms from charging excessively high prices: price cap regulation and rate-of-return regulation. The rate-of-return approach is used in Canada, Japan, and the United States, where regulatory agencies fix the rate of return that a utility can earn on its assets. They set the price the utility can charge so as to allow it to earn a specified rate of return—and no more. The regulated price can be adjusted upward if the utility starts making a lower rate of return, and it will be adjusted downward if the utility makes a higher rate.

Over the past decade or so, the price cap approach has become increasingly common internationally because it is thought to give firms stronger incentives to be efficient. Under this approach, the regulated price is adjusted each year by the rate of inflation plus or minus some predetermined amount and without regard to changes in the firm's profits. In the United Kingdom, for example, utilities are permitted to increase their prices by the change in the consumer price index plus or minus a specified amount. In gas and electricity, the price-setting rule is called $RPI - X$, where RPI is the retail price index and X represents the expected annual gain in the utility's efficiency. In water, the rule is $RPI + K$, where K represents both expected productivity gains and a permitted annual increase in the real price of water to allow for quality improvements (think of it as $RPI - X + Q$, where Q stands for the quality improvement). Since 1989, price caps have also been used in the United States to adjust the prices charged by the long-distance telephone company AT&T. In New Zealand, a price cap is used to adjust Telecom New Zealand's rental charge for a residential phone line. Price caps are also used in some developing countries. Malaysia, Mexico, and Peru, for example, use them for telecommunications, and Argentina uses them for gas and electricity as well.

In practice, price cap and rate-of-return regulation are less different than they might seem. First, a rule like $RPI - X$ considers only how prices should be changed from year to year; it doesn't tell a regulator how to set them in the first year. A regulator wanting to use price cap regulation for a new service would need to set the initial price in some way, and one obvious option is to consider the price the firm needs to charge to earn a satisfactory rate of return. Second, a price cap needs to be periodically reviewed: a regulator cannot reliably predict what changes in productivity will be possible in, say, ten years. In the United Kingdom, price caps typically are reviewed every five years. And during a review,

TABLE 1 AVERAGE INFRASTRUCTURE FIRM BETAS, BY COUNTRY, SECTOR, AND TYPE OF REGULATION, 1990–94

	Electricity		Gas		Combined gas and electricity		Water		Telecoms	
Country	Regulation	Beta	Regulation	Beta	Regulation	Beta	Regulation	Beta	Regulation	Beta
Canada	—	—	—	—	ROR	0.25	—	—	ROR	0.31
Japan	ROR	0.43	—	—	—	—	—	—	ROR	0.62
Sweden	—	—	—	—	—	—	—	—	Price cap	0.50
United Kingdom	—	—	Price cap	0.84	—	—	Price cap	0.67	Price cap	0.87
United States	ROR	0.30	ROR	0.20	ROR	0.25	ROR	0.29	Price cap (AT&T)	0.72
									ROR (others)	0.52

— Not available or not applicable.

Note: The betas are asset betas that control for differences in debt-equity ratios between firms. ROR is rate-of-return regulation.

Source: Oxford Economic Research Associates, "Regulatory Structure and Risk: An International Comparison" (London, 1996).

the regulator naturally takes into account the regulated utility's rate of return. If it is high, the price cap is likely to be reduced; if it is low, the price cap may be relaxed.

But as long as price cap reviews are sufficiently infrequent (say, every five years), price cap and rate-of-return regulation should have different effects on regulated firms. In particular, a price cap subjects businesses to more risk. For example, under price cap regulation, if a firm's costs rise, its profits will fall because it cannot raise its prices to compensate for the cost increases—at least until the next price review, which may be several years away. Under rate-of-return regulation, however, the business would seek—and typically be granted within a year or so—a compensating price rise, so its profits would not change much. But if the firm's costs fall, price cap regulation is more advantageous to the firm than rate-of-return regulation, because it would retain more of the resulting benefits as profits. Thus, under rate-of-return regulation, consumers bear some of the risk that firms bear in price cap systems. This difference in impact means that firms subject to price cap regulation have a stronger incentive to lower their costs because they keep more of the cost savings than they would if they were subject to rate-of-return regulation. But the increased risk they bear tends to raise their cost of capital.

Empirical evidence on risk and the regulatory system

The risk that affects a firm's capital cost can be measured by a statistic called the firm's beta. Betas are used by investors worldwide and are an important factor in their decisionmaking. A firm's beta measures the extent to which the firm's returns vary relative to those of a diversified portfolio of equity holdings. It indicates whether an investor with a diversified portfolio would take on more risk by investing in a particular firm. The higher the beta, the bigger the increase in the riskiness of the investor's portfolio.

Several studies that compared the betas of British firms subject to price cap regulation with those of U.S. firms subject to rate-of-return regulation found that the U.S. firms have lower betas, as expected. But the results leave room for doubt because it is unclear whether it is the difference in regulation that's at work or something else, such as a difference in the degree of competition in the British and U.S. markets. But new empirical work done by Oxford Economic Research Associates for the World Bank tends to confirm the earlier conclusions. This study measured the betas of more than 100 infrastructure firms in many countries. Table 1 summarizes the results of the study, by country, for companies subject to price cap or rate-of-return regulation. (Some countries in the study have

been omitted from the table because they use discretionary regulatory regimes that do not follow a price cap or rate-of-return rule, or because the data were not comparable.) The results show that price cap regulation is associated with higher betas than rate-of-return regulation in Canada, Japan, and Sweden, as well as in the United Kingdom and the United States. Rate-of-return regulation is associated with betas ranging from as little as 0.2 in the U.S. gas industry to 0.62 in Japanese telecommunications, while price cap regulation is associated with betas ranging from 0.5 in Swedish telecommunications to 0.87 in British telecommunications. Overall, and as explained below, the differences in betas imply that firms subject to price cap regulation have to pay about an extra percentage point for their capital.

Why betas matter

To understand why betas matter, note that different firms face different costs of capital. Some firms must offer an expected rate of return of, say, 20 percent to attract investors, while others can get all the money they need by offering only 15 percent. Although the precise reasons for these discrepancies are not known with confidence, one critical factor is risk. Investors tend to be risk averse: other things equal, they prefer safer investments to risky ones. That means that firms have to compensate them for taking on more risk.

Investment risk, in the sense in which it is used here, relates only to bottom-line profits—the net impact on a firm's profits of all the separate risks facing the firm, such as operating risk, inflation risk, interest rate risk, foreign exchange risk, and political risk. Investment risk is not all downside. Risky projects are those with both a higher-than-average chance of turning out exceptionally badly and a higher-than-average chance of turning out well. Thus, when investors say they want to be compensated for taking on risk, what they mean is not just that they prefer an investment with a certain return of 10 percent to one that will probably make 10 percent but might make less. They mean that they prefer the safe 10 percent return to an investment offering, say, equal chances of 5 percent and 15 percent returns. Investment risk, then, has to do with the variability of returns.

Much investment risk can be eliminated by diversification. To see why, consider a racetrack analogy. Suppose you have no information on how fast the horses can run. You could bet all your money on one horse, or you could bet a little on each horse in the race. The two strategies have about the same expected, or average, return: two people, each using one of the two strategies for, say, a thousand races, would probably end up with roughly the same amount of money. For any one race, however, the two strategies pose different degrees of risk. The strategy of betting on just one horse is riskier: you could do well, but you're more likely to lose everything you bet. But when you bet on every horse, you almost certainly will lose a little, because the racetrack has to make a profit.

Under rate-of-return systems, consumers bear some of the risks that firms bear in price cap systems

As with betting on horses, investing in many firms eliminates much risk without significantly reducing the expected return. Thus, professional investors do not worry about the sort of risks that can be eliminated by portfolio diversification. But the risks of professional investment are different from those in racetrack betting. At the racetrack, you can eliminate almost all investment risk by betting on every horse. The same isn't true of investing in firms. Some years are good, and in these years, most firms do well. In other years, most firms do badly. So, on average, firms' returns tend to move in the same direction, and even if you've invested in every firm, the return on your portfolio is uncertain. This risk that remains after diversification is the risk that professional investors are most concerned about.

Professional investors are particularly interested in the likelihood that a firm's returns will move

with the returns on a completely diversified portfolio—that is, a portfolio that includes investments in enough firms so that further diversification would not significantly reduce risk. In one possible scenario, a firm's returns might be expected to vary in equal proportion to the diversified portfolio, so that, for example, when the returns on the portfolio increase by 10 percent, the returns on the investment also are expected to increase by 10 percent. In this scenario, beta equals 1, and the investment neither increases nor reduces the total riskiness of an investor's portfolio. As a result, investors will demand a moderate rate of return when investing in the firm, and the firm's cost of capital will be moderate.

In another scenario, a firm's returns might vary disproportionately with those of the diversified portfolio, so that a 10 percent increase in the portfolio's returns would be associated with, say, a 20 percent increase in the firm's returns, and a 10 percent decrease in the portfolio's returns with a 20 percent decline in the firm's. Here, beta equals 2. Because investing in such firms increases total risk, investors demand an above-average rate of return as compensation, and capital costs these firms more than it does the average firm.

In a third scenario, a firm's returns might vary less strongly with those of the diversified portfolio, with a 10 percent increase in the portfolio's returns associated on average with, say, a 5 percent increase in the firm's returns. Here, beta equals 0.5. Because investing in such firms reduces total risk, investors are willing to give up some return to invest in them. For these firms, the cost of capital is lower than average.

Betas and regulation revisited

Equipped with this measure of investment risk and the cost of capital, consider the returns available from investing in a utility subject to rate-of-return regulation. Because prices are adjusted each year to keep the rate of return roughly constant, investments in the firm are subject to little risk, particularly the market-related risk that investors worry about. If returns in the market as a whole rise, the regulated utility's returns won't rise much (though they can rise a little in the period before the regulator requires a price cut). But if the market turns bad and returns fall, the utility's returns won't fall below the target set by the regulator for long. Thus, firms subject to rate-of-return regulation tend to have low betas and a lower-than-average cost of capital.

Price cap regulations don't have the same effect. Because in the short run the regulator sets no target rate of return, the regulated company's return can vary from period to period and is free to vary with the returns on the market. Even under price cap regulation, utility firms often have a fairly safe business, with returns that are affected less by economywide shocks than are those of other firms. As shown in table 1, their betas are still lower than 1, the average for all firms. But they are higher than the betas of firms subject to rate-of-return regulation. So investors will demand a higher return for investment in a firm subject to price cap regulation.

Conclusion

This does not imply that price caps are less desirable than rate-of-return regulation. It simply means that regulators need to take account of the effect of regulation on the cost the regulated firm has to pay investors for capital. Regulators using rate-of-return regulation can set the target rate of return lower than that earned by the average firm and still expect investors to be interested, because the returns are subject to less risk than those of an average firm. Regulators using price cap regulation need to give firms under their jurisdiction the opportunity to make somewhat higher returns, because those returns are riskier. If they don't, the firms will be unable to attract new investment capital, and the quality of their service will eventually suffer.

This Note is based on work by Ian Alexander at Oxford Economic Research Associates.

Ian Alexander, London Economics, London, and Timothy Irwin (tirwin@worldbank.org), Private Sector Development Department

The Note series is an open forum intended to encourage dissemination of and debate on ideas, innovations, and best practices for expanding the private sector. The views published are those of the authors and should not be attributed to the World Bank or any of its affiliated organizations. Nor do any of the conclusions represent official policy of the World Bank or of its Executive Directors or the countries they represent.

To order additional copies please call the FPD Note line to leave a message (202-458-1111) or contact Suzanne Smith, editor, Room G8105, The World Bank, 1818 H Street, NW, Washington, D.C. 20433, or Internet address ssmith7@worldbank.org. Previous issues are also available on-line (http://www.worldbank.org/html/fpd/notes/notelist.html).

♻Printed on recycled paper.

Regulation by Formula

Tools to facilitate changing utility economics.

By Michael Schmidt



These are challenging times for the electric and gas utilities. Reliability projects, renewable portfolio standards, greenhouse-gas emissions control, AMI, smart-grid investments, and conservation programs—all these things add to costs, but might bring in no additional revenue. Moreover, there will be unprecedented capital investment in transmission, renewable generation projects, and replacement of old facilities from the 1950s and 1960s. Thus, earnings likely will be more closely watched and traditional general rate cases might not be able to keep up.

Regulatory Options

There are a number of regulatory tools available to address rapid changes in utility operating and capital costs. First is the possibility of using a projected test year instead of a historical test year; but there's nothing really new here, and it still requires a general rate case. A full blown general rate case is time consuming and costly for the utility, the regulator, and intervenors. Cost estimates for a projected test year add to disputes among the parties.

A second option, decoupling commodity sales from base rate revenues serves to tie base rates to the number of customers, equivalent to allowing the utility to recover a fixed amount of money per customer regardless of sales. A balancing account tracks any over or under-collections during the year. The balancing account is closed out by adjusting base rates in the next year. Any commodity costs (*e.g.*, fuel, purchased power, natural gas) are billed to the customer as a separate line item on the bill. However, decoupling does nothing to incorporate increased operating costs or increases in rate base.

A third option, performance-based regulation or ratemaking (PBR) attempts to adjust rates annually in response to a pre-specified escalation factor reduced by a productivity factor. The formula for PBR is: $Pt = Pt\text{-}1\ ((1 + esc.) - X)$, where *Pt* equals the new price in time period *t* after adjustment; *Pt-1* equals the old price in the prior time period *t-1* before adjustment; *esc.* equals a predetermined escalator, for example the consumer price index (CPI); and *X* equals a predetermined productivity factor that serves to reduce the escalator and account for productivity improvements in the time period *t-1*.

PBR, although it provides for a rate increase each year, might or might not cover all operating cost increases, cost of capital increases, or significant increases in rate base. Moreover, absent decoupling, declines in sales likely will result in a revenue requirement shortfall. In addition, the *X* factor is difficult to establish and the escalator not necessarily correlated with utility cost increases. Finally, the PBR usually has a three- to five-year time limit before another full-blown rate case is triggered.

A fourth option, ratemaking by formula, allows for changes in rates to reduce or increase rates in response to changes in earnings and rate base.

Planning for Change

Ratemaking by formula addresses two types of issues: those relative to earnings, and those relative to revenues. In summary, a rate-of-return band is established around the allowed return on equity (ROE) from the utility's last general rate case. Earnings outside the band are shared with customers through rate

---

Michael Schmidt has held senior management positions with electric and gas utilities, consulting firms and a state regulatory agency. He wrote several books on utility issues, including PUR's *Performance-Based Ratemaking: Theory and Practice* (2000). Email him at michaelrschmidt@msn.com.

decreases for excess earnings and rate increases for deficient earnings. The revenue aspect of the model incorporates new assets and other significant cost items into rates without a full blown general rate case. Significant sales reductions can be incorporated into rates, as well as revenue losses from conservation activities.

Under ratemaking by formula, earnings up to a pre-defined level are retained by the utility. Earnings in excess of that level are shared with customers by reducing rates—referred to as asymmetrical ratemaking. Earnings below a pre-defined level result in an increase in rates to bring earnings up to the pre-defined level—referred to as symmetrical ratemaking.

Ratemaking tied to earnings isn't new. For example:

- *Sheffield Gas Act* approved by the British Parliament in 1855 limited dividends when price of gas higher than a preset level, but allowed a higher dividend when prices were below a preset level;
- Consumers' Gas of Toronto in 1877 refined the process to identify a specific return target allowing for price decreases above the target and price increases below the target (included a $1 million dead band); and
- Recent earnings-based ratemaking by formula applications include New Mexico 1975; MidAmerican (Iowa) 2001; Entergy LA 2005; Entergy Gulf States 2006; Entergy Mississippi 2006; and Southwestern Electric Power Co. 2007.

The mechanics of the earnings side of ratemaking by formula is tied to the allowed rate of ROE. The calculation is a three-step process: First, establish earnings targets; second, calculate jurisdictional regulated income; and third, determine whether jurisdictional regulated income exceeds either or both earnings targets.

The earnings targets are applicable to earnings between the allowed ROE (Target ROE) as established by the state regulatory commission up to a preset number of basis points above that target (*i.e.*, 80 basis points). The first target for an asymmetrical application is calculated as follows: *Equity Portion of Jurisdictional Rate Base x Allowed ROE = Equity Earnings Target 1*. The second target is calculated as follows: *Equity Portion of Jurisdictional Rate Base x (Target ROE + 0.8 percent) = Equity Earnings Target 2*. The first target for a symmetrical application (Target 3) is the same as for asymmetrical applications. Target 3 is calculated as follows: *Equity Portion of Jurisdictional Rate Base x Allowed ROE - 0.8 percent = Equity Earnings Target 3.*

The next step in the process is the determination of earnings sharing with customers through pre-established sharing bands. For example: earnings between Target 2 plus 0.5 percent are shared 50:50 between shareholders and customers; earnings between Target 2 plus 0.5 percent up to 1 percent are shared 60 percent with customers, 40 percent with shareholders; and earnings above Target 2 plus 1.5 percent go to customers.

For example, with an ROE of 10 percent, Target 2 is 10.8 percent establishing a 10-percent to 10.8-percent dead band with all earnings going to shareholders; the earnings from an ROE of 10.8 percent to 11.3 percent are shared 50:50 between customers and shareholders; an ROE of 11.3 percent to 12.3 percent is shared 60-percent customers, 40-percent shareholders; and all additional earnings above 12.3 percent go to customers.

Under the formula, earnings up to a pre-defined level are retained by the utility, while earnings in excess are shared with customers.

Sharing an earnings deficiency works the same way, but deals with earnings below the allowed ROE. For example, for an earnings deficiency between Target 1 and Target 3, shareholders absorb all of the earnings deficiency; deficient earnings from Target 3 less an additional 0.5 percent share the deficiency 50:50 with customers and shareholders; and the deficiency in earnings from Target 3 minus 0.5 percent down to 1 percent are absorbed 40 percent by shareholders, and 60 percent by customers. Additional earnings deficiencies below Target 3, less 1.5 percent, are all made up by customers. For example, between 10 percent to 9.2 percent, all of the earnings deficiency is absorbed by shareholders; the portion of the earnings deficiency from 9.2 percent to 8.7 percent is made up 50:50 between customers and shareholders; the earnings deficiency between 8.7 percent to 7.7 percent is made up 60 percent by customers, 40 percent by shareholders; and any earnings deficiencies below 7.7 percent are all made up by customers.

The process depends on an annual audit and review by commission staff. Hearings or settlement talks would be triggered if there are intervenors. The filing requirements would include an earnings monitoring report using book amounts as adjusted based on the results of the utility's most recently completed general rate case and subsequent settlements resulting from the annual audits. The filing should include: 1) schedules for regulated income showing revenue and expenses by FERC Uniform System of Accounts; 2) a cost-of-service study to allocate expenses and rate base; and 3) cost of capital calculated on an end-of-year basis. To simplify the refunds to, or

recovery from, customers, revenue adjustments might be allocated to each applicable rate class based on that class' base revenue as a percentage of total base revenues eligible for refund or recovery of deficiency from customers if a symmetrical plan is used.

The revenue side of ratemaking by formula provides for recovery or refund of incremental costs and revenue requirements for a limited number of items. Examples include additions to capacity; significant changes in sales both positive and negative; purchased power capacity costs; and the loss of any revenue from interruptible rate programs. Thus, the process acts like a fuel adjustment clause in regard to revenue in that other cost changes are ignored as long as earnings Target 2 is not exceeded. The key word is incremental in that the process is sort of a mini-rate case for new additions (or deletions) in the revenue requirements associated with certain predefined cost elements, such as:

- Purchased power capacity: the revenue requirement associated with the capacity costs portion of purchased-power agreements and interruptible contracts if excluded from base rates;
- Lost revenue (if any) from conservation programs;
- Environmental costs: in the event the company incurs new environmental costs resulting from the imposition of regulations that previously haven't been imposed;
- New major construction projects: recover the retail revenue requirement with major construction projects not fully recovered through current base rates (commission preapproval for the project is necessary);
- Storm reserve: in the event the company's approved storm reserve reaches a balance below a pre-specified level, the company should be allowed to make an adjustment necessary to increase the storm reserve to the commission-approved level (*i.e.*, the increase must be grossed up for applicable taxes);
- Exceptional changes: it's recognized that from time to time a utility might experience increases or decreases in costs, or decreases in revenues, that occur as a result of actions, events, or circumstances beyond the control of the company; and
- Loss of customers or customer load: a utility might also, from time to time, experience significant reductions in its base revenue due to loss of customer load targeted to some level of ROE reduction (*i.e.*, 0.5 percent).

The process is sort of a mini-rate case for new additions (or deletions) in revenue requirements.

Avoiding Pitfalls

Ratemaking by formula can present certain challenges for utilities and regulators.

Most notably, it might mitigate utility efficiency. The stimulus of regulatory lag is lost and the utility might be lulled into a cost-plus situation. The regulator must carefully review the annual cost-of-service study required as part of the utility's annual filing to try to identify imprudent expenditures—especially an increase in expenditures as the higher sharing bands are reached. Performance incentives (*e.g.*, for reliability and customer satisfaction) tied to financial rewards and penalties also would be helpful. In addition, the regulator periodically should hold hearings that deal only with cost allocation and rate design, since over time with ratemaking by formula there is likely to be a mismatch between costs and allocation or rate structure. The cost of equity capital also could get out of whack with capital market conditions and should be the subject of separate hearings.

Tasks for the regulator, among others, include: establishing and monitoring operating parameters; conducting the annual review process; monitoring rate-base level; monitoring rate of return; evaluating any incentive results; pre-approving major capital expenditures; reviewing cost of service each year; pre-approving purchased power and natural gas contracts; establishing and monitoring commodity purchase incentives; and pre-approving major sales adjustments. Thus the regulators' focus is proactive, devoted to pre-approvals and incentives. F

Did you receive SPARK last month?

If not, visit www.pur.com to see what you missed. Complete the order form to ensure this month's issue arrives via e-mail.

SPARK is an electronic newsletter exclusively available to Fortnightly subscribers.

Doctrine for Cybersecurity*

Deirdre K. Mulligan† Fred B. Schneider‡

May 15, 2011

Abstract

A succession of doctrines have been advocated in the past for enhancing cybersecurity: prevention, risk management, and deterrence through accountability. None has proved effective, and their failings are discussed. Proposals are now being made to view cybersecurity as a public good or to adopt mechanisms inspired by those used for public health. This landscape is surveyed through the lens that a new doctrine—public cybersecurity—provides.

1 Introduction

Governments, business, and individuals are growing increasingly concerned about the security of networked computing systems. This concern is justified. Press reports of successful attacks grow ever more frequent: cross-site scripting used to pilfer consumers' passwords, large-scale breaches of corporate customers' personal information, massive distributed denial of service attacks on web sites, cyber espionage aimed at classified documents, and all manner of attacks on civil critical infrastructures.

Computer Scientists and those who fund them are, consequently, investing heavily in technological means for improving cybersecurity. But technological solutions are useless if they are not deployed or if operating

*Supported in part by AFOSR grant F9550-06-0019, National Science Foundation grants 0430161, 0964409, CNS-0524745 (ACCURATE), and CCF-0424422 (TRUST), ONR grants N00014-01-1-0968 and N00014-09-1-0652, and a grant from Microsoft. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of these organizations or the U.S. Government.

†School of Information, University of California, Berkeley 102 South Hall, Berkeley, CA 94720–4600.

‡Computer Science Department, Cornell University, Ithaca, New York 14852.

practices allow them to be circumvented by attackers. Policy must create incentives for system developers, operators, and users to act in ways that enhance rather than weaken system security. Moreover, neither technologists nor policy makers have the luxury of starting with a clean slate. All must labor in the shadows of legacy networks and end-systems that are not secure (nor easily made so) and in the context of extant policy that reflects societal values from a time when dependence on networked information systems was minimal.

Enhanced levels of cybersecurity can entail tensions involving cost, function, convenience, and societal values such as openness, privacy, freedom of expression, and innovation. Absent some widely accepted doctrine, evaluation of proposals for improvement is difficult and debate about their adoption can be neither compelling nor conclusive. The utility of a doctrine is thus determined by the extent to which it offers a framework for resolving these tensions but does not impose, ignore, or rule out possible technical or policy solutions.

We thus conclude that a prerequisite for achieving enhanced cybersecurity is articulating a *cybersecurity doctrine*, which specifies *goals* and *means*.

- **Goals.** These define some agreed upon kinds and levels of cybersecurity, characterizing who is to be secured, at what costs (monetary, technical, convenience, and societal values), and against what kinds of threats. The goals might be absolute or they might specify a range of permissible trade-offs. In allowing trade-offs, we acknowledge the political nature of cybersecurity and the need for conversations among those affected when goals are set.

- **Means.** These might involve technological, educational, and/or regulatory *measures*. We should expect a means to include policy that creates incentives—which might range from market-based to coercive—that foster adoption and/or deployment of the measures it proposes.

Through incentives provided as part of its means, a cybersecurity doctrine can address barriers to market production of cybersecurity that others[1] have aptly noted reflect a lack of will rather than a lack of ability. Incentives also can prompt continued improvement to address the constantly emerging landscape of threats and the new needs that arise as a growing range of applications are being migrated to networked information systems.

[1] R. Anderson and T. Moore. The Economics of Information Security, *Science 314* (5799), 610-613, October 27, 2006.

One candidate for such a cybersecurity doctrine is the subject of this paper. Our doctrine of *public cybersecurity* is rooted in the thesis that cybersecurity is a public good. The focus of this doctrine, then, is on the collective rather than on any single individual's or entity's computer, network, or assets. Comparisons with public health—another public good—are apt; these are made below.

The next section (§2) analyzes the limitations of various cybersecurity doctrines that have been proposed to date. Then §3 discusses in detail our new doctrine of public cybersecurity. The production of cybersecurity by building systems that have fewer vulnerabilities is the subject of §4. Subsequent sections discuss approaches for managing insecurity: diversity (§5), surveillance (§6), installation of patches (§7), isolation (§8), and the role of intermediaries (§9). Finally, §10 puts this work into perspective and §11 offers some conclusions.

2 Cybersecurity Doctrines: Past and Present

The advent of time-sharing in the 1960's meant that computations on behalf of multiple users were interleaved on a single computer. Each user's computation and data thus had to be protected from misbehavior by programs being run by other users. Confronted by a problem born of technology, engineers of early time-sharing systems sought solutions in technology. Therefore, the focus of early cybersecurity doctrine was on developing new technology. Societal values could be and were ignored, because users of these early computing systems had shared values which meant those values would be respected without explicit discussion.

Technological solutions for creating the needed isolation were beyond our capabilities, especially when users could be motivated and capable adversaries bent on disrupting another user's computation or stealing information. Improved technology, however, is not the only way to solve problems that technology has created, and subsequent cybersecurity doctrines focused on policy to leverage those technological solutions that were at hand. These doctrines too were unsuccessful. But had they succeeded, they ultimately would have been inadequate because the problem was changing.

Computer systems were becoming pervasive, which had two broad consequences. First, the information technology sector became a significant economic force. Concerns about freedom to innovate and success in the marketplace come with this first consequence. The second consequence was that computer systems increasingly touched the lives of ordinary people, as

citizens whose records were stored electronically, and as workers who used information technology to be more efficient. Eventually, computer networks and the web were created; they changed how people shopped, communicated, socialized and engaged in politics. As a consequence, privacy and other societal values grew in importance. The requirement for a cybersecurity doctrine to take account of societal values became crucial.

Because their effectiveness has been limited, it is instructive to review the three doctrines that have dominated cybersecurity thinking for the past 50 years—the doctrines of prevention, risk management, and deterrence through accountability. In particular, by analyzing the measures these doctrines propose, we might hope to gain insights into properties that cause measures to be effective.

Doctrine of Prevention. The goal of this doctrine is systems that are completely free of vulnerabilities. Absent vulnerabilities, attacks are not possible, so the resulting system is secure.

Such *absolute cybersecurity* is worthwhile but unlikely ever to be achieved. For systems that incorporate humans as users and operators, we would need some way to prevent social engineering and intentional insider-malfeasance. Prevention, here, requires overcoming the frailty of humans, which is likely to involve more than technology.

If we ignore humans in a system then the problem is different but not easier. Software systems today are just too large and complicated to be verified using formal logics. Researchers, assisted by computers, have been able to devise formal proofs for small systems[2] (i.e., under 10,000 lines of code), and software producers regularly employ automated checking for relatively simple properties of code and for analyzing specifications. Techniques are also being explored that reduce the size of code bases for certain key systems, because a smaller code base is more amenable to formal verification.[3] But revolutionary advances are needed before formal verification could be used to verify the entire code base found today on a desktop system or server, yet that would be necessary for implementing the Doctrine of Prevention.

[2]Klien et al. seL4: Formal Verification of an OS Kernel. *Proceedings of the ACM SIGOPS 22nd Symposium on Operating Systems Principles*, SOSP '09, (Big Sky, Montana, USA), ACM, 207–220.

[3]For example, "prerendering" can reduce the code required to generate the user interface in a voting system, thereby simplifying the vote-entry software and making it more amenable to verification. Yee, Ka-Ping. *Building Reliable Voting Machine Software.* Ph.D. Dissertation, Computer Science Department, University of California at Berkeley, Fall 2007.

System testing is the obvious alternative for gaining assurance that a system has no vulnerabilities. Tests, however, only can reveal the presence of vulnerabilities—not their absence. Exhaustive testing would be necessary for demonstrating the absence of vulnerabilities, and the amount of work involved here is prohibitive even for small components, much less for large systems.

Formal proofs and testing are necessarily relative to some expectations about what the system must do and the environments in which the system will operate. So we must discharge the Doctrine of Prevention relative to some expectations, which is tantamount to establishing the absence of vulnerabilities only if certain assumptions hold. Unfortunately, assumptions about the environment that hold one day might subsequently be invalidated. Attacks evolve in sophistication in response to better defenses. And threats emerge to exploit new opportunities for disruption that are created when cyberspace provides access to new forms of value. So a system that was once considered secure might not remain so for very long

In light of this dynamic, expectations about the environment must be periodically revisited and, if necessary, revised. Thus the Doctrine of Prevention involves a recurring expense. That recurring expense is inconsistent with the business model employed by many of today's software providers, which favors reuse and extension of existing hardware and software in order to lower the cost of producing new systems.

The adoption of mandatory standards can be seen as a way to support the Doctrine of Prevention by increasing the chances that what is built and/or deployed has fewer vulnerabilities. Some standards directly concern what functions an artifact must or must not support, some govern its internal structure, while others concern the process by which the artifact is constructed or maintained, and yet others govern qualifications of the personnel who are involved in creating the artifact. Examples include the TCSEC[4] (also known as the Orange Book), its successor the Common Cri-

[4]DoD Computer Security Center. *Department of Defense Trusted Computer System Evaluation Criteria*, CSC-STD-001-83, August 1983.

teria[5] as well as security provisions in information privacy laws[6] [7] [8], the Federal Information Security Management Act[9], and the Voluntary Voting System Guidelines[10]. Current market activity suggests that such mandates show value in some areas. However, a correlation between the absence of vulnerabilities and standards compliance has not yet been documented. So the stated goal for the Doctrine of Prevention is unlikely to be achieved through these measures.

Doctrine of Risk Management. Absolute cybersecurity is not affordable but, fortunately, it is also not needed for most systems. The Doctrine of Risk Management responds by stipulating a more modest goal—that investments in security reduce expected losses from attacks. To adopt this doctrine is to admit that all vulnerabilities are not equal; one should focus only on vulnerabilities whose exploitation is (i) sufficiently likely by the perceived threats and (ii) could enable system compromises that are expensive (by some cost measure). In contrast to the Doctrine of Prevention, the more modest goal of defending a smaller body of code against a more restricted set of threats is likely within our capabilities. Moreover, to maintain that steady state, fewer assumptions about the environment would have to be revisited periodically.

In theory, the Doctrine of Risk Management seems quite sensible. But the lack of information about vulnerabilities, incidents, and attendant losses makes actual risk calculations difficult. Companies and individuals do not know how to value (i) confidentiality of information, (ii) integrity of information, or (iii) the pain of dealing with recovery from an attack's effects (e.g., bad credit ratings). And these costs seem to vary tremendously. Also,

[5] US National Institute of Standards. *Common Criteria for IT Security Evaluation*, 1999. ISO Standard 15408. http://csrc.nist.gov/cc/.

[6] Health Insurance Portability and Accountability Act of 1996 (HIPAA), Pub. L. No.104-191, 110 Stat. 1936 (1996) (regulating the use and disclosure of "Protected Health Information").

[7] Title V of Gramm-Leach-Bliley Act (GLBA), Pub. L. No. 106-102, 113 Stat. 1338 (1999) (codified at 15 U.S.C. §§6801-6827 (2006)), 15 U.S.C. §§6801, §§6805 (empowering various agencies to promulgate data-security regulations for financial institutions).

[8] Children's Online Privacy Protection Act, 15 U.S.C. §6501 et seq. (prohibiting the collection of personally identifiable information from young children without their parents' consent).

[9] Federal Information Security Management Act of 2002 ("FISMA", 44 U.S.C. §3541, et seq.).

[10] Election Assistance Commission. *Draft Voluntary Voting System Guidelines*, Version 1.1, May 27, 2009. Available at www.eac.gov/testing_and_certification/voluntary_voting_system_guidelines.aspx

people have difficulty reasoning about extremely low-probability events. Finally, investment incentives are missing when costs are borne by third parties, materialize only well after the breach occurs, and causation is difficult to discern much less prove.

Accurate information about threats and attacks may not be publicly available because those with that knowledge fear tarnishing their reputations or compromising their intelligence methods and/or sources. Even were that information available, deployment of replacement systems and upgrades changes the uses to which systems are being put and alters the set of vulnerabilities, which in turn can lead to the creation of new attacks. These differences mean that the past is not a good predictor of the future. As a consequence, actuarial models cannot be constructed, so insurance to transfer risk is impossible to price in a way that ensures profits to the policy underwriter.[11]

Were there some way to analyze a system mechanically and obtain a quantity that indicates just how secure that system is, then we could have a basis for assessing what is gained from specific investments made in support of cybersecurity. But today such cybersecurity metrics do not exist. Quantities derived entirely from empirical observations also don't work for justifying investments. The absence of detected system compromises could indicate that investments in defenses worked, attacks haven't been attempted, or the compromise escaped notice (e.g., theft of confidential information). So whether or not prior security investments were well targeted is impossible to know, leaving security professionals to justify investments based solely on non-events.

Approaches based on risk management are further confounded by externalities that derive from the emergent nature of cybersecurity in networks. Individuals and entities here can neither fully reap the benefit of their security investment nor entirely control their vulnerability through investments.[12] For example, a single compromised system anywhere in a network can serve as a launching point for attacks on other systems connected to that network. So local investment in defenses not only provides local benefits but also benefits others; and under-investment in defenses elsewhere in the network could facilitate local harms. Absent coordination, the only

[11] Rainer Boehme and Galina Schwartz. Modeling Cyber-Insurance: Towards A Unifying Framework. Working paper presented at Workshop on Economics of Information Security, Harvard University, June 2010. http://weis2010.econinfosec.org/papers/session5/weis2010_boehme.pdf

[12] R. Anderson and T. Moore. The Economics of Information Security, *Science 314* (5799), 610-613, October 27, 2006.

logical strategy would be to invest in insurance (were it available), because here an entity can reap the entire benefit of that investment.[13] However, this strategy does nothing to improve security and, as noted above, viable long-term business models for insurance do not exist today.

The situation for risk management is not completely bleak, though. In the policy arena, state security breach notification laws[14] can be viewed as a risk management intervention. Significant costs are incurred to notify individuals and to manage the adverse publicity surrounding reportable breaches. These potential costs act as a proxy for the costs of security failures to customers, forcing companies to internalize previously externalized risks of security failures. The price tag on breaches also means that these laws have created a set of data to use in risk and return on investment calculations. However, the laws focus on only a narrow set of breaches and, as a result, might artificially skew the focus of investments.

Doctrine of Deterrence Through Accountability. This doctrine is concerned with treating attacks as crimes, and it therefore focuses on infrastructure to perform forensics, identify perpetrators, and prosecute them. We deter attacks by increasing the chances that perpetrators will be found and prosecuted.[15] Implementations of this doctrine require strong authentication technologies and surveillance of network activity. Robust forms of user identity allow us to overcome the loose binding that exists today between individuals and machines.[16]

Absent an effective means for retribution, this doctrine has no teeth and fails. And efforts to punish perpetrators of cyber-attacks are not always feasible in today's global environment. Attribution of actions by machines to individuals is complicated, agreement about illegality is illusive, and cross-

[13] J. Grossklags, N. Christin, J. Chuang. A game-theoretic analysis of information security games. *Proceedings of the 17th International World Wide Web Conference* WWW 2008, (Beijing, China, April 2008).

[14] Security breach notification statutes that require companies to notify individuals when certain personal data has been accessed or disclosed without authorization are in place in 46 states, the District of Columbia, Puerto Rico and the Virgin Islands. The first was California's Notice of Security Breach Law Cal.Civil Code §1798.29 (2002).

[15] Butler W. Lampson. Computer security in the real world. *IEEE Computer* 37(6), June 2004, 37–46.

[16] South Korea currently requires internet users to attach their real names and resident ID numbers when they post messages on the Internet. To accomplish this, Web sites that allow posting must collect and confirm names and resident IDs with a government server. Se Jung Park, Yon Soo Lim, Steven Sams, Sang Me Nam, and Han Woo Park, "Networked politics on Cyworld: The text and sentiment of Korean political profiles," *Social Science Computer Review*, September 21, 2010.

border enforcement requires more cooperation than is likely to emerge between important constellations of nations. Recent attacks against U.S. and other systems suggest that we cannot ignore non-nation state actors that engage in terrorism and large-scale financial crimes. The very features that make the internet a profitable environment for criminals—world-wide reach, connectedness, neutral treatment of packets, and weak binding of machines to individuals—make it difficult for law enforcement to identify and catch them, while other features of the international landscape make it difficult to bring them to justice.[17]

There are also conceptual obstacles that limit the effectiveness of the doctrine. First, this doctrine is punitive. Like most criminal law, it is aimed primarily at punishing with expectations of producing both general and specific deterrence. This does little to keep networks up and running when they are under siege nor does it prompt proactive security investments. Second, the doctrine could require individuals to sacrifice privacy and, in the extreme, abandon the possibility of anonymity and the protections for freedom of speech and association it affords.

Nevertheless, many attacks are indeed criminals plying their trade, and it makes good sense that criminal activity in cyberspace face the risk of retribution that we employ to deter criminal activity in the physical world. The Doctrine of Deterrence Through Accountability thus brings value here. But in cyberspace, unlike in the physical world, terrorists or state actors are difficult to distinguish from common criminals.[18] Deterrence through accountability is not necessarily effective against these trans-national threats. Other doctrine is also required.

3 A New Doctrine: Public Cybersecurity

Cybersecurity is non-rivalrous and non-excludable so, by definition, it is a *public good*. It is non-rivalrous because one user benefiting from the security of a networked system does not diminish the ability of any other user to benefit from the security of that system. And it is non-excludable because users of a secure system cannot be easily excluded from benefits security brings. Doctrines to foster the production of public goods thus constitute a sensible starting point in our search for doctrines that promote cybersecurity.

[17] J. Franklin, V. Paxson, A. Perrig, and S. Savage. An inquiry into the nature and causes of the wealth of internet miscreants. *Proceedings of the 14th ACM Conference on Computer and Communications Security* (CCS '07). ACM, New York, NY, USA, 375-388.

[18] William J. Lynn III. Defending a new domain. The Pentagon's Cyberstrategy. *Foreign Affairs*, September/October 2010.

Economists define a *common good* as one that is rivalrous and non-excludable. The sea, outer space, and air are examples. In so far as common goods (by definition) are inherently different from public goods, doctrines for common goods are likely to be unsuitable for fostering cybersecurity. Indeed, we can see this irrelevance in laws for protecting common goods, which typically aim to ensure rights of equal use, and in the mechanisms these laws introduce, which are intended to manage the depletion and inequitable consumption by first comers or more sophisticated users. The production of cybersecurity has little to do with any of these issues.

3.1 Public Health as an Exemplar

Public health—the prevention of disease and promotion of good health in populations writ large—is a public good. It is non-rivalrous because having the population healthy implies a lower prevalence of disease, which in turn decreases the chances any member can fall ill. And it is non-excludable because nobody can limit an individual's ability to profit from the health benefits that living amongst a healthy population brings.

The essential characteristics of public health law are a focus on the health of the population as a whole and the singular responsibility of government in that enterprise.[19] To discharge these responsibilities, various agencies are mandated to engage in a broad set of activities, which include the following.

- Public education about the cause and effects of disease, as well as methods of prevention, since that education empowers individuals to act in ways that optimize their own health, which in turn furthers public health.

- Creation and use of methods for the prevention and treatment of specific diseases. This could involve (i) providing subsidies to procure care needed by those who could not otherwise afford it[20] or (ii) imposing specific health standards as eligibility requirements for various societal

[19]The mission of public health was defined by an influential Institute of Medicine Committee as "fulfilling society's interest in assuring conditions in which people can be healthy". *The Future of Public Health*, Institute of Medicine, 1988 p. 7. A rich description of the legal framework is set out in Lawrence O. Gostin, *Public Health Law: Power, Duty, Restraint*, University of California Press, January 2001.

[20]For example, after several outbreaks of measles among primarily unvaccinated children, a federal law was passed to provide free vaccines to certain groups of children and to provide funds to states to support efforts to enhance vaccination levels.

benefits.[21]

- Identification and management of disease and infected individuals though surveillance, information gathering, and analysis—including mandatory reporting requirements for certain diseases and conditions, mandatory testing or screening for others, symptom surveillance that seeks to identify non-obvious public health threats in masses of routine records, and mandatory treatment.

While interests of an individual and the public often align, it is the points of conflict that public health law speaks to—by setting out the rights of individuals as sovereigns over their physical bodies versus the obligation of the state to protect the population as a whole and by offering frameworks to mediate conflicts between the two.

As an example, public health mandates that children be vaccinated because in a generally healthy population such vaccinations cannot be justified based on the benefit to the individual. In fact, the optimal choice for any given child might be to avoid vaccination and thus avoid the risk of side-effects. What mandatory vaccination creates is *herd immunity*, which benefits the collective by reducing the total number of hosts available to carry a disease, thereby decreasing the risk to individuals who have not been vaccinated. However, if too many individuals act in their own self-interest and eschew vaccinations, then the herd immunity that allowed a non-vaccinator to freeload may disappear. This is a "tragedy of the commons" where individuals acting rationally leave everyone worse off.

Every state in the U.S. conditions a child's attendance at school on satisfying some specified regimen of vaccinations. And, in addition, vaccine manufacturers are indemnified from liability for side-effects users might experience. Specifically, the Vaccine Injury Compensation Program (VICP) provides certain payouts from public coffers to children injured as a result of a vaccine. VICP also provides a compensation mechanism outside seeking damages from the vaccine manufacturers, thereby providing an environment conducive to both the production and willing use of vaccines. While this cannot fully compensate for negative health consequences from vaccinations, it is an important component of the overall public health strategy.

Public health is a logical outgrowth of having disease detection and prevention mechanisms, which transformed societal perception of health from a primarily private concern to a concern of the collective. Ultimately, this

[21] As discussed later, children in the United States are eligible to attend public school only after receiving a mandated set of vaccinations.

led to public health being seen as a public good that the government should enable.[22] With a focus on the collective, health now becomes intertwined with societal values. For example, we impinge on societal values when we introduce mandatory reporting and surveillance systems that (i) alert individuals at specific risk so they can be tested and treated, and (ii) allow isolation, quarantine, and even mandatory treatment to be imposed. At the same time, public health interventions aim to minimize their intrusiveness due to the chilling effect they may have on access to healthcare. Thus, we pursue programs such as anonymous HIV testing and needle exchange.

This same public health framework (*viz*, laws, agencies, and measures) applies equally well to weaponized pathogens. This is not to suggest that motive is irrelevant in considering public health strategies. For example, weaponized pathogens may change more quickly than those that evolve in nature, and certainly the transmission vectors may differ when pathogens are used as weapons. But the basic tools of public health still apply: public education (to minimize exposure and facilitate early detection), investments to create means for prevention and treatment (antidotes and vaccines), and surveillance and analysis (facilitating isolation and quarantine as defenses).

3.2 From Public Health to Public Cybersecurity

Public health and cybersecurity both aim to achieve a positive state (health or security) in a loosely affiliated but highly interdependent network. With one, it is a network comprised primarily of people existing in an environment over which they have some limited control; with the other, the network is comprised of people, software, and hardware (for communications, storage, and processing). And because this positive state is ultimately unachievable, both struggle with how to manage in its absence as well as with how to prompt its production. Success ultimately depends not only on technical progress but on reaching a political agreement about (i) the relative value of some public good in comparison to other societal values and (ii) the institutions granted authority to resolve conflicts (and the methods they use).

We define a *doctrine of public cybersecurity* to be any cybersecurity doctrine whose goals are (i) producing cybersecurity and (ii) managing insecurity[23] that remains, where political agreement balances individual rights and

[22] *The Future of Public Health*, Institute of Medicine, 1988 p. 3.

[23] Systems that employ technical means to enable continued operation in the face of attacks are sometimes called *intrusion tolerant*. A sampling of specific techniques for achieving intrusion tolerance is discussed in *Foundations of Intrusion Tolerant Systems*

public welfare. There is no single doctrine of public cybersecurity. Different definitions for what is meant by "cybersecurity" and "insecurity" as goals[24] lead to different doctrines of public cybersecurity. Also, different choices of measures and incentives that together constitute the means result in different doctrines of public cybersecurity. But none of the doctrines discussed in §2 has all of the elements we require in a doctrine of public cybersecurity.[25]

The analogy with public health inspires cybersecurity measures like prevention, containment, mitigation, and recovery—strategies that direct resources toward production and preservation of cybersecurity. But modern public health doctrine does not compensate victims of disease, so by analogy, a doctrine of public cybersecurity would not focus on restitution. Indeed, restitution is economically efficient only when attacks are infrequent, and that assumption is not realistic today.

Modern public health also does not punish victims of disease, though there is some nuance. Quarantine, in response to disease, benefits the collective by depriving an individual of certain freedoms. Such a response could be seen as harsh consequences, which is one definition of "punishment". By analogy, a doctrine of public cybersecurity could dictate responses that deprive individuals of actions, but only if those responses benefit the collective. Punishments solely for retribution could not be part of a public cybersecurity doctrine (since retribution does not benefit public welfare), though nothing precludes implementing a doctrine of public cybersecurity alongside a cybersecurity doctrine that does admit retribution. Finally, the parallel with public health also suggests that prevention be preferred to recovery.

With regard to incentives, ensuring that actors contribute to public cybersecurity requires interventions to overcome positive and negative externalities that lead rational individuals to under-invest, just like for public health. And when incentives are insufficient to prompt private provisioning, the public interest requires making value-ridden choices to interfere with the rights and interests of individuals and organizations. Those choices would be embodied in goals that reflect political agreement about the good in

(Jaynarayan Lala, editor), IEEE Computer Society, Order Number PR02057, ISBN 0-7695-2057-X, 2003. Uses of the term "managing insecurity" in this paper are intended to denote something broader, admitting non-technical means as well as intrusion tolerance techniques.

[24]H. Nissenbaum, Where Computer Security Meets National Security. *Ethics and Information Technology*, Vol. 7, No. 2, June 2005, 61-73.

[25]Doctrine of Prevention is not concerned with managing insecurity; Doctrine of Risk Management and Doctrine of Deterrence Through Accountability are not concerned with producing cybersecurity. And none concern trade-offs of individual rights for public welfare.

question (its definition), the socially desirable level given competing priorities and values, and provisions for determining when the individual's desires yield to the collectives' need. For example, an agreement might stipulate that state coercion is permitted only when certain incursions into the rights and interests of individuals are tightly circumscribed.

Public health solutions don't always translate into sensible support for public cybersecurity, but we have been pleasantly surprised at how often the former inspires strategies for the latter. Below, we explore some examples. These also illustrate how doctrines of public cybersecurity can be useful for evaluating current cybersecurity proposals. Our choice of examples should not, however, be seen as an endorsement for any particular proposed set of interventions.

4 Producing Security during Development

Underutilized approaches (e.g., formal methods, testing, and improved software engineering processes and standards) that were developed, in part, to serve the Doctrine of Prevention are effective in producing cybersecurity (by reducing the number of vulnerabilities present in a system), even if they cannot produce absolute cybersecurity. Thus, methods do exist that could serve as a means for a doctrine of public cybersecurity, just like disease prevention through vaccination and the monitoring of our food and water supplies fosters public health. The question is: What incentive structures would ensure that these methods are used?

Education could play a key role in defect reduction. Knowledgeable developers are less likely to build systems that have vulnerabilities. They are also better able, thus more likely, to embrace leading-edge preventions and mitigations. There is, however, no agreement about what should be taught. To reach that agreement would require a dialog among universities and with practitioners.

Once agreement has been reached about a body of knowledge for cybersecurity practitioners, we could require that mastery of this material be certified as a condition for practice. But the details of how certification is handled can be subtle. Possession of a certificate does not by itself compel the use of best practices, and it is easy to imagine certified system-builders who cut corners by choice (out of laziness, for example) or by mandate (because management is trying to reduce costs). Moreover, unless the certification process imposes a continuing education requirement to ensure that certificate holders will stay current with new developments, it might impede

rather than promote the spread of innovation. And even when continuing education is mandated, old habits die hard; physicians, for example, who have been shown new methods, empirically demonstrated to be superior, nevertheless exhibit a tendency to stick with familiar practices.[26]

Employing practices to reduce defects during system development and employing better-educated practitioners will mean systems become more expensive to produce. Today's software-procurement market does not provide incentives that would prompt developers to incur those additional expenses. Moreover, purchasers are unable to predict the costs of a system's vulnerability to attack and, without ways to measure a system's security, have no way to rationalize paying higher prices—the same underlying reasons that the Doctrine of Risk Management failed.

Law could force system producers and/or purchasers to make the necessary investments. Software distributors currently disclaim liability beyond the purchase price for damages caused by their product. This probably reduces the time and energy that developers devote to eliminating defects, as evidenced by the number of buffer overruns and other exploitable coding errors still being discovered and exploited by attackers. The law could, for example, be revised to disallow limits on damages flowing from attacks that exploit poor coding practices that lead to buffer overflows and other easily exploited vulnerabilities. Limits on liability could depend on the use of formal methods, type safe languages, or specific forms of testing (e.g., fuzz testing[27]). Creation of a class of certified security professionals also could provide the basis for a professional duty of care supporting liability for shoddy security.

Law also could require that software developers adhere to security standards. Or safe harbor provisions could be created to protect software developers against future findings of liability for those systems built according to specified standards. In fact, the law arguably already mandates that companies follow certain standards regarding personally identifiable information. Through a set of settlement agreements, the Federal Trade Commission established a de facto standard that requires a company collecting and handling personal information of consumers (i) to establish reasonable security processes and (ii) to mitigate system vulnerabilities that are known in the marketplace and for which mitigations exist. A first step in deter-

[26] Deborah G. Mayo and Racehelle D Hollander. *Acceptable Evidence: Science and Values in Risk Management*. Oxford University Press, 1994

[27] In *fuzz testing*, a system is exposed to random inputs of unexpected kinds. This form of testing reveals inadequacies in the input validation routines of a system. Several classes of attacks are blocked by implementing stringent input validation.

mining whether law should more broadly mandate the adoption of security standards might be research that identifies connections between security development processes and good security outcomes.[28]

5 Managing Insecurity: Diversity

Monocultures in nature risk extinction from pathogens and have less chance of adapting to changing conditions. Diversity—of the individuals within each species and by virtue of many species co-existing within an ecosystem—is one way nature creates a resilient ecosystem. Public health benefits from individuals in a population having different inherent resistance to pathogens and, by virtue of the different exposures[29] to diseases that each experiences, having different immunities.

Although nature abhors monocultures, cyberspace seems to favor them. A collection of identical computing platforms is easier, hence cheaper, to manage because we need master only one interface and make one set of configuration decisions. In addition, user training costs are reduced when job transfers do not have the overhead of learning yet another operating system and suite of applications; investments in education about how to use or manage a system also now can be amortized over a larger user base in a monoculture. Finally, interoperability of a few different kinds of systems is far easier to orchestrate than integrating a diverse collection, standards not withstanding. So networking is usually easier to support with a monoculture. Mindful of these advantages, the public and private sectors both tend to adopt procurement policies that foster creating computer monocultures.[30]

[28] For a case-study comparison of four vulnerability reduction techniques see K. Buyens, B. De Win, W. Joosen: Empirical and statistical analysis of risk analysis-driven techniques for threat management. ARES 2007: 1034-1041, *Proceedings First International Workshop on Secure Software Engineering* (SecSE 2007). For a theoretical comparison of two high-profile development processes (Microsoft SDL and the Open Web Application Security Project's Comprehensive, Lightweight Application Security Process (CLASP) see Johan Gregoire, Koen Buyens, Bart De Win, Riccardo Scandariato, and Wouter Joosen: On the secure software engineering process: CLASP and SDL compared. In *SESS'07: Proceedings of the 3rd International Workshop on Software Engineering for Secure Systems*, Minneapolis, MN, USA, May 2007.

[29] Vaccination works by causing exposure to a relatively benign form of the disease against which protection is being sought.

[30] A notable example is U.S. Office of Management and Budget policy memorandum M-07-11 "Implementation of Commonly Accepted Security Configurations for Windows Operating Systems" which lists the permitted versions of Windows that should be used by certain civilian federal agencies. See www.cio.gov/documents/FDCC_memo.pdf.

Methods do exist, however, to artificially and automatically create diversity in software systems—and do so without sacrificing the advantages given above for having a monoculture. These methods involve tools that randomly transform code and/or stored information in semantics-preserving ways. Due to the *artificial diversity*, internal details of an individual system are no longer predictable. So an attack that exploits knowledge of internal details is more likely to cause a system crash after a small number of instructions than to cause an attacker to receive control. In many settings, having a system crash is preferable to it being controlled by attackers. Moreover, a platform that crashes in response to an attack cannot then help propagate that attack to other platforms. The crash also (implicitly) signals system operators that something is wrong, creating an opportunity for initiating other means to block an attack's spread.

As with the diversity found in nature, artificial diversity is inherently a probabilistic defense, because an attack against any given individual component might not be derailed by the specific random transformations that were made to that component. Also, by converting some attacks into crashes, artificial diversity can adversely affect a system's availability.

Despite these limitations, artificial diversity does facilitate public cybersecurity because it provides a way to cope with residual vulnerabilities, thereby providing a way to manage insecurity. Today, artificial diversity is often used in operating systems but less so in applications[31] (even though increasingly it is applications that attackers target). However, the various legal approaches discussed above in §4 for incentivizing defect reduction during development are equally well suited for incentivizing system producers to support artificial diversity. So a rich space of incentives is at hand for encouraging broader adoption of the measure.

6 Managing Insecurity: Surveillance

Surveillance is used extensively in support of public health, where data is collected through a variety of means. The data enables disease containment and mitigation through

- the dissemination of information that facilitates individual actions to avoid exposure to infection,

[31] D. Ladd et al. The SDL Progress Report. Microsoft Corporation, 2001. page 23–24. Available at http://www.microsoft.com/downloads/en/details.aspx?FamilyID=918179A7-61C9-487A-A2E2-8DA73FB9EADE&displaylang=en.

- isolation and quarantine, which limit the interaction of affected individuals with the rest of the population, and
- mandatory treatment to reduce danger to the public.[32]

Data collection for public health occurs at many levels. At the lowest level is the self-serving inclination of individuals themselves to assess their well-being. Education equips individuals with a basic level of knowledge about health indicators—normal body temperature, pulse, blood pressure, and respiratory rate—as well as with simple precautions to limit infection and the spread of disease (e.g., frequent hand washing). Other data is collected by primary care providers in conjunction with annual check-ups and, when symptoms require further analysis, at hospitals and other more-advanced diagnostic facilities. Each successively higher level is concerned with the overall health for a larger population and, thus, provides a natural venue for constructing and analyzing larger data aggregations.

By minimizing the disclosure of information about an individual's health, public health law strives to reduce one potential source of reluctance about seeking health care—fear an individual might have about being shunned because of a publicized health condition. Generally, identifying information flows away from primary health care providers only in instances where aggregation and/or analysis is necessary to identify significant trends. And when information does flow, efforts are undertaken to protect each individual's privacy.

Surveillance in Networked Systems. In contrast to public health, coordinated surveillance is not used extensively today in support of cybersecurity. Yet it would be feasible and advantageous to do so. Low-level indicators about the basic "health" of a computer can be made available by running built-in checking software (e.g., virus scanners and intrusion detection systems). And each of the Internet Service Providers (ISPs) that constitutes the network has an infrastructure that facilitates monitoring of events internal to its network as well as interactions with other networks.

Surveillance of network traffic (volume, distribution over time and destinations, etc.) could be a powerful potential source of information about certain attacks and vulnerabilities. Denial of service attacks, for example, have an obvious manifestation and a natural mitigation based on traffic filtering by ISPs. However, the source(s) of such attack packets, the target(s),

[32]The general rule allows individuals to refuse treatments. Some states mandate treatments for communicable diseases, such as tuberculosis, that pose a danger to the public.

and the intermediaries are likely to span multiple ISPs, which would have to share data and coordinate for mitigation. Data sharing among ISPs is, unfortunately, inhibited today by competition and, in some cases, varied interpretations of privacy law.[33] So ISPs are less able to have situational awareness that could enable them to suppress packets delivering attacks. Widespread sharing of information, however, can introduce a risk if chances are now increased that attackers learn about vulnerabilities for specific sites.

Just as there are privacy issues with collecting data about an individual's health, network traffic surveillance raises privacy concerns. The extent to which collecting packets actually does impinge on privacy depends on what information is recorded, how long it is stored, how it is used, who can have access to the information, etc. For example, realtime responses to protect networks can be accomplished by authenticating machines, a far less politically fraught solution than proposals for "internet drivers licenses" and other tight bindings between machines and individuals.[34]

ISP cooperation and information sharing is less likely to raise privacy concerns than collection of information by centralized government organizations. Yet the packets themselves can be invaluable to a government seeking situational awareness about threats in cyberspace, and defense of its citizens is a clear responsibility of government. Unfortunately, packet inspection is also easily abused by a government for spying on citizens; critics cite this fear (among others) when discussing the Einstein[35] systems recently deployed by the U.S. Government for monitoring all federal civilian agency connections to the internet. As with public health, political agreement must be reached based on the expected benefits (backed by sound research and field experience) of surveillance and the risk it poses to other values.

An understanding of the kinds vulnerabilities found in systems is a form of situational awareness of potentially great value to system builders. In the absence of mandatory reporting requirements for cybersecurity incidents, a diverse collection of public and private reporting mechanisms have evolved. The U.S. Department of Homeland Security (US-CERT) and the NIST Computer Security Division maintain databases of common vulnerabilities.

[33] M. Van Eeten and J. M. Bauer. Economics of malware: security decisions, incentives and externalities. OECD STI Working Paper 2008/1. OECD. Available at www.oecd.org/dataoecd/53/17/40722462.pdf.

[34] See, D. D. Clark and S. Landau. Untangling Attribution. *National Security Journal* Vol. 2, Issue 2 (2011). http://harvardnsj.com/2011/03/untangling-attribution-2/.

[35] Center for Democracy and Technology, Einstein Intrusion Detection System: Questions That Should Be Addressed, July 28, 2009. http://cdt.org/security/20090728_einstein_rpt.pdf.

Many organizations contribute, but these are not the only such databases and none provides anything close to a complete view. Some vulnerabilities do not get reported to any. For example, a private sector community of "security researchers" search systems for vulnerabilities but report their findings to middlemen, who offer it for sale to companies that build and sell anti-malware or intrusion prevention/detection products.[36] The ad hoc Conficker Working Group[37] is an example of a rather successful coordinated private-sector activity formed to defend against a particular attack.

7 Managing Insecurity: Patching

A *patch* is an update that can be applied to an installed system in order to eliminate one or more previously identified vulnerabilities. Exploitation of an unpatched vulnerability on a computer could target that machine and assets of the individual and, therefore, be fully internalized. Or the exploitation could target the machines and systems of others, producing a negative externality.[38] This uncertainty about consequences means that self-interest of machine owners is not a strong incentive to apply patches.

Various policy interventions could raise patch rates, though. Choosing among them requires additional information about why people and businesses delay or outright fail to apply patches.

- Research might conclude that low patch rates in the consumer market are caused by an under-appreciation of the risks. Public education that applying patches improves cybersecurity might then dramatically increase patching.

- We might find that individuals lack awareness of vulnerabilities present on their machines. Here, built-in software to check whether all current patches have been applied might suffice for triggering consumers to be more attentive to downloading patches for their machines.

- Feedback about what others do, thereby creating new norms of behavior, might lead to better patching practices. Researchers in other areas

[36] The risk that such vulnerabilities might be sold or disclosed to irresponsible or hostile parties is cause for concern about this kind of market, though the structure flourishes in the absence of alternatives.

[37] See www.confickerworkinggroup.org.

[38] For example, unpatched machines can be co-opted by attackers into a botnet. Such collections of remotely-controlled machines are used today for a variety of illicit activities, including generation of spam email and distributed denial of service (DDoS) attacks.

have found that showing individuals how their behavior compares to others taps into a competitive and/or social consciousness. So simple statements that a great percentage of others have patched their machines and are doing their part for cybersecurity just might be enough to push the laggards into applying patches.

- Research might find that individuals or enterprises hesitate to install patches for fear of destabilizing their other software. Reluctance by individuals might be overcome if software vendors were more transparent about what specific configurations and applications they tested. And enterprises that depend on home-grown software could have have their fears somewhat assuaged if test suites could be provided for use by patch developers.[39] As a final safety-net, we might require that all software contain mechanisms whereby a patch that has been applied can easily be removed, and the system and data restored to the pre-patch state.

- If the impediment to installing security patches is the time or expertise needed then an obvious mitigation is to have vendors configure defaults that automatically download and apply security patches.

- Consumers being charged for Internet access in proportion to the amount of bandwidth they use will incur lower costs by not downloading patches, which could be an excuse to forgo installing security patches. One solution, here, is to subsidize the bandwidth required for such downloads; tariffs that distinguish between different kinds of traffic is another solution.

- Those who run pirated software might hesitate to install patches for fear that the installation process would disable the illegal software or detect and report it. Regulations could address this by prohibiting security-patch installation from implementing functionality in support of license-enforcement or any other form of intellectual property protection.

Incentives to apply patches could also have a useful indirect effect. If patch installations are frequent and disruptive then consumers have reason to prefer products that have fewer security vulnerabilities from the start. This, in turn, pressures software producers to build and deploy more secure products.

[39] However, some enterprises regard their software and data as proprietary. They might not be comfortable providing test suites to patch developers.

Mandates to apply patches raise concerns about subsidizing the installation of patches[40] and compensating injured parties when patches cause harm. Losses from applying mandated patches, particularly where unacknowledged and uncompensated, will breed suspicion and resistance to patching efforts. Thus, it seems advisable to consider back-stop measures, analogous to what VICP provides to incentivize the use and production of vaccines and to the process used at the Food and Drug Administration to ensure vaccine efficacy.

8 Managing Insecurity: Isolation

Geological features, like mountains and oceans, have proved valuable in protecting individuals and populations. And when natural boundaries are absent, we build our own—fences to surround buildings and nations, often with guards to control who is allowed to transit the border. Such boundaries protect activities on one side from activities occurring on the other. A boundary might limit travel in one direction or in both directions, and it might be completely impervious or it might be selective about what may pass. Neither is a panacea. An impervious boundary could block the good with the bad; a selective boundary must employ some kind of *filter*, and that filter might block things it shouldn't or pass things it should.

Firewalls, so-called network guards, intrusion detection/prevention systems, and "air gaps" are examples of mechanisms that implement boundaries in networked systems.[41] Data collected through surveillance can serve as the basis for *signatures*, which are then used to define filters, effectively creating dynamic boundaries. Surveillance thus can lead to automatically-imposed quarantines. And since attacks in networks propagate rapidly, automatic response is especially attractive.

Ideally, we would deploy selective boundaries that block attacks but nothing else. In practice, though, filters will be far from perfect.

- Filters that inspect packet payloads (known as *deep packet inspection*) in addition to checking packet headers are ineffective when packet payloads are encrypted or otherwise obfuscated. Encryption is not

[40] Richard Clayton. Might governments clean up malware? *Proceedings Ninth Annual Workshop on Economics and Information Security* WEIS10, (Cambridge, MA, June 2010).

[41] The term "air gap" originally referred to isolation caused when no wires are connected to a given component. With the advent of wireless networking, the term's meaning has broadened to denote isolation caused when the Laws of Physics ensure no signal can reach the component.

today used extensively in networks, but that could easily change. And encryption is often used by attackers to evade detection. Moreover, it is malware variants that often are being spread, where each variant is obfuscated by applying a different random set of semantics-preserving transformations. It is quite difficult and often impossible to construct a signature that matches all variants by generalizing from a few.

- A filter might be designed either to (i) block packets and protocols corresponding to known attacks or (ii) pass packets from protocols or conveying content that is known to be normal. Filters that implement (i) are fooled by new attacks (in addition to suffering limitations described in the previous bullet). Filters that implement (ii) could block previously unseen protocols and kinds of packets, which stifles innovation.

- Whether a packet is part of an attack could depend only on sender intent. Consider a large number of request packets being sent to a web server. Are many people trying to access the same particularly topical content or is a denial of service attack is in progress? Sender intent is the sole differentiator.

There is also a human element to consider. Boundaries and filters must be installed, configured, and managed by human operators, and people do make mistakes. Moreover, when such a mistake allows unimpeded flow, then the error might be difficult to detect until after it is too late.

Network providers are understandably reluctant to publicize details of defenses, because revealing that information could help attackers. Yet we do see example defenses in today's commercial networks, which create and reference so-called "black lists" of sites whose communications will be ignored and "white lists" of sites that are known to be trustworthy. Some ISPs create a competitive advantage by offering a service to their customers whereby suspicious inbound-traffic spikes directed at that customer's site will automatically prompt upstream filtering to block those suspicious packets. Denial of service attacks in such networks are harder to undertake, as a result. Other ISPs monitor each endpoint, disconnecting it if outgoing traffic suggests that endpoint is compromised.[42]

A boundary might be deployed around a system (be it a single computer or a network) that must be protected from attacks or it might be deployed

[42]OECD. Malicious software (malware): A security threat to the internet economy. Ministerial Background Report DSTI/ICCP/REG(2007)5/FINAL, June 2008. Available at http://www.oecd.org/dataoecd/53/34/40724457.pdf.

around a system that is likely to harbor attackers. Different incentives are effective in each case. One natural scenario for direct government investment occurs when security boundaries coincide with national ones. Systems in different countries are subject to different laws, typically reflecting different societal values. A government might therefore justify installing a boundary whenever systems subject to its laws are being connected to systems located in a jurisdiction that allows system behavior the first considers an attack.

Boundaries are more likely to accepted and effective when instigated by the collective rather than by individuals. First, an individual is unlikely to have the necessary authority for mandating changes to defenses on all of the remote systems that could be involved in creating a quarantine. Second, the possibility of freeloading means externalities limit the incentives that might prompt owners or operators of networks or individual systems to make the investments to support enforced isolation. Finally, an agent of the collective would have the fuller picture that enables better signatures to be defined for filters.

We see an example of such boundaries in recent proposals for deterrence through accountability. Some have suggested that the internet be partitioned into national or multi-national enclaves. Those enclaves that serve the population we are trying to protect (i) run protocols that enable packet-sender tracing and (ii) do not carry traffic from enclaves where packet-sender tracing is not supported or cannot be trusted. Accountability of attack packets back to an individual machine can now be supported in enclaves that serve populations of concern.

With powerful enough filters, boundaries have the potential to intrude on societal values. One concern arises when the defining filters not only block packets that contain attacks but can be configured to block other kinds of packets. Such a filter could be used to prevent data from leaving an enclave, which makes it well suited for protecting confidential information against theft. But content filters also could allow governments to implement censorship, as illustrated by the firewalls China has installed to protect that nation's computing systems from receiving information in violation of local laws about allowed speech. Deployed in the reverse direction, a content filter could block someone from sharing information with others, thereby stifling debate.

So there may be trade-offs: societal values as well as potential benefits for the collective versus constraints on activities by individuals and business. Moreover, no criteria for deciding where a system should be segregated will be infallible. The result is a complex risk-management decision procedure that society must prescribe, with imperfect information and unknowable

consequences.

9 Managing Insecurity: ISPs as Intermediaries

The public health system leverages health professionals and other institutions to influence individuals' behavior. For example, health professionals educate individuals about benefits of vaccinations, schools demand conformance with vaccination schedules, and airports screen passengers for symptoms during some infectious disease outbreaks. Intermediaries clearly play an important role in public health strategies.

Intermediaries also have an important role to play in fostering cybersecurity. For example, in many cases today, network operators, such as employers and universities, require that all machines on their networks run virus detectors or malware detectors (with up-to-date signature files). These intermediaries could require that all machines are up-to-date on security patches. Similarly, some ISPs have chosen to notify subscribers when a computer seems to be infected.[43] At least one ISP restricts web surfing until the machine is cleaned-up, while another ISP reportedly quarantines any compromised machine until it is clean.[44]

ISPs are well positioned to facilitate patching and, by monitoring traffic, to enforce isolation of machines harbouring certain malware. Yet ISPs today have little incentive to engage in such behaviors, because in doing so they incur the bulk of the costs while any costs from the infected machines is more widely dispersed. Moreover, an ISP that disables or limits a machine's access to the internet will likely bear the burden of assisting that customer as she attempts the necessary repairs. Analysis[45] suggest that the cost incurred by an ISP in fielding a customer's tech-support call approaches the ISP's annual revenue from that customer. So by making this sort of monitoring and clean-up a mandatory obligation for ISPs, we would not

[43] Elinor Mills, "Comcast pop-ups alert customers to PC infections" Cnet News, October 8, 2009 (discussing Comcast October 2009 announcement of a trial of an automated service that will "warn broadband customers of possible virus infections" called Comcast Constant Guard the pop-up notice of possible infection directs customers to resources to rid their machine of infection, a similar service by Qwest, Customer internet Protection Program that displayed a Web page warning to customers provided ways to remove the infection for free before allowing them to surf the Web; and a similar older, now discontinued service of SBC that quarantined computers until they were cleaned)

[44] Id

[45] Richard Clayton. Might governments clean up malware? *Proceedings Ninth Annual Workshop on Economics and Information Security* WEIS10, (Cambridge, MA, June 2010), page 5, footnote 2.

only force action by the ISPs but we would also prevent consumers from shopping around for ISPs with weaker security requirements.

More daunting are the potential costs an ISP might incur from making an incorrect decision to disconnect a customer.[46] To limit spam email, for example, an ISP might block all bulk sending of email. But missives sent by a political organization might then be blocked, resulting in unwanted attention from advocacy groups and the press.[47] And while the law is evolving to provide ISPs who take steps designed to protect security with immunity from suits brought by providers of malware, those users who suffer losses after installing required patches or system upgrades or suffer due to isolation might also file legal complaints. In summary, the costs of ISP intervention present a formidable barrier to such action, but the law could remove these disincentives.

A recent proposal[48] by Lichtman and Posner argues that expanding ISPs liability "for violations of cyber-security" would improve cybersecurity because (i) individual attackers are often either beyond the reach of the law or are judgment-proof and (ii) because ISPs "can detect, deter, or otherwise influence the bad acts in question." The proposers speculate that ISPs could detect cybersecurity violations by building profiles of their users and looking for traffic anomalies, in a manner analogous to the monitoring of cardholder purchases done by credit card companies. But anomaly detection with usable levels of fidelity has eluded cybersecurity researchers for decades (as Lichtman and Posner openly admit), so implementing the proposal is not at present feasible.

Still, a policy holding ISPs liable for the damage caused by infected machines running on their networks might encourage more diligence in monitoring and fixing their subscriber's machines. The details are subtle and depend on the standard for liability—strict, knowledge based, etc. One (undesirable) outcome would be if the ISP undertakes less monitoring as a way to avoid its duty to intervene.

Alternatively, we could provide indirect or direct subsidies to foster cybersecurity-preserving activities by ISPs. For example, the government

[46]For a discussion of some of the issues raised by an ISP decision to cut off broadband access due to infection see, George Ou, "Comcast heading the right direction on cybersecurity" Digital Society Web site, October 9, 2009 http://www.digitalsociety.org/2009/10/comcast-heading-th-right-direction-on-cybersecurity/.

[47]http://www.eff.org/wp/noncommercial-email-lists-collateral-damage-fight-against-spam.

[48]See generally Doug Lichtman and Eric Posner, Holding Internet Service Providers Accountable, 14 S. Ct. Econ. Rev. 221, 233-34 (2006).

could create a centralized service for hosting patches or it could subsidize bandwidth to all endpoints in order to ensure that cost or delay to download a patch cannot become an impediment to installing that patch.

Given the decentralized and private provisioning of network resources in this country and many others, figuring out the role of intermediaries in driving cybersecurity is essential. As in other areas, such as copyright, the challenge is to establish policies that incentivize the desirable behavior while minimizing impact on other values.

10 Discussion

A biological basis for cybersecurity has been discussed by computer scientists for at least two decades. Their thrust has been to understand whether computer networks can benefit from implementing defenses like those that have worked so well for protecting living things. So we see intrusion detection systems that mimic pathogen detection by immune systems[49] and software defenses based on artificial diversity.[50] A recent white paper from DHS[51] describes how a human immune system's response mechanisms might serve as the blueprint for software that defends individual computers and networks against cyber-attacks; much research remains to be done, though, before those ideas are reduced to running code.

In contrast to this biological metaphor, which focuses on technical measures for blocking cyberattacks, the analogy between public health and cybersecurity is primarily concerned with new policy and new institutions. The notion of a Cyber-CDC, for example, has been attracting considerable interest.[52] Inspired by the existing Centers for Disease Control and Prevention, the proposed Cyber-CDC would be a government institution that provides leadership in organizing public and private sector strategies to enhance cybersecurity. It would also undertake data collection about threats and attacks, analyze and disseminate that information (perhaps in partner-

[49] S. Forrest, A.S. Perelson, L. Allen, and R Cherukuri. Self-nonself discrimination in a computer. *Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy* (Oakland, CA, May 1994).

[50] S. Forrest, A. Somayaji, and D. Ackley. Building diverse computer systems. *Proceedings of the Sixth Workshop on Hot Topics in Operating Systems* (Cape Cod, Mass, 1997).

[51] Department of Homeland Security. Enabling distributed security in cyberspace. http://www.dhs.gov/xlibrary/assets/nppd-cyber-ecosystem-white-paper-03-23-2011.pdf.

[52] National Cyber Leap Year Summit 2009, Co-chairs Report. http://www.nitrd.gov/NCLYSummit.aspx

ship with the private sector), serve as a repository for technical remedies, and educate the public about best practices, defenses, and remedies.

An IBM white paper[53] broadens the analogy. It proposes not only borrowing from public health but also from public safety. The broader analogy leads to additionally recommending that a Cyber Federal Emergency Management Agency be established and that a Cyber National Response Framework be devised. Independently, Microsoft's Scott Charney has advocated "device health" measurement and using a device "health certificate" as a basis for controlling access by devices to network resources.[54] Hunker, instead of focusing on institutions, suggests that public health could be a model for norms of behavior.[55] Individuals are expected to satisfy those norms; government institutions should focus on supporting the norms.

None of the aforementioned work includes a compelling argument for why the analogy to public health is a suitable starting point for cybersecurity. Public health informs people's behaviors (seemingly an obvious route to enhanced cybersecurity) but so does religion (which is not being advocated as a cybersecurity solution). We argued the analogy between public health and cybersecurity by using economic theory to highlight their shared status as public goods, since economics provides a vocabulary for talking about externalities and incentives. So while viewing cybersecurity as a public good is not new[56] [57] [58] we seem to be the first to use economics insights to justify the analogy between public health and cybersecurity.

Our public cybersecurity doctrine also goes beyond prior work that explores cybersecurity counterparts for institutions and policies that have served public health well. Public cybersecurity derives from identifying cybersecurity counterparts to the *goals* of public health—not the *institutions* of public health. First, through public health law, we gain and have

[53] Daniel B. Preito and Steven Bucci. Meeting the cybersecurity challenge: Empowering stakeholders and ensuring coordination. White Paper, IBM U.S. Federal Division, Feb 2010.

[54] Scott Charney. Collective Defense. Applying Public Health Models to the Internet. White paper available at http://www.microsoft.com/security/internethealth

[55] Jeffrey Hunker. U.S. International policy for cybersecurity: Five issues that won't go away. *Journal of National Security Law & Policy* 4, 197–216.

[56] Benjamin Powell. Is cybersecurity a public good? Evidence from the financial services industry. Working Paper Number 57, The Independent Institute, March 15, 2001.

[57] Bruce H. Kobayashi. An economic analysis of the private and social costs of the provision of cybersecurity and other public security goals. Working paper 26, George Mason University School of Law, 2005. Available at http://law.bepress.com/gmulwps/gmule/art26.

[58] Brent Rowe and Michale Gallagher. Private sector cyber security investment strategies: An empirical analysis. RTI International, March 2006.

exploited a powerful framework for balancing collective versus individual interests. Second, just as managing disease is an important goal of public health, managing insecurity is an important goal of public cybersecurity. We thus advocate that the siren call for the production of "secure" systems and networks be augmented with a mandate to manage the inevitable insecurity that comes from the constant vulnerabilities and adversaries we face.

The goals of public cybersecurity focus on the collective. Individual high-consequence systems, such as those that control critical infrastructures, are not singled out. Why not instead just focus our efforts on a seemingly smaller problem—making only the high-consequence systems secure? In public health terms, why not focus only on keeping important [sic] people healthy? Isolation is not a realistic proposition for either. Public health teaches that it is easier to keep specific individuals healthy when everyone is healthy. And so it will be with cybersecurity. If we foster the production of cybersecurity generally and our networks are able to manage insecurity then it will be easier to ensure that our high-consequence systems are secure.

Cybersecurity, like security in so many other contexts, involves trade-offs with other values.[59] So conflicts will have to be resolved between public cybersecurity and other values or interests of specific individuals, entities, and society at large. And a cybersecurity doctrine is obliged to provide principles and processes to negotiate and resolve these. Public health already offers principles to resolve such conflicts for a public good.

First, an individual's decision is most drastically prescribed by the state where that decision might directly impact the health of others, and the state generally lacks the ability to coerce, when the health impact is primarily on the individual. Change "health" to "security" and we obtain sensible guidelines for public cybersecurity.

Second, public health guidance applied to how public cybersecurity should deal with externalities suggests the following.

- The state's obligations and abilities to shape and override private choices should turn on the extent to which they have a direct impact on the security of the public broadly rather than the security of an individual or entity.

- Where their decision may be poor due to a lack of information or a misapprehension of the risks, the state's role should be aimed at facilitating better individual decisions through information or gentle

[59] H. Nissenbaum, Where Computer Security Meets National Security. *Ethics and Information Technology*, Vol. 7, No. 2, June 2005, 61-73.

interventions that influence the perception of risk, not supplanting them.

- Where security choices of the individual will impact the security of others, then the state can use a wider array of tools to alter behavior.
- Even where state action is permissible, impact on other societal values must be considered in choosing among solutions.
- Whenever possible, opt for minimal interventions implemented in a decentralized manner, so as to limit the negative impact they may have on willingness to participate.

11 Conclusions

Cybersecurity is the obstacle to success of the information age. Though the problem resides in technologies, the solution requires policies. It requires intervention in the private choices of individuals, hard trade-offs, and political agreements that could span nations.

We believe that a doctrine of public cybersecurity can be the basis for those policies. The doctrine establishes a framework for state incentives and coercion that we believe is rational, defensible, and legitimate. It alters the focus of cybersecurity to be on the collective rather than on the individual; and it advocates that systems be built with fewer vulnerabilities but acknowledges that systems will have vulnerabilities and must still be resilient in the face of attacks. If adopted, public cybersecurity will reorient public policy and discourse toward the proper goals of encouraging collective action to produce the public good of cybersecurity and managing the insecurity that remains.

Acknowledgments. We benefited from comments on early drafts of this paper and discussions with Marjory Blumenthal, Aaron Burstein, Scott Charney, John Chuang, Dave Clark, Craig Fields, Kelly Garrett, Jens Grossklags, Joshua Gruenspecht, Joseph Lorenzo Hall, Carl Landwehr, Susan Landau, Steve Lipner, Helen Nissenbaum, Shari Pfleeger, Audrey Plonk, Ashkan Soltani, participants at the 2009 Workshop on the Economics of Securing the Information Infrastructure, and attendees at several Daedalus meetings.

Harvard Journal of Law & Technology
Volume 17, Number 1 Fall 2003

MODULARITY, VERTICAL INTEGRATION, AND OPEN ACCESS POLICIES: TOWARDS A CONVERGENCE OF ANTITRUST AND REGULATION IN THE INTERNET AGE*

*Joseph Farrell*** & *Philip J. Weiser****

TABLE OF CONTENTS

I. INTRODUCTION 86

II. OPEN ARCHITECTURE AND MODULARITY 90
A. The Creation of the Internet and Its End-to-End Architecture 90
B. The Transformation of the Computer Industry 92
C. The Development of Competition in Telecommunications 93
D. Modularity and the Logic for Open Access Regulation 95

III. INTEGRATION AND EFFICIENCIES: PUTTING THE MODULARITY MOVEMENT ON ICE 97
A. Integrative Efficiencies 97
B. ICE and the Rationale Against Open Access Regulation 100

IV. HOLES IN THE ICE: WHEN ITS LOGIC CAN FAIL 105
A. Baxter's Law 105
B. Price Discrimination 107
C. Potential Competition 109
D. Bargaining Problems 112
E. Incompetent Incumbents 114
F. Option Value 117
G. Regulatory Strategy Considerations 118

---

* This work stems from the University of Colorado's Silicon Flatirons Telecommunications Program's first two policy conferences, "Telecommunications Law for the 21st Century" and "The Regulation of Information Platforms," which spurred the discussion that led to this collaboration. The authors thank Tim Bresnahan, Julie Cohen, Dale Hatfield, Marty Katz, Mark Lemley, Doug Melamed, Jon Nuechterlein, Robert Pitofsky, Steven Salop, Marius Schwartz, Jim Speta, Steve Williams, and Chris Yoo for their helpful comments and encouragement.

** Professor of Economics and Chair of the Competition Policy Center, University of California at Berkeley. Professor Farrell served as Deputy Assistant Attorney General and Chief Economist at the Antitrust Division of the U.S. Department of Justice (2000–01) and as Chief Economist at the Federal Communications Commission (1996–97). He thanks SIEPR for financial support through the Cain Fellowship.

*** Associate Professor of Law and Telecommunications, and Executive Director of the Silicon Flatirons Telecommunications Program, University of Colorado. Professor Weiser served as Senior Counsel to the Assistant Attorney General in charge of the Antitrust Division of the U.S. Department of Justice (1996–98). He acknowledges the generosity of a Law and Public Affairs Fellowship at Princeton University (2001–02) and a summer research grant from the University of Colorado.

H. Incomplete Complementarity 119

V. LESSONS FROM ICE AND ITS EXCEPTIONS, AND TOWARDS A COGENT FRAMEWORK FOR OPEN ACCESS POLICIES 119
A. Regulatory Strategies to Facilitate Modularity 120
B. Considerations for Regulatory Policy 125
C. Regulatory Philosophies 126

VI. THE FRAMEWORK IN ACTION 127
A. Microsoft 128
B. The Computer Inquiries 129
C. Broadband Policy 133

VII. CONCLUSION 134

I. INTRODUCTION

Just as the dust settles from the Microsoft case, the Federal Communications Commission ("FCC" or "the Commission") is preparing to craft rules to regulate broadband networks.[1] Taken together, these developments may mark the beginning of a new model of regulation for the Internet age. This regulatory regime will govern when a firm must provide "open access" to its platform — be it an operating system, a telecommunications service, or some other technology that facilitates Internet content or services — and will significantly influence the future development of the Internet.

A critical challenge for this emerging model of regulation will be whether and how to integrate antitrust policy and telecommunications regulation into a coherent whole. Antitrust and regulation have starkly contrasting traditions on mandated access. As the Internet, computer software, and telecommunications ("New Economy"[2]) industries converge, affected firms will increasingly seek clear and consistent legal rules.[3] Moreover, courts reviewing the FCC's decisions in this area are increasingly pressuring the Commission to devise a regulatory regime more compatible with economic theory and antitrust policy.[4]

---

1. *See, e.g.*, Inquiry Concerning High-Speed Access to the Internet over Cable and Other Facilities, 17 F.C.C.R. 4798 (2002) (Declaratory Ruling and Notice of Proposed Rulemaking) [hereinafter High-Speed Declaratory Ruling].

2. *See* Richard A. Posner, *Antitrust in the New Economy*, 68 ANTITRUST L.J. 925, 925 (2001).

3. For a discussion of technological convergence and its impact on telecommunications regulation, see NAT'L RESEARCH COUNCIL, BROADBAND: BRINGING HOME THE BITS 9 (2002), which states that "with convergence, everything — video, audio, text, and so forth — has become a digital stream that can be transported across the Internet."

4. For an example of increasing judicial insistence on careful economic analysis by regulators, see United States Telecom Ass'n v. FCC, 290 F.3d 415, 422–28 (D.C. Cir. 2002), in which the court criticized the economic rationale behind the FCC's rules for unbundling the local telecommunications network. *See also* Warren G. Lavey, *Inconsistencies in Applications of Economics at the Federal Communications Commission*, 45 FED. COMM. L.J. 437,

To do so, however, the FCC must develop a framework for regulating what economists call "vertical relations": how a firm relates to other firms in adjacent markets and whether it integrates into those markets.

In broad-brush terms, antitrust policy viewed much vertical conduct as suspect until the 1970s. By the late 1970s, however, the Chicago School of economics influenced mainstream antitrust thinking by establishing that vertical integration (e.g., mergers) and many kinds of vertical contracts had efficiency benefits and were unlikely to harm competition.[5] While post-Chicago School scholarship of the 1980s and 1990s has weakened that view,[6] current antitrust doctrine still generally presumes that vertical agreements, vertical extension, and vertical mergers are unobjectionable unless a fact-intensive investigation shows otherwise.

By contrast, in similarly broad-brush terms, early telecommunications policy positively encouraged integration and close coordination into "one network" under the regulated AT&T monopoly. Starting in the 1970s, however, a series of FCC and court decisions adopted a policy of developing and protecting open interfaces. This open architecture philosophy held that powerful firms at one level should not be allowed to leverage that power into — or perhaps even participate in — adjacent competitive segments. Likewise, the United States government's early support for the Internet encouraged the development of an open architecture based on modular standards.[7]

These contrasting traditions of analyzing open access leave telecommunications policy unsettled. Technological convergence and emerging competition in telecommunications blur the lines between industries regulated primarily by antitrust (notably computing) and those subject to telecommunications law, and telecommunications regulators increasingly pledge fealty to antitrust approaches.[8] The

---

439–40 (1993) (calling for increased judicial efforts to promote consistency in the use of economic theory to justify regulation).

5. The landmark event in the rise of Chicago School thinking was the Supreme Court's decision in Continental T.V., Inc. v. GTE Sylvania, Inc., 433 U.S. 36, 48 n.15, 55–56 (1977), which cited heavily to Chicago School criticisms of the Court's earlier doctrine.

6. *See* David S. Evans & Michael Salinger, *Competition Thinking at the European Commission: Lessons from the Aborted GE/Honeywell Merger*, 10 GEO. MASON L. REV. 489, 512 & n.58 (2002) (discussing the impact of post-Chicago School scholarship).

7. As we explain in more detail below, "modularity" is a means of managing complexity. As one commentator defined the term, modularity involves "breaking up a complex system into discrete pieces — which can then communicate with one another only through standardized interfaces within a standardized architecture — [to] eliminate what would otherwise be an unmanageable spaghetti tangle of systemic interconnections." Richard N. Langlois, *Modularity in Technology and Organization*, 49 J. ECON. BEHAV. & ORG. 19, 19 (2002).

8. For two discussions of the impact of convergence on regulatory policy, see Philip J. Weiser, *The Imperative of Harmonization Between Antitrust and Regulation*, *in* TELECOMMUNICATIONS CONVERGENCE: IMPLICATIONS FOR THE INDUSTRY AND FOR THE PRACTICING LAWYER 73 (PLI Intellectual Property Course, Handbook Series No. G-698,

clash of traditions and of arguments on open access is particularly sharp in one of today's central telecommunications problems: the regulatory treatment of broadband transport and its close complements. Broadband transport, usually provided by cable modems or telephone digital subscriber lines ("DSL"), promises to transform the Internet by vastly speeding up downloads and by permitting high-bandwidth applications.[9] Some commentators — most notably Lawrence Lessig — have urged regulators to impose modularity on this market by requiring broadband transport providers to share their facilities with Internet service providers ("ISPs").[10] Others, echoing the Chicago School perspective, argue that the market will facilitate open access to the extent that open access is efficient.

The open access question is even more ubiquitous than it may first appear, as policymakers and commentators often use different terms to describe the issue. Antitrust commentators discuss the "primary" (or "bottleneck") market and the "secondary" (or "complementary") market. In telecommunications, participants talk of "conduits" and "content." This Article, adopting the terminology used in the computer industry, will discuss "platforms" (often "information platforms") and "applications."[11] The essence of the issue is the complementarity between applications and platforms, whether the application is an input to the platform, a buyer of the platform, or neither.[12]

2002) and Philip J. Weiser, *Law and Information Platforms*, 1 J. TELECOMMS. & HIGH TECH. L. 1 (2002) [hereinafter Weiser, *Information Platforms*].

9. Though the definition of "broadband" will evolve over time, the FCC's current dividing line is 200 kilobits per second — bandwidth sufficient "to change web pages as fast as one can flip through the pages of a book and to transmit full-motion video." Inquiry Concerning the Deployment of Advanced Telecommunications Capability, 14 F.C.C.R. 2398, ¶¶ 20, 22 (1999) (Report); *see id.* ¶ 25 (noting that the definition will evolve); *see also* Inquiry Concerning the Deployment of Advanced Telecommunications Capability, 17 F.C.C.R. 2844, ¶ 7 (2002) (Third Report) (adhering to the 200 kilobits per second dividing line). *But see* NATIONAL RESEARCH COUNCIL, *supra* note 3, at 78–80 (proposing alternative definition).

10. *See* LAWRENCE LESSIG, THE FUTURE OF IDEAS 147–67 (2001). This argument builds off a prior piece that addressed critics of mandated modularity. *See* Mark A. Lemley & Lawrence Lessig, *The End of End-to-End: Preserving the Architecture of the Internet in the Broadband Era*, 48 UCLA L. REV. 925 (2001) (engaging arguments made in Phil Weiser, *Paradigm Changes in Telecommunications Regulation,* 71 U. COLO. L. REV. 819, 831 (2000) and James B. Speta, *Handicapping the Race for the Last Mile?: A Critique of Open Access Rules for Broadband Platforms*, 17 YALE J. ON REG. 39, 77–90 (2000)).

11. For a further explanation of the information platform concept and how it can frame technology policy debates, see Weiser, *Information Platforms*, *supra* note 8, at 3–8. A notable example of an information platform from the computer industry is the Microsoft Windows operating system, which exposes Application Programming Interfaces ("APIs") that can be used by applications developers to "call" certain functions provided by the operating system. *See* United States v. Microsoft Corp., 253 F.3d 34, 53 (D.C. Cir. 2001).

12. In part for this reason, we define "applications" broadly, not distinguishing between software applications and hardware products (such as peripherals), both of which may connect to an underlying platform. Rather, we will use the term "applications" for all complementary products or services used in conjunction with a platform.

This Article aims to help regulators and commentators incorporate both Chicago School and post-Chicago School arguments in assessing whether regulation should mandate open access to information platforms. Much discussion on such questions focuses on the degree of competition among platforms. By contrast, the central analytical tool — not necessarily the victor — in our discussion is a Chicago School-style argument we call *internalizing complementary efficiencies* or "ICE." ICE claims that even a monopolist has incentives to provide access to its platform when it is efficient to do so, and to deny such access only when access is inefficient. ICE is often a persuasive argument, yet its logic admits several cogent exceptions. Unfortunately, regulators and commentators seldom do justice to the nuances of this principle: some ignore ICE, while others embrace it and underestimate its exceptions. Only by addressing both ICE and its exceptions can regulators make full use of economics in analyzing open access requirements.[13]

In its broadband proceedings, the FCC has an opportunity to embrace the insights of ICE and its exceptions in developing a framework to evaluate independent providers' claims for mandated access to a platform such as broadband transport.[14] Ideally, such a framework would harmonize telecommunications regulation with antitrust policy and guide regulation in related contexts, such as unbundling policy for local telecommunications networks.[15] The FCC could thus more accurately apply economic principles to information platforms and satisfy judicial demands for a better economic explanation of its regulatory policies.

This Article proceeds in five main parts. Part II recounts experiences of the Internet, computer, and telecommunications industries, illustrating the powerful benefits of modularity that inspire proponents of open access regulation. To explain the Chicago School skepticism of such regulation, Part III first discusses how close (i.e., other than arm's-length) vertical relationships can yield important efficiencies.

---

13. Christopher Yoo makes a similar observation in his exposition of a project related to ours. *See* Christopher S. Yoo, *Vertical Integration and Media Regulation in the New Economy*, 19 YALE J. ON REG. 171, 177 & n.19, 178 (2002) (describing the project's focus on cable television, broadcast, and broadband markets, but disclaiming any application to telecommunications markets).

14. Such a framework would provide more guidance than prior ad hoc FCC decisions in this area, which typically arose from merger reviews. *See* James B. Speta, *A Common Carrier Approach to Internet Interconnection*, 54 FED. COMM. L.J. 225, 226 (2002) ("[T]he only legal rules governing Internet interconnection are a limited number of company-specific conditions imposed in some merger reviews."); Philip J. Weiser, *Internet Governance, Standard Setting, and Self-Regulation*, 28 N. KY. L. REV. 822, 844 (2001) ("In terms of setting a precedent for future regulation of information platforms, the FCC's AOL/Time Warner Order failed to set forth a principled model of analysis . . . .").

15. *See* United States Telecom Ass'n v. FCC, 290 F.3d 415 (D.C. Cir. 2002) (remanding the development of the standard for unbundling the local telecommunications network back to the FCC).

Part III then explains the ICE principle: even monopoly platform providers have at least some incentive to operate in a modular fashion when it is efficient to do so, because they *internalize complementary efficiencies*. Part IV describes eight holes in the ICE logic: reasons why a monopoly platform provider might inefficiently close its platform. We do not see comparable reasons why such a monopoly might inefficiently open its platform. Part V outlines regulatory tools often used to facilitate open access, discusses factors that regulators should consider when contemplating open access policies, and offers three possible regulatory philosophies consistent with our discussion. Finally, Part VI applies the ICE framework to the FCC's Computer Inquiries, the *Microsoft* case, and the current broadband proceedings, illustrating how the subtleties of ICE and its exceptions, if not carefully addressed, can lead to policy instability. In conclusion, the Article urges the FCC to adopt a coherent model of platform regulation that takes account of ICE and permits a more harmonious convergence between antitrust and regulatory policy.

II. OPEN ARCHITECTURE AND MODULARITY

This Part focuses on the benefits of modularity. Sections A, B, and C explain how the Internet, computing, and telecommunications industries all came to be organized in a relatively modular fashion. Section D then discusses the benefits of modularity in general and the rationale for making it a guiding light for information policy.

A. The Creation of the Internet and Its End-to-End Architecture

The Internet's development was a triumph of United States technology policy. The Internet grew from the Defense Department's Advanced Research Projects Administration's ARPANET and later relied on support from the National Science Foundation. From its early days in the late 1960s until the early 1990s, the Internet remained a government project, relying on the academic and research community for its development.[16] By the time commercial entities developed Internet services and products in the 1990s, its basic architecture was already in place. This architecture reflects the Internet pioneers' conscious strategy that the platform should not anticipate what applications would rely on it, and that no central gatekeeper should decide which applications could be provided.

The Internet can be understood as being comprised of four layers: content, applications, logical, and physical.[17] At its center lies the

16. *See* JANET ABBATE, INVENTING THE INTERNET 54–65 (1999).

17. There are various ways to describe the layers of Internet architecture. Lawrence Lessig, for example, suggests a definition of the content layer that includes what others call the

III. INTEGRATION AND EFFICIENCIES: PUTTING THE MODULARITY MOVEMENT ON ICE

Perhaps partly recognizing the efficiency and competitive benefits of modularity, antitrust policy until the 1970s was wary of allowing dominant firms to integrate into adjacent markets and create closed relationships between complementary products.[47] Over the last twenty-five years, however, antitrust policy has accepted the Chicago School argument that close (even closed) vertical relationships can yield and be motivated by integrative efficiencies. Furthermore, economists' better understanding of how complements boost demand for the primary good has taught antitrust that powerful firms, recognizing the merits of a modular industry structure, will often institute modularity voluntarily. The question for regulators therefore is not whether modularity is good — it very often is — but whether modularity is likely to be good *even when* it will not emerge (or survive) spontaneously, as it often will when it is most valuable to consumers.

This Part explains the logic behind allowing firms (even monopolists) to decide whether or not to integrate vertically into — or, more broadly, depart from an arm's-length relationship with — complementary markets.[48] Section A outlines some important efficiency benefits that can stem from a vertical relationship closer than an arm's-length one (or, equivalently, inefficiencies of arm's-length relationships). Section B goes on to explain the powerful concept of *internalizing complementary efficiencies*. ICE suggests that even a platform monopolist often has incentives to make efficient choices about when to maintain modularity and when to get involved in an adjacent market.

A. Integrative Efficiencies

Palm, which introduced the first successful personal digital assistant, later decided to separate its operating system and software applications divisions from its hardware division.[49] It did not want to follow Apple, which failed to commit to an open licensing strategy for its operating system and subsequently lost its initially strong market

47. *See, e.g.*, United States v. Arnold Schwinn & Co., 388 U.S. 365, 379 (1967) (recognizing the per se illegality of certain types of vertical restraints).

48. For brevity we sometimes talk as if platform firms choose between full integration and an arm's-length modular relationship with a complement, but of course there is a spectrum of vertical relationships, including partial integration (e.g., joint ventures), tie-ins, partial equity investments, long-term contracts, and affiliate relationships.

49. *See* Pui-Wing Tam, *For Palm, Splitting in Two Isn't Seamless*, WALL ST. J., June 27, 2002, at B4.

share.[50] Palm presumably preferred to emulate Microsoft, which has benefited enormously from modularity in the PC market.

The Apple and Microsoft examples might make Palm's decision to vertically separate seem like a no-brainer, but it was not. By separating its operations vertically, Palm lost control of some important aspects of its product deployment. For instance, Palm's reliance on outsiders and an "inability to crack the whip on its far-flung programmers" contributes (according to some observers) to its "slow pace of innovation" in applications.[51] By contrast, Sega developed the operating system, equipment, and leading games (such as Sonic the Hedgehog) for its Sega Genesis system all in-house, presumably in order to control its product offerings and drive consumer demand for its system.[52]

Because the platform and the applications made for it are economically interdependent, an arm's-length relationship can involve contractual hold-up hazards (on both sides, though especially threatening to competitive applications providers).[53] A closer vertical relationship can be an efficient response to such hazards.[54]

An arm's-length relationship can also lead to what economists call "double marginalization." The classic formulation, offered by Augustin Cournot in 1838, is that separate complementary monopolies, each imposing a monopoly markup, wind up with a final product price that exceeds the overall monopoly price. As a result, both consumers and the producers are worse off than they would be if the two firms merged and charged a monopoly price for the two goods together.[55] More generally, this insight explains that firms providing complementary activities or products are in a mutual position of "vertical externality." When Microsoft, for example, improves its software or lowers its price, more consumers buy Intel's complementary microprocessor; similarly, when Intel improves its hardware or lowers

---

50. *See id.*

51. Erick Schonfeld & Ian Mount, *Beating Bill*, BUS. 2.0, June 2002, at 36, 39, *available at* http://www.business2.com/articles/mag/print/0,1643,40438,FF.html (last visited Oct. 29, 2003).

52. *See* ADAM M. BRANDENBURGER & BARRY J. NALEBUFF, CO-OPETITION 237–41 (1996). One possible explanation for these differing approaches is that the proprietary strategy is most effective in launching a new system, but, as Palm is discovering, it is difficult to determine when or whether integration has outlived its usefulness and when, if ever, to separate integrated divisions that once worked well together. *See* Tam, *supra* note 49.

53. *See* Yoo, *supra* note 13, at 262–64 (noting that vertical integration guards against free riding, hold-up problems, and other strategic behaviors by vital complementors).

54. More precisely, the hazards arise when fully effective modularity is not available, so that ex post haggling is likely. There may then be an intriguing positive feedback: when modularity works well, it is appealing and may be stable, but when it starts to break down, a platform supplier's best response may eventually be to integrate — perhaps killing off whatever imperfect modularity remains.

55. *See* AUGUSTIN COURNOT, RESEARCHES INTO THE MATHEMATICAL PRINCIPLES OF THE THEORY OF WEALTH 103 (Nathaniel T. Bacon trans., MacMillan 1927) (1838).

its price, demand for Microsoft's operating system rises. Thus, when complementors move closer to maximizing joint profits — whether through integration or through a closer contractual relationship than arm's-length pricing — it tends to encourage innovation and price-cutting.[56]

Innovation can require changing the platform/application interface, which can be a slow process if an industry relies on open standards and open interfaces. In such cases, hand-in-glove coordination between the platform sponsor and one or more complementors can accelerate innovation.[57] In particular, a new product that would require new interfaces may be most readily launched in a hand-in-glove, even integrated, fashion. Indeed, Palm first launched its product in an integrated manner before moving to modularity through its voluntary split. Moreover, such coordination can give a platform provider more scope for penetration pricing and other start-up tactics aimed to encourage efficient use and adoption of its platform,[58] particularly when the product is newly introduced and relatively unknown.[59]

Integration or hand-in-glove coordination also helps assure consumers that complementary products will work well, because the platform sponsor retains control over quality and interoperability. Antitrust law, even at the height of its hostility to vertical tie-ins, appreciated this point in a case involving the rollout of cable television and related equipment.[60]

Analyzing a firm's choice of vertical structures is a focus of the "new institutional economics" ("NIE"). Building on insights of Nobel Laureate Ronald Coase,[61] NIE "seeks to extend and enrich understanding of the microanalytic details of business behavior and the in-

56. For development of this point and some important refinements of it, see Joseph Farrell & Michael L. Katz, *Innovation, Rent Extraction, and Integration in Systems Markets*, 48 J. INDUS. ECON. 413 (2000). Of course, when competitors — in contrast to complementors — move closer to maximizing joint profits, the result can readily be anticompetitive.

57. For a development of this theme and a discussion of the virtues of proprietary platform competition, see Philip J. Weiser, *The Internet, Innovation, and Intellectual Property Policy*, 103 COLUM. L. REV. 534 (2003). For a further discussion of how developing stable interfaces can be too expensive and time consuming to merit the effort, see Langlois, *supra* note 7, at 23.

58. *See* Douglas Lichtman, *Property Rights in Emerging Platform Technologies*, 29 J. LEGAL STUD. 615, 616–17 (2000).

59. *See* JEFFREY H. ROHLFS, BANDWAGON EFFECTS IN HIGH-TECHNOLOGY INDUSTRIES 197 (2001).

60. *See* United States v. Jerrold Elec. Corp., 187 F. Supp. 545, 556–57 (E.D. Pa. 1960), *aff'd per curiam*, 365 U.S. 567 (1961) (allowing leeway for bundling in introducing a new product when reputation matters).

61. *See generally* Ronald H. Coase, *The Nature of the Firm*, 4 ECONOMICA 386 (1937); Ronald H. Coase, *The Problem of Social Cost*, 3 J.L. & ECON. 1 (1960). The subject is also known as "transactions cost economics."

dustry settings that shape firm conduct."[62] Usefully, if tautologically, NIE suggests that firms will vertically integrate or depart from arm's-length market dealing when such arm's-length dealing would be more costly.[63] Thus, firms will sometimes opt for modularity as a means of bringing maximum imagination and diversity to the problem of developing applications on a platform, and minimizing the need for complex coordination. Conversely, firms will sometimes opt for vertical integration in order to facilitate complex coordination and strengthen incentives for product development and deployment.[64]

B. ICE and the Rationale Against Open Access Regulation

In an ideal world, a firm could obtain the benefits of vertical integration while still employing some degree of modularity to spur independent innovation. In attempting such strategies, platform providers who integrate into applications development often take pains "not to compete with customers" so as to minimize any ill effects of integration on independent applications.[65] But because getting the best of both worlds in this way is hard, firms may give up and stick to their core business. In this respect, Palm's decision to divest its operating system can be seen as a step to reassure its licensees that it can be trusted as a steward of the standard, that it will not leverage its control of the platform into related markets, and that it will remain focused on serving the needs of independent developers — particularly now that Microsoft's rival handheld operating system is offered on a modular basis (i.e., without a hardware component).[66] Similarly, AT&T divested its equipment manufacturing arm, Lucent, perhaps to reassure

---

62. Timothy J. Muris, Chairman, Fed. Trade Comm'n, Remarks at George Mason University Law Review's Winter Antitrust Symposium (Jan. 15, 2003), *at* http://www.ftc.gov/speeches/muris/improveconfoundtio.htm (last visited Oct. 29, 2003).

63. For discussions of this point and citations to relevant literature, see Alan J. Messe, *Tying Meets the New Institutional Economics: Farewell to the Chimera of Forcing*, 146 U. PA. L. REV. 1, 50–66 (1997) and Gregory J. Werden, *The Law and Economics of the Essential Facility Doctrine*, 32 ST. LOUIS U. L.J. 433, 462–64 (1987).

64. Because it is seldom obvious which of these two strategies is superior, antitrust courts have waded carefully into the area of "technological tying," requiring plaintiffs to establish that any competitive harms outweigh the efficiencies produced by such integration. *See, e.g.*, United States v. Microsoft Corp., 253 F.3d 34, 92–94 (D.C. Cir. 2001); ILC Peripherals Leasing Corp. v. IBM, 458 F. Supp. 423, 443–44 (N.D. Cal. 1978), *aff'd sub nom.* Memorex Corp. v. IBM, 636 F.2d 1188 (9th Cir. 1980); Telex Corp. v. IBM*,* 367 F. Supp. 258, 347 (N.D. Okla. 1973), *rev'd on other grounds*, 510 F.2d 894 (10th Cir. 1975).

65. This phrasing is most natural when applications developers buy the platform product and then sell a combined product downstream. Nonetheless, the same issues arise whether this is the market structure, or whether the platform provider buys from the applications developers, or whether end users or intermediaries buy both products.

66. *See* Ian Fried & Dawn Kawamoto, *Two Palms Better Than One?* (Feb. 4, 2002), *at* http://news.com.com/2100-1040-828446.html (last visited Oct. 29, 2003).

equipment customers who competed with other parts of AT&T that Lucent would not favor the latter.[67]

If a monopoly platform provider chooses to stick to its core platform business, it would prefer that applications — the complements to its product — be cheaply, innovatively, and efficiently supplied. Thus, in choosing how to license interface information, certify complementors, and otherwise deal with developers, such a firm has a clear incentive to choose the pattern that will best provide it or its customers with applications. That is, a firm will *internalize complementary efficiencies* arising from applications created by others. Although antitrust law has not always appreciated it,[68] we call this point Obvious ICE.

Obvious ICE can be illustrated with a numerical example involving a platform monopolist in the game console market.[69] Assume that competition in the market for applications (video games) will yield a selection of applications such that each user of the platform values it at $100, while a monopoly in applications will yield platform valuations of only $70.[70] Under these conditions, if the platform provider were to monopolize the applications market, the platform's value to a buyer would fall by $30; consequently, the platform provider would have to either sell fewer platforms or lower its platform price by $30. In that way, the platform provider *internalizes* the *complementary efficiencies* (here $30) from a better performing applications market.[71]

---

67. *See* T. Randolph Beard et al., *Why Adco? Why Now? An Economic Exploration into the Future Structure for the "Last Mile" in Local Telecommunications Markets*, 54 FED. COMM. L.J. 421, 457 n.83 (2002).

68. Judge Posner makes this point sharply in discussing the antitrust rule governing minimum resale price maintenance. *See* RICHARD A. POSNER, ANTITRUST LAW 177–78 (2d ed. 2001).

69. In antitrust, a company need not control 100% of a market (and even "market" is a nuanced term of art) to be considered a "monopolist"; rather, a "monopolist" is a company with considerable control over prices and output (and/or the ability to exclude competitors). *See id.* at 195–96 (noting that courts use market shares of 50% to 70% as threshold indicators of when a firm is a monopolist); United States v. DuPont & Co., 351 U.S. 377, 391 (1956) (defining "monopoly power" as "power to control prices or exclude competition"); *see also* ABA SECTION OF ANTITRUST LAW, ANTITRUST LAW DEVELOPMENTS 235–36 (4th ed. 1997) (noting the 50% and 70% benchmarks and citing supporting federal case law); *id.* at 238 (listing factors relevant to monopoly power determinations such as "presence and degree of barriers to entry or expansion, technological superiority resulting in cost advantages, economies of scale and scope, ability to price discriminate, the relative size of competitors, competitors' performance, pricing trends and practices, homogeneity of products, potential competition, and the stability of market shares over time"); United States v. Syufy Enters., 903 F.2d 659, 664 (9th Cir. 1990) ("A high market share, though it may ordinarily raise an inference of monopoly power, will not do so in a marker with low entry barriers . . . .").

70. Such valuations reflect the quality, variety, and price of the available applications, and are measured assuming that the platform is already purchased.

71. The argument as formulated yields a slightly sharper conclusion than is usually stated. First, it is the incremental value to the marginal platform purchaser that counts. Second, if the platform provider chooses a different price strategy than that described, it will more than capture the advantage of the more efficient downstream organization. *See* Joseph

Obvious ICE neither proves nor assumes that competition in applications markets is efficient. If, for instance, it is exceptionally hard to avoid spillovers of innovation among applications developers, then competition among developers might lead to less rather than more innovation. Or, if consumers cannot easily judge the quality of applications, fly-by-night entry into applications could spoil the market. If, for such reasons, a competitive applications market would yield less value than a monopolized one, the monopoly platform provider would gain by efficiently preventing competition in the market for applications. Thus, Obvious ICE does not say what structure of the applications market is optimal, but simply observes that the unintegrated platform monopolist has an incentive to favor whichever form of organization of applications is most efficient (or delivers the most value to users).

But often a platform monopolist *does* integrate into (and remain in) the market for applications for its platform.[72] For at least three reasons, it will often be able to take a dominant position in that business. First, it has a stronger incentive than an independent firm to work harder on its applications: while innovators can seldom capture all their incremental value through simple pricing, the integrated provider (as ICE reminds us) can capture some — perhaps all — of the residue in its platform sales. Second, even if a platform provider truly tries to cooperate with independent applications developers, it is unlikely to be as open with them as with its own applications division (unless it builds a "Chinese wall" to keep information from the latter). Third, if the integrated firm wants to hamstring applications rivals, it might be very easy to bias interface design, the timing of new releases, pricing policy, and other choices. Moreover, such subtleties would only be necessary if blunter means (e.g., rendering an application inoperable) were unavailable. For these reasons, a platform provider's decision to integrate vertically may well hurt independent complementors, seemingly posing formidable competitive concerns.[73]

Obvious ICE does not address these concerns. But a stronger and less obvious version of ICE claims that platform monopolists will act efficiently even in deciding whether or not to integrate into adjacent markets. Similarly, this version of ICE claims that if a platform monopolist integrates into an adjacent market, it will still welcome value-added innovations by independent firms. Thus, according to this form of ICE, such close vertical relationships do not raise economic policy concerns.

Farrell, *Integration and Independent Innovation on a Network*, 93 AM. ECON. REV. 420 (2003).

72. This need not be literal integration; alliances with particular applications developers could have similar effects. Therefore, this Article sometimes refers to "close vertical relationships" instead of using the traditional term "vertical integration."

73. *See* Farrell & Katz, *supra* note 56, at 421–26; Farrell, *supra* note 71, at 421–23.

Suppose, in the hypothetical above, that the platform provider could integrate into the applications market, and by participating in that market improve the platform's value to users from $100 to $105, while breaking even on its applications. Then, it will be able to charge $5 more for its platform and sell as many as before. Whether or not it chooses just that pricing, it will profit from vertical integration, as it should, since by hypothesis integration increases value. Suppose, on the other hand, that the platform provider contemplates integrating into applications, monopolizing that market, and making a profit of $20 per user there while users value the platform at $70 rather than at $100. Because the $20 profit is less than the $30 harm created by this action — harm that is in the first instance to applications buyers, but that redounds to the platform monopolist's bottom line because consumers will be willing to pay less for the platform — it will lose by such a strategy, as it should since, by hypothesis, this strategy leads to lower overall value. To be sure, a platform provider would choose to monopolize the applications market if it could make $40 (per user) rather than $20 in doing so, but only because the assumptions imply that this monopolization would somehow increase rather than decrease total value.[74]

ICE maintains that the platform monopolist cannot increase its overall profit by monopolizing the applications market, because it could always have charged consumers a higher platform price in the first place; it has no incentive to take profits or inefficiently hamper or exclude rivals in the applications market because it can appropriate the benefits of cheap and attractive applications in its pricing of the platform. To the contrary, ICE claims that a platform monopolist has an incentive to innovate and push for improvements in its system — including better applications — in order to profit from a more valuable platform.[75]

For the reasons discussed above, firms may hesitate to enter an applications market where they must compete with the platform provider. More generally, efficient applications competition can be problematic if one of the competitors controls the platform.[76] In such cases, ICE teaches that platform providers may choose to stay out of (or exit from) the applications market altogether as a means of ensur-

74. That is, the platform provider makes an extra $40 per user at the cost of only $30 per user of reduced value. Admittedly, the assertion that this increases total value rides on an assumption that excluded applications firms do not capture more than the $10 difference in pure profits.

75. *See, e.g.*, United States v. Microsoft Corp., 84 F. Supp. 2d 9, 26 (D.D.C. 1999) ("[I]f there are innovations that will make Intel-compatible PC systems attractive to more consumers, and those consumers less sensitive to the price of Windows, the innovations will translate into increased profits for Microsoft.").

76. *See, e.g.*, Farrell, *supra* note 71, at 421–24.

ing efficient competition in that market. (Palm's recent break-up may illustrate such a motive.)

The more ambitious version of ICE is a close kin to the "one monopoly profit theory,"[77] which dates back to early Chicago School thinking and the later work of Richard Posner and Robert Bork.[78] But the "one monopoly profit" label[79] captures only part of ICE. It claims that a platform monopolist cannot gain by inefficiently leveraging its market power into applications: this is ICE's claim that where competition in the applications market is efficient, the platform monopolist will protect it. But ICE goes further, stressing the broader principle that the platform monopolist gains from an *efficient* applications market — whether that be unbridled competition, integration without independents, licensing of a limited set of independents, or some attempt to combine these or other structures. The "one monopoly profit" label fails to suggest this broader point. In sum, ICE better conveys the claim that the platform monopolist has an incentive to be a good steward of the applications sector for its platform[80] and thus better captures the argument for laissez-faire vertical policies.

The stronger form of ICE largely explains modern antitrust law's reluctance to worry broadly about spillovers and leveraging of market power. It also underlies the basics of Chicago School doctrine, as well as its more ambitious arguments for the per se legality of tying ar-

---

77. Judge Posner has outlined the argument succinctly for the case where the complement is an input into the platform product:

> But the bare fact that a firm has monopoly power in Market X does not imply that it will have an incentive to obtain monopoly power over Y, an input into X. In general a monopolist like any other firm wants to minimize its input costs; the lower those costs are, the greater the monopoly profits it will be able to make. Therefore the rational monopolist will usually want his input markets to be competitive, for competition usually will minimize the costs that he has to pay for his inputs.

Olympia Equip. Leasing Co. v. W. Union Tel. Co., 797 F.2d 370, 374 (7th Cir. 1986); *see also* POSNER, *supra* note 68, at 200–02.

78. The classic statement of the Chicago School position came in Ward S. Bowman, Jr., *Tying Arrangements and the Leverage Problem*, 67 YALE L.J. 19 (1957). The orthodox restatements of it came in RICHARD A. POSNER, ANTITRUST LAW: AN ECONOMIC PERSPECTIVE 171–84 (1976) and ROBERT H. BORK, THE ANTITRUST PARADOX 372–75 (1978).

79. *See* BORK, *supra* note 78, at 229 ("[V]ertically related monopolies can take only one monopoly profit."); RICHARD A. POSNER & FRANK H. EASTERBROOK, ANTITRUST 870 (2d ed. 1981) ("There is only one monopoly profit to be made in a chain of production."). Judges, too, have used the "one monopoly profit" label. *See, e.g.*, Town of Concord v. Boston Edison Co., 915 F.2d 17, 23 (1st Cir. 1990) (Breyer, J.); W. Resources, Inc. v. Surface Transp. Bd., 109 F.3d 782, 787 (D.C. Cir. 1997).

80. In this spirit, some commentators have argued that intellectual property holders should be able to control the development and deployment of complementary products. *See* Edmund W. Kitch, *The Nature and Function of the Patent System*, 20 J.L. & ECON. 265, 277–78 (1977); *see also* Lichtman, *supra* note 58, at 615. *But see* SUZANNE SCOTCHMER, INNOVATION AND INCENTIVES (forthcoming 2004) (manuscript ch. 5, on file with authors).

rangements.[81] Surprisingly (and, as we see below, not always correctly), it suggests that antitrust and regulation should generally not worry even if an integrated firm engages in behavior within the applications market that is plainly exclusionary.

IV. HOLES IN THE ICE: WHEN ITS LOGIC CAN FAIL

ICE is a central organizing principle for the analysis of vertical competitive effects. But its claims do not always hold. In this Part, we explain eight ways in which it can fail: (1) Baxter's Law; (2) price discrimination; (3) potential competition; (4) bargaining problems; (5) incompetent incumbents; (6) option value; (7) regulatory strategy; and (8) incomplete complementarity. There are other exceptions,[82] but we find these eight particularly relevant to the information industries.

A. Baxter's Law

Even classical Chicago School adherents concede an exception to ICE where the platform (the core monopoly) is subject to regulation but the applications market is not.[83] The economics of "Baxter's Law" echo the ICE argument itself: ICE argues that a monopolist can capture in its platform profits improvements in consumer value in applications, but it generally cannot do so if the platform price is regulated. Thus, regulated platform prices can lead a monopolist to relate differently to the applications market than ICE would ordinarily suggest.

Two simple economic reasons underlie Baxter's Law. First, suppose that there is an "ideal" price cap that constrains the price of the platform product and that will not respond if platform-level profits change over time. Now consider how the regulated platform monopolist will view an opportunity to raise the price of applications and take a profit there. Assuming fixed one-to-one proportions between the platform and the applications market, suppose that the platform pro-

81. *See* BORK, *supra* note 78, at 288 (arguing that all vertical restraints should be per se legal).

82. *See* Steven C. Salop & R. Craig Romaine, *Preserving Monopoly: Economic Analysis, Legal Standards, and Microsoft*, 7 GEO. MASON L. REV. 617, 625 (1999).

83. *See* Olympia Equip. Leasing Co. v. W. Union Tel. Co., 797 F.2d 370, 374 (7th Cir. 1986) ("There are, however, special circumstances in which a rational monopolist may want to restrict competition in an input market; as it happens, one of those circumstances is where the monopolist's rates are regulated."); Jefferson Parish Hosp. Dist. No. 2 v. Hyde, 466 U.S. 2, 36 n.4 (1984) (O'Connor, J., concurring) ("In a regulated industry a firm with market power may be unable to extract a supercompetitive profit because it lacks control over the prices it charges for regulated products or services. Tying may then be used to extract that profit from sale of the unregulated, tied products or services." (citations omitted)). Bowman's initial argument contemplated this exception, *see* Bowman, *supra* note 78, at 22, but later commentators criticized this argument as too deferential to the regulatory process. *See* Louis Kaplow, *Extension of Monopoly Power Through Leverage*, 85 COLUM. L. REV. 515, 522 n.26 (1985).

vider can take an additional profit of $1 per unit in the applications market by monopolizing that market. As ICE stresses, this profit potential lowers the profit-maximizing price for its platform by $1 (in the simplest case), given the level of platform sales. But whereas this "normally" lowers platform profits by $1, it may have a far smaller effect on platform profits when the platform price is already regulated below the profit-maximizing level.[84] In a sense, the platform provider can compensate for the fact that its platform is priced below the profit-maximizing price by taking additional — and perhaps otherwise inefficient — profits in the applications market.

The second reason for Baxter's Law does not apply under an ideal price cap but does hold under some other common forms of price regulation. Suppose that the platform provider is regulated in a rate-of-return fashion, or by a price cap that responds over time to changes in platform profits. Then, by raising the price of its application product by $1 and gaining profits there, a platform provider would benefit on balance even if in the short term its profits in the platform market would fall by the full $1, because the regulatory process will over time make its platform operations whole and restore that "lost" $1.

This exception to ICE has figured prominently in telecommunications policy.[85] In particular, the Bell System allegedly leveraged its way to market power in complementary markets, denying equal access to its network to competitors in long distance and equipment manufacturing.[86] By excluding such competitors, AT&T could rent telephones to its customers and sell equipment from its Western Electric affiliate to its operating companies or telephone subscribers at inflated rates. Such a strategy was *available* to AT&T because of its network-level market power, but ICE would claim the option should be *unattractive* because it would decrease demand for telephone sub-

---

84. The loss of demand is the $1 divided by the absolute slope of the demand curve, so it is -dx/dp, or (-dx/dp)/x per unit sales. Multiplying by the gross margin (p-MC) gives (p-MC)(-dx/dp)/x, or [(p-MC)/p] * (-p/x dx/dp). This is the Lerner markup index times the absolute elasticity of demand; this amounts to 1 if p is profit-maximizing, and is less than 1 if p is below the profit-maximizing level.

85. This issue also emerged in cases involving railroad regulation. *See* N. Pac. R.R. v. United States, 356 U.S. 1, 8 (1958) (noting that land grant sales conditioned on "preferential routing" might well be an example of a tie used as a substitute for an unlawful rebate); *see also* Kaplow, *supra* note 83, at 522 n.26.

86. *See* Roger G. Noll & Bruce M. Owen, *The Anticompetitive Uses of Regulation:* United States v. AT&T, *in* THE ANTITRUST REVOLUTION 328, 339–44 (John E. Kwoka, Jr. & Lawrence J. White eds., 2d ed. HarperCollins 1994), *available at* http://www.oup.com/us/antitrustrevolution (last visited Oct. 29, 2003). In theory, an ideal "global price cap" could restore ICE, but modern telecommunications regulation rarely focuses on this goal, instead aiming to deregulate workably competitive segments. For more extensive discussions of the relationship of ICE to regulation, see JEAN-JACQUES LAFFONT & JEAN TIROLE, COMPETITION IN TELECOMMUNICATIONS (2000); B. Douglas Bernheim & Robert D. Willig, The Scope of Competition in Telecommunications (1996) (unpublished manuscript, on file with authors); and MARK ARMSTRONG ET AL., REGULATORY REFORM: ECONOMIC ANALYSIS AND BRITISH EXPERIENCE (1994).

scription. But that decrease did not deter AT&T because of the price regulation of local telephone service. In the Carterfone decision and its aftermath, the FCC imposed an "unbundling" requirement on AT&T to prevent it from requiring consumers to rent phones, and thereby opened the customer premises equipment ("CPE") market to competition.[87] This issue was at the heart of the government's antitrust case against AT&T, even though AT&T's long-distance rates — like its local ones — were regulated.[88] Because then-Assistant Attorney General Baxter highlighted this hole in ICE in championing the consent decree that broke up AT&T, this exception is termed "Baxter's Law" or the "Bell Doctrine."[89]

B. Price Discrimination

Participating in, or dominating, the applications market can help a platform monopolist to price discriminate; this objective may make even inefficient vertical leveraging profitable.[90] Control over applications can help a platform monopolist to engage in price discrimination, charging different markups on combinations of the platform with different sets of applications.[91] It can customize its offerings for different buyers, separating "inframarginal" customers who are willing to pay more, from "marginal" customers who would switch to other alternatives in the face of a price increase.[92] Price discrimination is familiar in airline travel, where airlines use various means to segment

87. *See supra* note 34 and accompanying text.

88. AT&T was federally regulated as a dominant carrier in the interstate long-distance market until 1995. *See* Motion of AT&T Corp. to be Reclassified as a Non-Dominant Carrier, 11 F.C.C.R. 3271, ¶¶ 10–13 (1995) (Order) (ending rate regulation of AT&T in the long-distance market). The fact that AT&T faced regulation in its complementary markets — both in long distance and, in some cases, in CPE — suggests that the justification for regulatory action based on Baxter's Law is more complicated than often appreciated. Three possible variations on this classic explanation might address this complication. First, areas that regulation did not address — such as certain CPE markets, particularly where AT&T sold equipment to itself — may have been open to abuses. Second, imperfections of regulation may have enabled the AT&T monopoly to take greater advantage of consumers by providing both the monopoly and complementary service — i.e., the end of vertical integration helped consumers by facilitating better regulation. Finally, the ability to prevent competition might have helped AT&T to forestall innovation in complementary markets that would force it to depreciate its sunk investments more quickly than it wished.

89. *See supra* note 40 and accompanying text.

90. Proponents of the "leverage theory" of tying regularly invoke this explanation. *See, e.g.*, Lawrence A. Sullivan, *Section 2 of the Sherman Act and Vertical Strategies by Dominant Firms*, 21 SW. U. L. REV. 1227, 1237 (1992); *see also* Kaplow, *supra* note 83, at 523 ("[P]ractices merely increasing profits to an existing monopoly, without 'extending' it, can increase the welfare loss that results.").

91. For example, in the Internet environment, customer identity might be more readily tracked through the complement than through the platform product.

92. For a discussion and explanation of the difference between "marginal" and "inframarginal" customers, see James A. Keyte, *Market Definition and Differentiated Products: The Need for a Workable Standard*, 63 ANTITRUST L.J. 697, 739–45 (1995).

the market and extract premium prices from inframarginal business travelers who cannot plan in advance. In telecommunications, both incumbents and entrants practice price discrimination by offering different tiers of packages or sets of offerings to different customers.[93]

Price discrimination need not in itself be inefficient or anti-consumer, but the platform monopolist's desire to price discriminate can outweigh ICE and lead it to exclude efficient innovation or price competition in complementary products. In the classic case, the monopolist does so more or less intentionally because control of the complementary market allows it to maximize profits through large markups on complementary goods — for example, the substantially higher ticket prices charged to first class airline passengers for better meal service.[94] In other cases, profit maximizing price discrimination involves *below*-cost pricing of complements. The platform provider has no motive to exclude such offerings per se and probably would be delighted if independent complementors were to offer cheap and innovative offerings; independent developers, however, may refrain from providing such products where the platform provider offers its own complements below cost.[95]

Economists recognize that price discrimination can either harm or benefit consumers overall (and is likely to harm some and benefit others).[96] Some forms of price discrimination, like Ramsey pricing,[97] can raise profits at the lowest possible cost to consumers as a group, and

---

93. *See, e.g.*, Tiffany Kane, *Legislators Laud Debut of Covad's Service* (June 19, 2002) (reporting on Covad's tiered pricing structure), *at* http://news.com.com/2100-1033-937523.html (last visited Oct. 29, 2003); *see also* CARL SHAPIRO & HAL R. VARIAN, INFORMATION RULES 300 (1998) (indicating that price discrimination is predictably prevalent in information industries).

94. This, of course, does not explain why the platform monopolist seeks to co-opt and tame independent innovation rather than welcoming it. But allowing independent innovation while restricting the available independent product offerings to facilitate price discrimination may prove either infeasible or unadministrable.

95. *See* Farrell & Katz, *supra* note 56 (formally modeling such an effect). A platform provider could alternatively offer a uniform subsidy to independent as well as its own complements, which might avoid this problem but raise others.

96. *See* Richard A. Posner, *The Chicago School of Antitrust Analysis*, 127 U. PA. L. REV. 925, 926–28 (1979) (explaining how price discrimination can reduce the "misallocative effects of monopoly"). Moreover, if price discrimination increases output and thus generates economies of scale and/or "learning by doing" efficiencies, unit cost of production will drop. *See* Jerry Hausman & Jeffrey MacKie-Mason, *Price Discrimination and Patent Policy*, 19 RAND J. ECON. 253, 257 (1988). Finally, even if price discrimination is harmful, policies to limit it may have unintended consequences, such as leading firms to use cruder means of achieving the same purpose. *See* Hal R. Varian, *Price Discrimination*, *in* 1 HANDBOOK OF INDUSTRIAL ORGANIZATION 597, 644–46 (Schmalensee & Willig eds., 1989); Hausman & MacKie-Mason, *supra*, at 257.

97. As Justice Breyer explained, "Ramsey pricing is a classical regulatory pricing system that assigns fixed costs in a way that helps maintain services for customers who cannot (or will not) pay higher prices." AT&T v. Iowa Util. Bd., 525 U.S. 366, 426 (1999) (Breyer, J., concurring in part and dissenting in part).

this is valuable where profits are an important spur to innovation.[98] Thus modern economics is not generally hostile to price discrimination.

As a result, some commentators do not see price discrimination as an exception to ICE.[99] But it is. Even where price discrimination itself *enhances* efficiency, the platform monopolist may impose highly inefficient restrictions on applications competition in order to engage in price discrimination.

To illustrate, consider the attitude of cable providers toward streaming video applications over their cable modems. ICE would suggest that cable providers should happily endorse this use of their platform, as it would make the platform more valuable to users and therefore more profitable. But a cable provider who allows video streaming will find it harder to engage in the profitable and customary price discrimination that sets high markups for premium cable programming. Thus, a cable provider might rationally, but inefficiently, try to stop this innovative method of distribution.[100]

C. Potential Competition

Platform monopolists will evaluate actions in complementary markets through two lenses. On the one hand, ICE reminds us that the platform franchise often is worth more when the complement is efficiently supplied. On the other hand, competition in the complement can sometimes threaten the primary monopoly.[101] Thus, even if a two-

98. *See* Hausman & MacKie-Mason, *supra* note 96, at 263 (allowing for price discrimination in the sale of a patented product can spur innovation and thus substitute for longer intellectual property protection).

99. *See, e.g.*, Town of Concord v. Boston Edison Co., 915 F.2d 17, 24 (1st Cir. 1990); BORK, *supra* note 78, at 241–42; POSNER, *supra* note 68, at 203–06.

100. For an anecdotal suggestion that cable providers may fear such effects, see David Lieberman, *Media Giants' Net Change Major Companies Establish Strong Foothold Online*, USA TODAY, Dec. 14, 1999, at B3 (reporting that Dan Somers, CEO of AT&T Broadband, dismissed suggestions that it would allow video streaming of programming on the ground that "AT&T did not spend $56 billion to get into the cable business 'to have the blood sucked out of our veins'").

101. Some have argued that this point adds a dynamic element to the analysis that the traditional Chicago School model lacks. *See* Kaplow, *supra* note 83, at 524, 527–29 (arguing that a monopolist's "motivation is to change the structural conditions it faces in the future in order that it may receive greater profits," and contrasting "dynamic" and "static" approaches); *see also* Steven C. Salop & R. Craig Romaine, *Preserving Monopoly: Economic Analysis, Legal Standards, and Microsoft*, 7 GEO. MASON L. REV. 617, 625–26 (1999) (discussing the "preserving monopoly theory" that posits that vertical integration can be used "to impede the efforts of firms that might reduce the monopolist's power and thereby cause it to reduce its prices, increase innovation or perhaps lose out to a superior rival"); Herbert Hovenkamp, *Antitrust Policy After Chicago*, 84 MICH. L. REV. 213, 261 (1985) (criticizing Chicago School orthodoxy as focused on "static" analysis and unable to take account of "strategic behavior"). For an economic model of tying strategy, see Dennis W. Carlton & Michael Waldman, *The Strategic Use of Tying to Preserve and Create Market Power in Evolving Industries*, 33 RAND J. ECON. 194 (2002).

level monopoly may not yield more than one monopoly profit, it can protect the monopolist against entry in several ways.[102]

First, if there are no independent applications suppliers, any potential platform rival would need to enter at both the platform and applications levels.[103] This "two-level entry" theory is familiar to both telecommunications regulation and antitrust policy. For example, the program access provisions of the Cable Policy Act of 1992 give satellite firms access to cable networks affiliated with rival cable operators in order to ensure that satellite providers can compete effectively with cable and are not hindered by a lack of programming availability.[104]

The two-level entry theory also underlaid the Justice Department's challenge to General Electric's licensing policies for medical imaging equipment. The company had contractually restricted hospitals from servicing the equipment of other hospitals. The DOJ argued that these restrictions illegally raised barriers to entry in the market for medical imaging equipment.[105] If hospitals' service staffs learned to service outside equipment, new equipment providers would need only to enter the equipment market, relying on hospital service staffs to service their own equipment and that of other hospitals.[106] Thus, this case fits our framework, with equipment playing the role of the "platform" and service the role of "applications."

Second, complements may ultimately make possible substitutes for the platform. In the *Microsoft* case, for example, Netscape's web

---

102. In his *Town of Concord* opinion, then-Chief Judge Breyer set out this justification:

> Insofar as it is more difficult for a firm to enter an industry at two levels than at one, the monopolist, by expanding its monopoly power, has made entry by new firms more difficult. And insofar as the monopolist previously set prices cautiously to avoid attracting a competitive challenge, the added security of a two-level monopoly could even lead that monopolist to raise its prices.

915 F.2d at 23–24; *see also* Matsushita Elec. Indus. Co. v. Zenith Radio Corp., 475 U.S. 574, 591 n.15 (1986) ("[W]ithout barriers to entry, it would presumably be impossible to maintain supracompetitive prices for an extended time."); POSNER, *supra* note 68, at 202 ("[T]he possibility that tying might discourage entry into the monopolized market for the tying product cannot be excluded altogether."). For an argument along these lines, see Jay Pil Choi & Christodoulos Stefanadis, *Tying, Investment, and the Dynamic Leverage Theory*, 32 RAND J. ECON. 52 (2001) and Jay Pil Choi, *Preemptive R&D, Rent Dissipation, and the "Leverage Theory,"* 111 Q. J. OF ECON. 1153 (1996).

103. Artificially created entry barrier issues emerged as an early post-Chicago School concern. *See, e.g.*, Jonathan B. Baker, *Recent Developments in Economics That Challenge Chicago School Views*, 58 ANTITRUST L.J. 645, 651–52 (1989).

104. *See* 47 U.S.C. § 548(c)(5) (2000); Implementation of Cable Television Consumer Protection and Competition Act of 1992, 16 F.C.C.R. 19074 (2001) (Notice of Proposed Rulemaking); News Release, FCC, FCC Extends Program Access Exclusivity Rules (June 13, 2002), *at* http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-223381A1.doc (last visited Oct. 29, 2003).

105. *See* Competitive Impact Statement at 4–5, United States v. Gen. Elec. Co., No. CV-96-121-M-CCL, 1999 U.S. Dist. LEXIS 598 (D. Mont. Jan. 11, 1999), *at* http://www.usdoj.gov/atr/cases/f1800/1842.htm (last visited Oct. 29, 2003).

106. *See id.*

browser was a complementary application in the short term, but could have facilitated operating systems competition in the long term.[107] By exposing its own application programming interfaces, the browser could ultimately "commoditize" the underlying operating system.[108] As the district court found (and the court of appeals affirmed), Microsoft concluded that this was a serious threat to its core monopoly and undertook a campaign to undermine Netscape's browser.[109]

Finally, independent providers of complements may themselves be likely entrants into the platform market. Carl Shapiro, a leading economist, recently concluded that while "network monopolies can be very strong, they are most vulnerable to attack by firms in a strong position in a widely-used *complementary product*."[110] Complementors know the market and have an economic interest in lowering the price of the underlying platform (lower platform prices will raise demand for their product). For the same reason, complementors need not fear a

---

107. Lessig has called this scenario a "partial substitute." *See* Brief of Prof. Lawrence Lessig as Amicus Curiae at 46–47, United States v. Microsoft Corp., 87 F. Supp. 2d 30 (D.D.C. 2000) (No. 98-1232), *at* http://www.lessig.org/content/testimony/ab/ab.pdf (last visited Oct. 29, 2003). This term, however, does not emphasize the temporal nature of what is often called "middleware." *See* James B. Speta, *Tying, Essential Facilities, and Network Externalities: A Comment on Piraino*, 93 NW. U. L. REV. 1277, 1282 (1999) (pointing out that Microsoft's predatory actions vis-à-vis Netscape can be explained on the ground that Microsoft viewed the browser as a partial substitute for the operating system); Michael D. Whinston, *Exclusivity and Tying in* U.S. v. Microsoft: *What We Know, and Don't Know*, J. ECON. PERSP., Spring 2001, at 63, 73.

108. For a discussion based on Microsoft's internal documents (as revealed by the trial), see Timothy F. Bresnahan, *Network Effects and Microsoft*, at 23–25, *at* http://www.stanford.edu/~tbres/Microsoft/Network_Theory_and_Microsoft.pdf (last visited Oct. 29, 2003); *see also* Timothy F. Bresnahan, *A Remedy That Falls Short of Restoring Competition*, ANTITRUST, Fall 2001, at 67, 67 [hereinafter Bresnahan, *Restoring Competition*] ("[T]he development of a spectacularly innovative complementary product . . . can lower entry barriers into the monopolized market and create an opening for substitutes to make inroads and competition to emerge.").

109. United States v. Microsoft Corp., 87 F. Supp. 2d 30, 38 (D.D.C. 2000) ("In this case, Microsoft early on recognized middleware as the Trojan horse that, once having, in effect, infiltrated the applications barrier, could enable rival operating systems to enter the market for Intel-compatible PC operating systems unimpeded. Simply put, middleware threatened to demolish Microsoft's coveted monopoly power."), *aff'd,* 253 F.3d 34 (D.C. Cir. 2001); *see also* Bresnahan, *Restoring Competition*, *supra* note 108, at 67–68 (describing Microsoft's campaign).

110. Declaration of Carl Shapiro at 6, United States v. Microsoft Corp., 87 F. Supp. 2d 30 (D.D.C. 2000) (No. 98-1232), *at* http://www.usdoj.gov/atr/cases/f4600/4642.pdf (last visited Oct. 29, 2003); *see id.* (listing examples). To address Microsoft's anticompetitive tactics to defeat a complementary product that threatened its monopoly platform, Shapiro's testimony recommended divesting Microsoft's applications division from its operating systems division in order to create additional competition in the operating systems market. *See id*. at 6–7. *But see* Howard A. Shelanski & J. Gregory Sidak, *Antitrust Divestiture in Network Industries*, 68 U. CHI. L. REV. 1, 99 (2001) (criticizing Shapiro's proposal). On entry by complementors in the computer industry more broadly, see Timothy F. Bresnahan & Shane Greenstein, *Technological Competition and the Structure of the Computer Industry*, 47 J. INDUS. ECON. 1 (1999).

platform monopoly's price cuts or quality enhancements in response to entry as much as a stand-alone entrant would.[111]

In television programming and distribution, the FCC's financial interest and syndication ("finsyn") rules effectively barred the major networks (then ABC, NBC, and CBS) from the programming market and kept the major studios (then Fox, Warner Brothers, and Paramount) out of the network market.[112] In court, however, the FCC failed to justify them and they were invalidated.[113] The studios — who had been the complementary providers of programming — then entered the platform market, creating three new networks. The existing networks likewise moved quickly to create their own programming.[114] Similarly, the Telecommunications Act of 1996 ("Telecom Act") arguably assumed that the long-distance providers — who rely on the local network — were likely entrants into the local telephone market and that the local providers were almost certain entrants into the long-distance market.[115]

D. Bargaining Problems

An independent innovator and a gatekeeping platform monopolist may fail to reach a mutually beneficial access arrangement. We identify two ways in which this can happen. In the simplest such bargaining problem, a complementor develops an innovative application, but transaction costs obstruct agreement with the platform gatekeeper, and the innovation lies fallow.[116] Thus, this problem has an immediate

111. *See* Joseph Farrell, *Prospects for Deregulation in Telecommunications,* 6 INDUS. & CORP. CHANGE 719, 727 (1997).

112. *See* Schurz Communications, Inc. v. FCC, 982 F.2d 1043, 1045–48 (7th Cir. 1992).

113. Judge Posner remarked of the FCC's justification for these rules: "Stripped of verbiage, the opinion, like a Persian cat with its fur shaved, is alarmingly pale and thin." *Id.* at 1050. Many commentators have concurred with Judge Posner's critical assessment. *See, e.g.*, Crandall, *supra* note 40, at 178–79.

114. In so doing, the networks often eschewed outside programming, only much later realizing the benefits of contracting out. *See, e.g.*, Bill Carter, *Ailing ABC Turns to HBO in Search of TV Hits*, N.Y. TIMES, Aug. 5, 2002, at C1 (reporting that, after its initial hesitation, ABC decided not to rely largely on its internal production of programming, but to solicit programming from outside sources). To a degree, therefore, the finsyn rules did protect modularity and prevent vertical integration, whether or not that was desirable.

115. *See* 47 U.S.C. § 271 (2003); *see also* Joel I. Klein, Address before the American Enterprise Institute (Nov. 5, 1997) ("In essence, then, the Act envisions that the local and long distance companies will enter each other's markets and offer new and improved services, including bundled offerings of local and long distance, at better prices to consumers."), *at* http://www.usdoj.gov/atr/public/speeches/1268.htm (last visited Oct. 29, 2003).

116. *See* Michael L. Katz, *Intellectual Property Rights and Competition Policy: Four Principles in a Complex World*, 1 J. TELECOMMS. & HIGH TECH. L. 325, 342 (2002).

impact, and also discourages independent innovations in the longer run.[117]

A second kind of bargaining problem arises if the platform provider threatens to withhold access to the platform unless the application inventor licenses its new application very cheaply. If the inventor reluctantly agrees, this may be an efficient solution after the fact, but the prospect of this outcome discourages future independent invention.[118] Invoking this theory, the Federal Trade Commission ("FTC") complained that Intel's demand for intellectual property licenses from its licensees (complementors) violated the antitrust laws.[119]

Such discouragement of efficient independent innovation might be a problem inherent in closed architectures.[120] In a fully modular structure without a gatekeeper, the innovation could quickly be introduced,[121] and the innovator would profit to an extent commensurate with its innovation. But, in the longer term, ICE suggests a possible self-correcting dynamic: if the platform sponsor thinks that more complementary innovation will be forthcoming as a result, it could set up a private commons or otherwise implement modularity. Microsoft exposes many of its APIs to independent developers, spending money and resources to cooperate with complementary (applications) providers.[122] Similarly, Intel carefully manages its complementors.[123]

---

117. Economists would call this the "short run" both because it is immediate and because it is inefficient given the set of applications that have been developed, in contrast to the problem of discouraging innovations.

118. *See* Farrell & Katz, *supra* note 56, at 430 (providing an economic model to support this conclusion).

119. *See* Carl Shapiro, *Technology Cross-Licensing Practices:* FTC v. Intel, *in* THE ANTITRUST REVOLUTION 350, 361–63 (John E. Kwoka, Jr. & Lawrence J. White eds., 4th ed. Oxford Univ. Press 2004).

120. To mitigate these potential barriers to innovation, intellectual property law has sought to develop certain open access doctrines. *See* Mark A. Lemley, *The Economics of Improvement in Intellectual Property Law*, 75 TEX. L. REV. 989 (1997). For an argument that copyright law should not allow complementors to access a platform standard through reverse engineering as a means of addressing transaction cost issues, see Lichtman, *supra* note 58, at 637–38. For an argument that patent law should give second-generation inventors legal protection to facilitate fair arrangements with original inventors and to address the hold-up problem, see Suzanne Scotchmer, *Standing on the Shoulders of Giants: Cumulative Research and the Patent Law*, J. ECON. PERSP., Winter 1991, at 29. Moreover, real property law also recognizes that "strategic behavior" can prevent a socially desirable arrangement and provides for flexibility in crafting appropriate relief to avoid this outcome. *See, e.g.*, Walgreen Co. v. Sara Creek Prop. Co., 966 F.2d 273, 276–79 (7th Cir. 1992) (Posner, J.).

121. In the Internet environment, for example, the openness of the logical standard allows developers like Napster to introduce applications without first reaching agreement with a network owner. *See supra* notes 20–21 and accompanying text.

122. *See* MICHAEL A. CUSUMANO & RICHARD W. SELBY, MICROSOFT SECRETS: HOW THE WORLD'S MOST POWERFUL SOFTWARE COMPANY CREATES TECHNOLOGY, SHAPES MARKETS, AND MANAGES PEOPLE 166–74 (1995). Under the proposed consent decree reached with the Justice Department, Microsoft would formalize — and be subject to judicial oversight related to — the disclosure of information on its otherwise proprietary interfaces. *See* Joe Wilcox, *Microsoft Tallies Antitrust Efforts* (Aug. 5, 2002), *at* http://news.com/2100-1009-948440.html (last visited Oct. 29, 2003).

A platform monopolist may find it hard to make a credible commitment to modularity. One way may be to stay out of the complementary sector altogether. Just as in the *AT&T* case, where the Justice Department was skeptical that equal access was credible without divestiture and quarantine, some platform gatekeepers think their complementors will find voluntary quarantine the best guarantee of fair treatment. This could involve spinning off divisions, as AT&T and Palm did.

E. Incompetent Incumbents

A platform monopolist will not behave as ICE predicts if it fails to understand ICE. Some applications of ICE are surprising even for professional economists.[124] Thus, even if there is only one monopoly profit, some may think otherwise and inefficiently seek a second.[125] And even when top management appreciates ICE, other employees may not.[126]

In our experience, businesspeople are often reluctant to help outside firms compete against internally supplied applications.[127] This may be particularly likely if the benefit of modularity comes in the form of "a hundred flowers" of diverse paths of innovation in the

---

123. On Intel's strategy, see GAWER & CUSUMANO, *supra* note 26, at 15–38. For related antitrust issues concerning Intel, see Intel Corp., 64 Fed. Reg. 20134 (Fed. Trade Comm'n Apr. 23, 1999) (Analysis to Aid Public Comment and Commissioner Statements) and Intergraph Corp. v. Intel Corp., 195 F.3d 1346 (Fed. Cir. 1999). For an economic model that explains "the theoretical rationale for the contrast between Intel's stated concern for complementors and its inability to fully commit not to behave aggressively towards them," see David Miller, *Invention Under Uncertainty and the Threat of Ex Post Entry* (June 20, 2003), *at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=319180 (last visited Oct. 29, 2003). *See also* Farrell & Katz, *supra* note 56.

124. For example, consider competition when a platform monopoly such as an incumbent local exchange company ("ILEC") charges an "access charge" above marginal cost to its downstream rivals (for example, long-distance providers). If demand is totally inelastic, ICE implies that no imputation rule is necessary to ensure that the ILEC should charge itself the same access charge as it charges rivals. *See, e.g.*, Farrell, *supra* note 71, at 423. In our experience, however, this "opportunity cost argument" is not obvious to policymakers, businesspeople, or even distinguished economists.

125. *See* Kaplow, *supra* note 83, at 548–49 ("[O]ne might argue that even if a leveraging strategy is unprofitable or doomed to complete failure in the long run, many firms cling to a misguided belief that they can succeed.").

126. Some courts have acknowledged this possibility. For example, the court in *Time Warner Entertainment Co. v. FCC* noted that a company may be "reluctant to ditch or curtail an inefficient in-house operation because of the impact on firm executives or other employees, or the resulting spotlight on management's earlier judgment." 240 F.3d 1126, 1138 (D.C. Cir. 2001). Such resistance to change might be rational for individual managers wanting to avoid the detection of their own mistakes, but irrational for the company as a whole, which would suffer from the lack of superior applications for its platform product. In economics, this is called a "principal-agent" problem.

127. *See, e.g.*, Bruce T. Allen, *Vertical Integration and Market Foreclosure: The Case of Cement and Concrete*, 14 J. L. & ECON. 251, 270–72 (1971) (offering this explanation for vertical integration).

complement.[128] Incumbents may fail to imagine the potential benefits of increased competition in the market for applications, and thus fail to implement modularity even when it would spur greater innovation and thus increase their platform profits.

ICE's insights for business strategy may be particularly hard to see for industries emerging from a regulated environment of end-to-end service.[129] ICE-aware business commentators have argued that the customer relationship business, the product innovation business, and the infrastructure business can be "unbundled" from one another to great efficiency benefits,[130] but that regulated incumbent firms often miss this opportunity.[131] Thus, two commentators claim that the local telephone companies have "deliberately limited the growth and profitability of their infrastructure businesses to protect their customer relationship businesses."[132]

ICE-savvy commentators also argue that Apple erred in the early 1980s by not licensing its operating system so that others could build computer systems around it.[133] Apple had developed an operating system widely viewed as better than Microsoft's MS-DOS (which IBM and others licensed),[134] but thought it could make more money by

---

128. This recalls Mao Tse-Tung's famous adage: "Let a hundred flowers bloom, let a hundred schools of thought contend." THE ENCYCLOPEDIA OF WORLD HISTORY (Peter N. Stearns ed., Houghton Mifflin CD-ROM 6th ed. 2001), *available at* http://www.bartleby.com/67/4149.html (last visited Nov. 30, 2003). For explorations of the economics of innovation and diversity, see Raaj Kumar Sah & Joseph E. Stiglitz, *The Invariance of Market Innovation to the Number of Firms*, 18 RAND J. ECON. 98 (1987) and Joseph Farrell et al., *Market Structure, Organizational Structure, and R&D Diversity*, *in* ECONOMICS FOR AN IMPERFECT WORLD (Richard Arnott et al. eds., 2003).

129. *See* Clifford Winston, *U.S. Industry Adjustment to Economic Deregulation*, J. ECON. PERSP., Summer 1998, at 89, 89 (stating that, under the influence of regulation, "managers and employees of regulated firms settle into patterns of inefficient production and missed opportunities for technological advance and entry into new markets"); *id.* at 98 (noting that it takes time for the management of formerly regulated monopolists to move to a more entrepreneurial culture).

130. *See* John Hagel III & Marc Singer, *Unbundling the Corporation*, MCKINSEY Q., 2000, No. 3, at 148, *available at* http://www.optimizemagazine.com/mckinsey/2002/0408.htm (last visited Oct. 29, 2003).

131. *See id.* This concern underlies the much-discussed proposal of imposing a wholesale-retail separation of the incumbent local telephone providers' operations. For debate on this, compare Beard et al., *supra* note 67, at 421 (2002) with Robert W. Crandall & J. Gregory Sidak, *Is Structural Separation of Incumbent Local Exchange Carriers Necessary for Competition?*, 19 YALE J. ON REG. 335 (2002).

132. Hagel & Singer, *supra* note 130, at 154.

133. *See* Charles R. Morris & Charles H. Ferguson, *How Architecture Wins Technology Wars*, HARV. BUS. REV., Mar.–Apr. 1993, at 86, 90 (noting how Apple's refusal to open the Macintosh platform hurt it in the marketplace); *see also* Joseph Farrell, *Standardization and Intellectual Property*, 30 JURIMETRICS J. 35, 42 (1989) ("As the IBM PC experience reminds us, moreover, a technology may be much more likely to set a standard if its owner chooses to renounce at least part of the prospective proprietary gains, by making the system 'open' or by widespread licensing."); Joseph Farrell & Nancy Gallini, *Second Sourcing as a Commitment: Monopoly Incentives to Attract Competition*, 103 Q. J. ECON. 673 (1988).

134. *See* Stanley M. Besen & Joseph Farrell, *Choosing How to Compete: Strategies and Tactics in Standardization*, J. ECON. PERSP., Spring 1994, at 117, 118.

bundling the operating system with its own computers. Thus, considered as an operating system platform provider, Apple bet on its own production and distribution channel rather than on a competitive hardware sector. Whether it failed to see that it was making this bet or simply overrated its hardware and distribution prowess, Apple lost the chance to be the leading producer of operating systems, realizing too late that it would have done better to promote an open architecture.[135]

If incumbents do not always fully understand ICE, what policy implications follow? Sensibly, public policy does not normally let courts or regulators tell a business how to maximize its profits.[136] Similarly, the antitrust laws and regulatory policy generally do not seek to correct business strategy failures.[137] Although we agree with this reluctance to second-guess platform providers' calculations of their best interests, one lesson does follow: the less we can count on a monopolist to be efficient even on its own terms, the more we should value platform-level competition, perhaps especially diverse competition.[138] In the case of Apple, for example, the presence of a rival platform protected customers; it also made the punishment for Apple's error more striking and more visible. Even monopolists who fail to understand ICE are punished with lower profits, but the punishment is sharper or at least more visible when there is competition among platforms. Thus, the arcane complexities of ICE and its implications boost the (already strong) case for platform-level competition.[139]

If, as Judge Posner claims, an economics-based approach has won in antitrust,[140] we urge that this salutary triumph be leavened by recognizing that competition protects not only against powerful firms with bad incentives (on which economics-based antitrust mainly focuses), but also against powerful firms with incompetent or dishonest management. When a firm fails to optimize modularity (or anything else) in a fully competitive industry, its shareholders suffer, but customers broadly do not. When a monopoly fails to do so, however, cus-

---

135. On the inferiority of Apple's strategy, see Langlois & Robertson, *supra* note 43, at 308–12.

136. For example, the "business judgment rule" used in corporate law instructs courts not to substitute their judgment for business decisions in assessing liability, provided that the decision at issue "can be attributed to any rational business purpose." *See* Sinclair Oil Corp. v. Levien, 280 A.2d 717, 720 (Del. 1971).

137. *See* William Baxter, *Legal Restrictions on Exploitation of the Patent Monopoly: An Economic Analysis*, 76 YALE L.J. 267, 318 (1966) (rejecting use of "antitrust laws to assure that private economic interests are perceived correctly"); Kaplow, *supra* note 83, at 549 (stating that the "purpose of the antitrust laws is not to improve the effectiveness of management").

138. The importance of such competition is elaborated in Weiser, *supra* note 57.

139. In evaluating, for example, mergers between platform providers, antitrust enforcers should be mindful of the competitive impact related to the loss of rival platforms and the associated experimentation that arises from such platform diversity.

140. *See* POSNER, ANTITRUST LAW, *supra* note 68, at ix.

tomers often suffer. Antitrust and regulation should thus aim to protect against incompetent monopolies as well as against rapacious ones.

F. Option Value

Perversely, fear of access regulation may itself discourage a firm from opening its platform. After a monopolist allows open access to its platform, it may not later be allowed to pursue a closed or fully integrated strategy. Under current antitrust jurisprudence, for example, a firm is far more likely to get into trouble for closing a previously open platform than for never opening it in the first place.[141] Some commentators and judges have noted the adverse ex ante effect of imposing liability for changing a cooperative practice and have cautioned courts against imposing such liability,[142] but the fear of such liability will not dissipate any time soon. Consequently, a firm may keep its platform closed even if opening it would be more profitable, if the option value of later being able to close it is important.

Thus, suppose that the platform provider can extract $10 of profits per customer in applications by monopolizing that market and knows the demand for its platform that will result, but is uncertain about how much more valuable the platform would be to its customers if applications were competitively supplied. Suppose in particular that the firm thinks it equally likely that customers will value the platform at only $6 more (the advantages of applications competition are small) or that customers will value the platform at $12 more (competitively supplied applications are very valuable). Then the efficient path, which also (as in ICE) maximizes the firm's overall profits, is to open the platform initially, learn how much customers value that openness, and leave it open if customers turn out to value open competition in applications at $12, but close it and take over the applications market if they turn out to value openness at only $6.

But if that path is prohibited (or will attract antitrust liability), the firm must choose between opening the market forever and keeping it closed. Note that an equal chance of a $6 or $12 boost to platform demand (from applications competition) is worth less than the $10 applications profits. Therefore, if denied the option to close the plat-

---

141. *See, e.g.*, Aspen Skiing Co. v. Aspen Highlands Skiing Corp., 472 U.S. 585, 603 (1985); Eastman Kodak Co. v. Image Technical Servs., Inc., 504 U.S. 451, 483 (1992). For a sense of the debate over these cases, compare Dennis W. Carlton, *A General Analysis of Exclusionary Conduct and Refusal to Deal — Why* Aspen *and* Kodak *Are Misguided*, 68 ANTITRUST L.J. 659, 668–71 (2001) (criticizing these cases) with Jonathan B. Baker, *Promoting Innovation Competition Through the* Aspen/Kodak *Rule*, 7 GEO. MASON L. REV. 495, 501–02 (1999) (endorsing these cases).

142. *See, e.g.*, Olympia Equip. Leasing Co. v. W. Union Tel. Co., 797 F.2d 370, 376 (7th Cir. 1986).

form later (should customers value openness at only $6), the firm will inefficiently close the platform *ab initio*.[143]

G. Regulatory Strategy Considerations

A second "iatrogenic"[144] exception to ICE arises if a firm thinks allowing open access in one context will increase its regulatory duties elsewhere. For instance, a broadband transport provider might refuse to open its platform even where open access increases its profits, because it does not want to risk having to provide access elsewhere. Thus, competitive provision of broadband Internet service might add value to a cable broadband transport product, but in another, related market — say, video content — competitive providers will hurt the cable company's core product offering. If the company believed that opening up its pipe to ISPs would make it substantially more likely to have to open up to video providers, it might rationally resist open access even for ISPs. Similarly, it is unclear why AT&T would have disliked the Hush-A-Phone itself, but it might well have feared that welcoming it would have created a precedent for other attachments. In this way, the likely response of law and regulation can affect a firm's stance toward modularity.

Some firms may be more inclined than others to believe that "their" regulators will extrapolate across markets. Certainly, regulators do sometimes do so, using benchmarking between regional monopolists in devising public policy.[145] Thus, in the cable market, they may seek to preserve cable firms with different sets of assets or business strategies on the view that some cable firms will be more willing than others to experiment with open access arrangements.

---

143. A version of ICE survives: with the efficient path unavailable, the firm chooses efficiently among those that remain. But this may be cold comfort to both the firm and the rest of society.

144. This term literally means "induced in a patient by a physician's activity, manner, or therapy." AMERICAN HERITAGE DICTIONARY 867 (4th ed. 2000), *available at* http://dictionary.reference.com/search?q=iatrogenic (last visited Oct. 29, 2003).

145. This rationale underlaid the creation of different regional companies as part of the AT&T consent decree. *See* United States v. W. Elec. Co., 993 F.2d 1572, 1580 (D.C. Cir. 1993) ("[T]he existence of seven [Bell Companies] increases the number of benchmarks that can be used by regulators to detect discriminatory pricing . . . in evaluating compliance with equal access requirements . . . ."). Similarly, in approving the SBC/Ameritech merger, the FCC recognized lost benchmarking opportunities as a harm caused by the merger. *See* Ameritech Corp., 14 F.C.C.R. 14712, ¶ 5 (1999) (Memorandum Opinion and Order) ("The merger will substantially reduce the Commission's ability to implement the market-opening requirements of the 1996 Act by comparative practice oversight methods."). In that proceeding, one of us (Farrell) made this very argument on behalf of Sprint Corporation. Note that differences among regional monopolies may be what causes them to choose different strategies, so it is arguable whether regulators should be willing, a priori, to impose on one monopoly what another seemingly similarly situated one finds acceptable.

H. Incomplete Complementarity

If applications can be valuable without the platform, platform providers may profit by monopolizing the applications market. As Michael Whinston has explained, this exception to ICE arises where (1) the platform is not essential for all uses of the application (creating the incentive), and (2) there are economies of scale or network effects in the application (creating the opportunity).[146] Consider for instance a restaurant on a beach resort that some travelers visit for day trips, while others stay for a week. A monopolist hotel might profitably bundle the complement (meals) with its core offering (rooms) if doing so would foreclose the restaurant market to rivals. As ICE insists, raising the price of meals lowers longer-stay visitors' willingness to pay for rooms (if they know about it before booking), but part of the profits on meals will be extracted from day-trippers and of course the higher meal prices cannot reduce their demand for rooms.

In the information industries there are often strong complementarities between platforms and applications, so we have assumed in our exposition that applications are strict complements with the platform. In reality, however, an application for one platform — say, broadband transport — may also be useful for another — say, narrowband transport — and this may lead the broadband transport provider to try to control the applications market. Thus, this exception could prove important.

V. LESSONS FROM ICE AND ITS EXCEPTIONS, AND TOWARDS A COGENT FRAMEWORK FOR OPEN ACCESS POLICIES

As discussed in Part III, modern antitrust generally supposes that ICE is broadly right with limited and fairly easily diagnosed exceptions, and thus usually permits even dominant firms to make their own vertical choices. Courts and commentators have often heeded the basic ICE argument for skepticism about claims that a monopolist would "leverage" its primary monopoly into a second market,[147] but have often adopted a simplistic form of this logic that does not fully address ICE's exceptions.

In contrast, as telecommunications policy moved away from its disposition toward regulated integration, it turned sharply toward mandating modularity or "openness." The Hush-A-Phone and Carter-

146. *See* Michael D. Whinston, *Tying, Foreclosure, and Exclusion*, 80 AM. ECON. REV. 837, 850–55 (1990).

147. *See, e.g.*, G.K.A. Beverage Corp. v. Honickman, 55 F.3d 762, 767 (2d Cir. 1995) ("Once having achieved the alleged bottling monopoly, therefore, appellees' sole incentive is to select the cheapest method of distribution."); Advo, Inc. v. Phila. Newspapers, Inc., 51 F.3d 1191, 1203 (3d Cir. 1995) (arguing that leveraging theory "makes no sense").

fone decisions, followed by the breakup of the Bell System and the Telecom Act's unbundling provisions, reflect this shift.

Thus, some take ICE very seriously, others take its exceptions very seriously,[148] but few integrate the two in a sophisticated manner. In light of this divide, a central question is whether ICE is the rule, with relatively rare or minor exceptions, or whether ICE is actually the exception.[149] This Part discusses how ICE and its exceptions can help frame and evaluate open access obligations.

In traditional telecommunications markets, the monopoly platform was generally price-regulated, and Baxter's Law provides that ICE does not apply to regulated monopolies.[150] And the Telecom Act's unbundling obligations can be viewed and justified within this tradition. In particular, the Act's ambitious effort to regulate "bottleneck" wholesale inputs, such as the local lines to residential telephone subscribers, aims in part to replace the legacy system of retail regulation.[151] But increasingly, as with broadband platforms, telecommunications regulators confront arguments for open access regulation where the market is not generally price-regulated. To address such arguments in a fashion that is consistent with antitrust policy, policymakers must understand the different regulatory tools for facilitating modularity, the difficult tradeoffs in developing a regulatory regime, and the possible regulatory philosophies for addressing the issue. This Part addresses each issue in turn.

A. Regulatory Strategies to Facilitate Modularity

When a regulator believes (despite ICE) that modularity is both efficient and yet threatened by actual or potential vertical integration, it may seek a remedy. Competitive remedies are often divided into "structural measures" and "conduct remedies."[152] Antitrust law tends

148. For two classic responses to Chicago School thinking, see Lawrence A. Sullivan, *Section 2 of the Sherman Act and Vertical Strategies by Dominant Firms,* 21 SW. U. L. REV. 1227 (1992) and Kaplow, *supra* note 83.

149. *See* Herbert Hovenkamp, *Post-Chicago Antitrust: A Review and Critique*, 2001 COLUM. BUS. L. REV. 257, 278–79 ("The principal difference between Chicago and post-Chicago economic analysis is" the prevalence in the latter of "a complex set of assumptions about how a market works, [which make] anticompetitive outcomes seem more plausible."); *see also* POSNER, *supra* note 68, at 194–95 (maintaining that policy deviations from ICE should be the exception, not the rule). Some "die-hard" Chicagoans believe that vertical arrangements can never have anticompetitive effects (i.e., they believe that there are no exceptions to ICE), but the heavy weight of economic opinion agrees that vertical integration and vertical market restrictions can injure competition in certain cases. *See* Posner, *supra* note 96, at 932 (discussing "die-hard" Chicagoans who refuse to accept subsequent refinements of early Chicago School ideas).

150. *See* Joskow & Noll, *supra* note 40, at 1249–50.

151. For a description of the Telecom Act's market opening strategy, see Philip J. Weiser, *Federal Common Law, Cooperative Federalism, and the Enforcement of the Telecom Act*, 76 N.Y.U. L. REV. 1692, 1733–67 (2001).

152. *See* Shelanski & Sidak, *supra* note 110, at 15–16.

to favor structural measures, both to avoid enmeshing itself in closely regulating behavior and to get at the heart of the incentive and opportunity for the unlawful conduct.[153] Nonetheless, as demonstrated during the debates over remedy in the *Microsoft* case, structural remedies pose their own risks, which may ultimately dispose even antitrust courts towards conduct relief.[154]

The classic and pure structural remedy is a "quarantine" that forbids the platform monopolist from participating in the applications sector. For those who distrust a platform monopolist's stewardship of an applications market, and yet also doubt regulators' ability to stop anticompetitive behavior by other means, this approach remains a favored option.[155] But it precludes (by definition) any integrative efficiencies.[156] In addition, unless the platform/applications boundary is clean and natural, a quarantine risks becoming clumsy and artificial, as (some have argued) the quarantine imposed on the Bell Companies under the AT&T consent decree illustrates.[157]

Recognizing such problems, regulators sometimes try to get the best of both worlds, allowing the platform provider to integrate but trying to ensure that it not abuse its position. The aim is to limit the platform monopolist's behavior in the applications market only to activities associated with integrative efficiency. A fundamental problem with such best-of-both-worlds regulatory strategies is that it is difficult to know whether the anticompetitive effects of a particular approach will outweigh the efficiencies it generates; by and large, stricter rules against anticompetitive problems also risk greater collateral damage to the integrative efficiencies that presumably motivated the rejection of a quarantine. Nonetheless, regulators often seek to develop compromise approaches between quarantine and vertical laissez-faire.

---

153. *See, e.g.*, United States v. AT&T, 552 F. Supp. 131, 193 n.251 (D.D.C. 1982), *aff'd sub nom*. Maryland v. United States, 460 U.S. 1001 (1983) (recognizing core economic incentives for anticompetitive conduct and discussing the limitations of antitrust courts in superintending regulatory relief).

154. *See, e.g.*, Shelanski & Sidak, *supra* note 110, at 73–90 (examining the weaknesses of divestiture relief in the *Microsoft* case).

155. *See, e.g.*, Charles H. Ferguson, *The United States Broadband Problem: Analysis and Policy Recommendations*, at 1, 6–8 (Brookings Inst., Policy Brief #105, July 2002), *at* http://www.brookings.edu/comm/policybriefs/pb105.pdf (last visited Oct. 29, 2003).

156. Restrictions on entry may well limit competition in the applications market, but because of the possible countervailing effect, this is not a certainty.

157. *See* United States v. AT&T, 552 F. Supp. 131, 227 (D.D.C. 1982), *aff'd sub nom.* Maryland v. United States, 460 U.S. 1001 (1983) (forbidding the Bell Companies from "provid[ing] interexchange telecommunications services or information services"). For the argument that this quarantine reflected an artificial distinction between local and long-distance telecommunications services, see Mark A. Jamison, *Competition in Networking: Research Results and Implications for Further Reform*, 2002 L. REV. M.S.U.-D.C.L. 621, 622–23.

One intermediate option is a structural separation requirement. Under the Telecom Act, for example, the Bell Companies may enter the long-distance market once certain conditions are met, but must do so through a structurally separate entity.[158] This form of regulation does not necessarily change a firm's ability to discriminate against rivals, but aims to make such discrimination easier to detect and prevent by requiring the firm to deal with its own affiliate at arm's length. This approach may require policing equal access arrangements and overseeing the management of the separate subsidiary (including the imputation of any access charges).[159]

Other intermediate approaches do not control scope or structure, but order the platform provider not to discriminate in certain ways, contrary (presumably) to its assessed incentives. Conduct remedies try to control behavior directly, often requiring ongoing supervision by a regulator or court.[160] Non-structural remedies also include mandated unbundling and compatibility.[161]

Mandated unbundling requires an integrated platform provider to offer the platform without (at least some of) its applications. In its simplest form, mandated unbundling is meant to protect applications competition even for a monopoly platform. When regulators fear that an integrated platform provider will inefficiently hinder independent applications on its platform — presumably because of an exception to ICE — they may mandate that the platform product be truly open to independent applications on terms comparable to those (perhaps only implicitly) given by the platform provider "to itself."

Telecommunications regulators used an unbundling strategy to facilitate competition in the terminal equipment (applications) market

158. *See* 47 U.S.C. § 272 (2003).

159. *See, e.g.*, Amendment of Section 64.702 of the Commission's Rules and Regulations (Second Computer Inquiry), 77 F.C.C.2d 384, ¶ 205 (1980) (Final Decision) (detailing measures imposed on telephone companies to facilitate monitoring of structurally separated subsidiaries) [hereinafter Computer II], *aff'd sub nom.* Computer & Communications Indus. Ass'n v. FCC, 693 F.2d 198 (D.C. Cir. 1982).

160. *See, e.g.*, Roland Mach. Co. v. Dresser Indus., Inc., 749 F.2d 380, 391–92 (7th Cir. 1984) (Posner, J.); *see also* Ken Auletta, *Final Offer*, THE NEW YORKER, Jan. 15, 2001, at 40, 43 (quoting Chief Judge Posner, mediator in the *Microsoft* case, as stating that conduct remedy-based consent decree provisions must be "sufficiently clear to be judicially administrable and that (even if clear) they would not impose an undue administrative burden on the district court, which would have to administer the decree").

161. For ease of exposition, we use the general term "open access" to describe all measures that require the platform provider to deal with other firms with whom it might otherwise choose not to deal. Some suggest that there is little need to parse the term more narrowly, but, like most commentators, we believe that the approaches discussed above are worth analyzing separately. *Compare* Lemley & Lessig, *supra* note 10, at 969 n.139 (quarreling with the suggestion that interconnection regulation and unbundling regulation are distinct approaches) *with* Gerald R. Faulhaber, *Access ≠ Access1 + Access2*, 2002 L. REV. M.S.U.-D.C.L. 677 (making this distinction), Speta, *supra* note 14, at 252 (same), Weiser, *supra* note 10, at 826 (same), *and* Joseph D. Kearney & Thomas W. Merrill, *The Great Transformation of Regulated Industries Law*, 98 COLUM. L. REV. 1323, 1350–57 (1998) (same).

by defining an interface to AT&T's telephone network (the platform), and permitting all customer premises equipment compatible with that interface and with certain requirements to plug into the network.[162] Similarly, MCI demanded and won the right to compete against AT&T in the long-distance component (an application) of a long-distance call,[163] complementing the Bell System's provision of local exchange access (the platform, or the first and last miles of such a call). In these cases, the exception to ICE was Baxter's Law, and the goal of unbundling was to protect competition in applications, which regulators thought likely to be efficient notwithstanding AT&T's opposition. And the Telecom Act's unbundling provisions[164] are in part intended to ensure competition in the parts of the network that have the potential to sustain competition, despite whatever natural monopoly network elements remain. Again, Baxter's Law disarms ICE.

In the antitrust realm, the Ninth Circuit required Kodak to cooperate with independent providers of service (applications) for its copiers (the platform).[165] Similarly, the FTC required AOL Time Warner to offer broadband transport (the platform) separately from Internet access (applications) and to cooperate with independent Internet access providers. To regulate this requirement, the FTC relied on a benchmark arrangement between AOL Time Warner and Earthlink and appointed a monitor to oversee other such arrangements.[166] In the *Kodak* case, the exception to ICE was a concern that Kodak had engaged in a hold-up strategy; by contrast, in the AOL matter, the FTC did not clearly identify any exception to ICE.

As the AT&T antitrust litigation and Telecom Act examples illustrate, unbundling often (though not always, as the CPE example shows) requires both complex regulation and difficult price-setting.[167] Because of this requirement and because most antitrust problems do not confront Baxter's Law, antitrust commentators are often skeptical of unbundling policies.

Antitrust is, however, open to unbundling remedies when the potential competition exception to ICE applies, as our discussion of the

---

162. *See* Proposals for New or Revised Classes of Interstate and Foreign Message Toll Telephone Service (MTS) and Wide Area Telephone Service (WATS), 56 F.C.C.2d 593, ¶ 16 (1975) (First Report and Order), *modified by* 58 F.C.C.2d 716 (1976) (Memorandum Opinion and Order), *modified by* 58 F.C.C.2d 736 (1976) (Second Report and Order), *aff'd sub nom*. N.C. Utils. Comm'n v. FCC, 552 F.2d 1036 (4th Cir. 1977).

163. *See* MCI Communications v. AT&T, 708 F.2d 1081, 1105 (7th Cir. 1983).

164. 47 U.S.C. §§ 251–52 (2003).

165. *See* Image Technical Servs. v. Eastman Kodak Co., 125 F.3d 1195, 1224–28 (9th Cir. 1997).

166. *See* Faulhaber, *supra* note 161, at 684–85.

167. As Justice Breyer explained, forced sharing regimes risk undermining investment incentives if prices for the shared facilities are set too low, and create considerable administrative costs if the regime is ambitious. *See* AT&T v. Iowa Utils. Bd., 525 U.S. 366, 427–30 (1999) (Breyer, J., concurring in part and dissenting in part).

General Electric case above shows. If complementors are important in providing potential platform competition, then unbundling may be required so as to increase the chance of such competition. In the *Microsoft* case, for example, the DOJ sought and obtained a judicially overseen regime for how Microsoft manages the APIs for its Windows operating system. In particular, the court imposed regulations aimed to ensure that rival "middleware applications" can be as compatible with Microsoft's Windows operating system as are Microsoft's applications.[168] The *Microsoft* remedy aims to restore the chance of platform competition indirectly facilitated by independent middleware. Likewise, the Telecom Act's unbundling provisions are meant in part as stepping stones for the many-level entry otherwise required in order to compete against the platforms of incumbent local exchange providers.

As in the CPE example, an unbundling remedy may require regulators to ensure that technical interface standards allow independent complementors to work with the platform. A different set of policies, directed at platform-level competition, also involve compatibility mandates. Such mandates can help make "small" platforms more effective competitors when economic network effects are important.[169] Regulators can flatly require compatibility or establish a right for any firm (or only for non-dominant firms) to request or ensure it.[170] The relevant kind of compatibility depends on the nature of the network effects.

Network effects sometimes arise directly from the size of a platform's customer base, in which case a compatibility mandate should ensure *access to customers*, requiring firms to share the benefits of their combined customer networks, even if one firm contributes the majority of customers. For example, with instant messaging, the value of the service rises as a subscriber can communicate with more users. In a compatibility mandate in that market, the FCC required AOL, as part of a merger approval, to develop an interoperable instant messaging system.[171] Similarly, the Telecom Act requires every telecommu-

168. United States v. Microsoft Corp., 231 F. Supp. 2d 144 (D.D.C. 2002).

169. Economists describe a greater value of a larger network as a "network effect." For an overview, see Joseph Farrell & Paul Klemperer, *Coordination and Lock-In: Competition with Switching Costs and Network Effects*, *in* 3 HANDBOOK OF INDUSTRIAL ORGANIZATION (Mark Armstrong & Robert Porter eds., forthcoming 2004) (on file with authors) and Mark A. Lemley & David McGowan, *Legal Implications of Network Economic Effects*, 86 CAL. L. REV. 479 (1998).

170. On rights of reverse engineering to ensure compatibility, for instance, see Pamela Samuelson & Suzanne Scotchmer, *The Law and Economics of Reverse Engineering*, 111 YALE L.J. 1575 (2002).

171. *See* Applications for Consent to the Transfer of Control of Licenses and Section 214 Authorizations by Time Warner Inc. and America Online, Inc., Transferors, to AOL Time Warner Inc., Transferee, 16 F.C.C.R. 6547, ¶¶ 191–200 (2001) (Memorandum Opinion and Order); *see also* Faulhaber, *supra* note 161, at 704–05 (discussing the interoperability mandate).

nications provider to terminate calls to its subscribers from other providers, thus "socializing" the network effect.[172]

In other cases the network effect arises from a greater variety of complements available for a particular platform — an example is the "applications barrier to entry" in the *Microsoft* case.[173] To address such a network effect, a compatibility requirement may be imposed that reduces porting costs and thus ensures that applications written for one platform are readily available on others. An extreme compatibility requirement would make the platform/applications interface public and common, and thus modularize the market.

B. Considerations for Regulatory Policy

Our analysis suggests that regulators should consider two basic questions: whether an exception to ICE exists, and, if this seems likely, how well the regulator can address the competitive harms that might result. A regulatory regime that addresses both questions will minimize the opportunity for anticompetitive conduct while also being less apt to chill efficient conduct. This Section will discuss each consideration in turn.

In assessing possible exceptions to ICE, regulators should consider error costs. Courts are accustomed and explicitly instructed to weigh error costs, for instance as part of a preliminary injunction inquiry.[174] Moreover, the law has adopted a number of doctrinal devices to guard against false positives, which, in antitrust, include the opportunity for a monopolist to offer an efficiency explanation for challenged conduct.[175] Error costs include both "false positives" and "false negatives." Some Chicago School commentators argue that policymakers should worry less about false negatives, because the marketplace can ultimately address regulatorily unremedied market power abuses whereas ill-conceived regulation faces no such self-correcting mechanism.[176]

---

172. For discussion and an analogy to intellectual property policy, see Joseph Farrell, *Creating Local Competition*, 49 FED. COMM. L.J. 201, 202–04 (1996).

173. *See* United States v. Microsoft Corp., 84 F. Supp. 2d 9, 18–23 (D.D.C. 1999).

174. *See, e.g.*, Am. Hosp. Supply Co. v. Hosp. Prods. Ltd., 780 F.2d 589, 593 (7th Cir. 1986) (noting that harm calculations should account for the probability of error); *see also* Frank H. Easterbrook, *The Limits of Antitrust,* 63 TEX. L. REV. 1, 3 (1984) ("[J]udicial errors that tolerate baleful practices are self-correcting while erroneous condemnations are not."); William F. Baxter, *Reflections upon Professor Williamson's Comments*, 27 ST. LOUIS U. L.J. 315, 320 (1983) (urging courts to be mindful of "error rates" and "false positives" in judging exclusionary conduct).

175. *See* Baker, *supra* note 141, at 518.

176. For an example of the debate on this score, compare Salop & Romaine, *supra* note 82, at 653–55 (discussing varying perspectives on the relative degree of harm associated with false positives and false negatives) with Ronald A. Cass & Keith N. Hylton, *Preserving Competition: Economic Analysis, Legal Standards, and Microsoft*, 8 GEO. MASON L. REV. 1, 30–33 (1999) (arguing that false positives are more harmful than false negatives).

Regulators should also evaluate how well they can address any identified anticompetitive conduct. As antitrust law recognizes, not all marketplace harms are easily remediable. Professor Donald Turner first made this point in regard to the difficulty of policing tacit collusion between oligopolists,[177] and remediability concerns continue to figure prominently in debates over whether and how antitrust law can address single-firm conduct, as in the *Microsoft* case.[178] In the regulatory arena, this concern is both less pronounced — as regulatory bodies have greater resources than courts — and also less well considered. Remedies can also have unintended negative side effects.[179] Remedies should aim to avoid chilling efficient conduct, creating large administrative costs, or allowing opportunities for rivals to engage in strategic behavior. One guard against overbroad regulatory remedies is to ask whether less intrusive measures could be equally effective at addressing the harmful conduct.

C. Regulatory Philosophies

Our analysis suggests three basic models for the regulation of vertical relations. Each of these models ultimately converges with antitrust policy by taking account of integrative efficiencies, appreciating the logic of ICE, and acknowledging its exceptions, but each proceeds from different basic premises. In particular, the models differ in their presumptions about the reliability of assessing claimed exceptions to ICE, about the importance of vertical efficiencies, and about the FCC's ability to administer vertical regulation.[180]

In the model closest to antitrust practice, the FCC could intervene only after careful investigation compellingly shows that ICE fails along the lines of an analytically coherent exception, and that the benefits of regulation likely outweigh its costs.[181] This model thus expects that exceptions to ICE can be fairly reliably diagnosed or predicted (placing the burden on the regulator to overturn the presump-

177. *See* Donald F. Turner, *The Definition of Agreement Under the Sherman Act: Conscious Parallelism and Refusals to Deal*, 75 HARV. L. REV. 655, 671 (1962).

178. *See* Weiser, *supra* note 39, at 14–21.

179. In criticizing the finsyn rules, Judge Posner made this very argument. *See* Schurz Communications, Inc. v. FCC, 982 F.2d 1043, 1045–48 (7th Cir. 1992).

180. Of course, these models could be used not only to consider new regulation but also to consider removing old regulations in light of changed market conditions or new economic learning. In some recent decisions, the Commission has lifted restrictive regulations based on this very logic. *See* Policy and Rules Concerning the Interstate, Interexchange Marketplace, 16 F.C.C.R. 7418, ¶¶ 10, 34, 35 (2001) (Report and Order) [hereinafter Unbundling Order].

181. One of us has advocated this approach previously. *See* Philip J. Weiser, *Changing Paradigms in Telecommunications Regulation*, 71 U. COLO. L. REV. 819, 835 (2000); *see also* W. Res., Inc. v. Surface Transp. Bd., 109 F.3d 782, 788 (D.C. Cir. 1997) (noting that the Surface Transportation Board took roughly this approach).

tion that ICE applies), and that regulators are reasonably good at predicting, or diagnosing and correcting, their own failures.[182]

The two other models, while differing in substance, both reflect pessimism about regulators' ability to diagnose exceptions to ICE. Such pessimism is hardly unreasonable, since some of the exceptions sketched above might be genuinely widespread, and yet might be colorably asserted even where they do not really arise. One response to such pessimism could be a categorical protection of modularity, as advocated by some commentators.[183] An opposite response is a categorical presumption that ICE applies, as in a hard-line Chicago approach. Stating the strategies in this manner suggests a helpful way to frame the contrast between an open architecture strategy and the Chicago School approach. Some Chicago scholars appear to trust ICE more than they trust imperfect regulators or courts to diagnose its exceptions.[184] Open architecture advocates, such as Lawrence Lessig, appear to trust the history and future prospects of successful innovation through modularity more than they trust either ICE or regulators' ability to diagnose its exceptions.

VI. THE FRAMEWORK IN ACTION

Agencies and courts are often asked to decide what vertical conduct should be regulated. ICE and its exceptions, as well as the considerations noted above, can help them towards a sophisticated and consistent treatment of platform monopolists.[185] Such sophistication

182. Such rules would thus focus on "readily observable conduct whose presence or absence is highly correlated with a conclusion a court would reach were it to conduct a full analysis." Baker, *supra* note 141, at 496.

183. *See, e.g.*, Francois Bar et al., *Access and Innovation Policy for the Third Generation Internet*, 24 TELECOMM. POL'Y 489, 496 (2000) (insisting that "open access to the network led to rich experimentation by many actors whose ideas had previously been excluded from shaping network evolution"). This categorical protection might also rely on two arguments that telecommunications uniquely justifies regulatory oversight that deviates from the logic of ICE's suggestion that platform providers can be trusted. First, network industries might create greater incentives for predatory strategies, particularly those that would raise entry barriers. *See* A. Douglas Melamed, *Network Industries and Antitrust*, 23 HARV. J.L. & PUB. POL'Y 147, 149–52 (1999) (suggesting this possibility). Second, telecommunications networks — as platforms for transporting ideas — might warrant open access not based on competition policy, but on First Amendment values.

184. A compatible argument, based on Joseph Schumpeter's theory of creative destruction, is that successive battles for dominance mean that any market power gained through predatory tactics will only be temporary and thus not worth addressing. *See, e.g.*, Shelanski & Sidak, *supra* note 110, at 10–12 (discussing Schumpeterian competition, in which "firms compete through innovation for temporary market dominance, from which they may be displaced by the next wave of product advancements"); *see also* Richard Schmalensee, *Antirust Issues in Schumpeterian Industries*, 90 AMER. ECON. REV. PAP. & PROC. 192, 195 (2000) (discussing Schumpeterian competition between Microsoft and Netscape in the Internet browser market).

185. Such treatment would not only mean better results, but also a reduction in regulatory uncertainty and its associated impact on investment incentives. *See* Warren G. Lavey, *Mak-*

Monday

May 21, 2012

Incentive Regulation and Its Application to Electricity Networks

PAUL L. JOSKOW *

Alfred P. Sloan Foundation and Department of Economics, MIT

Abstract

This paper examines developments since the publication of *The Economics of Regulation* in the theory of incentive regulation and its application to the regulation of unbundled electricity transmission and distribution networks. Conceptual mechanism design issues that arise when regulators are imperfectly informed and there is asymmetric information about costs, managerial effort, and quality of service are discussed. The design and application of price cap mechanisms and related quality of service incentives in the UK are explained. The limited literature that measures the effects of incentive regulation applied to electricity networks is reviewed.

1 Introduction

Alfred Kahn began to write what became *The Economics of Regulation* while I was an undergraduate at Cornell. He was my teacher and academic advisor at Cornell and is the one who stimulated my interest in both economics and the economics of regulation. Much has changed since *The Economics of Regulation* was published in 1970/71 (Volume 1 in 1970 and Volume 2 in 1971). Most of the industries that were thought to be "natural monopolies" and were subject to price, entry and service quality regulation at that time (for example, telecommunication, electricity, natural gas transportation, cable television, etc.) have been restructured and competition introduced into one or more of their horizontal segments.[1] Other industries, where the economic case for pervasive price and entry regulation was already increasingly being recognized as dubious by 1970, and where regulation and competition were often mixed together (for example, trucking, airlines, railroads, natural gas production), have been completely deregulated. The expanse of the economy in the U.S. and most other countries that is subject to price, entry and service quality regulation today has shrunk considerably since 1970.

There are many reasons for this trend, including changes in technology, poor performance exhibited by some regulated industries, changes in the political economy of regulation and associated changes in the power of different interest groups, and broader

* President, Alfred P. Sloan Foundation, 630 Fifth Avenue, Suite 2550, New York, NY. E-mail: joskow@sloan.org This paper is based on a longer study "Incentive Regulation in Theory and Practice: Electric Distribution and Transmission," Joskow (2006a) prepared for the National Bureau of Economic Research regulation project. http://econ-www.mit.edu/files/1181. See also Joskow (2006b). The views expressed here are my own and do not reflect the views of the Alfred P. Sloan Foundation, MIT or any other organization with which I am associated.

[1] Competition in Electricity remains a work in progress in the U.S. See Joskow (2006c).

ideological shifts favoring markets over regulation and state-owned enterprises. While the pendulum may be shifting back in some sectors (for example, financial market regulation, health and safety regulation, access of content providers to communications networks, etc.) the broad changes in the mix of regulated and competitive segments that we have observed in the last 30 years is unlikely to be reversed.

Even in countries that have gone the farthest in "liberalizing" previously regulated and state-owned enterprises, certain network segments of some of the historically regulated "natural monopoly" industries continue to be subject to price, entry and service quality regulation. These industries include electricity transmission and distribution networks, natural gas transmission and distribution networks, and water supply networks. In the case of electricity and natural gas transmission and distribution networks, the regulatory mechanisms applied to these networks have important implications for supporting wholesale competition (electricity generation, natural gas production, and associated wholesale marketing activities) and retail competition (competition to supply end-users) since they serve as platforms upon which competing wholesale and retail suppliers depend and also implement market and non-market mechanisms to maintain network reliability. Thus, good performance of the competitive segments depends on good performance of the remaining regulated network segments.

The application of sound regulatory mechanisms that affect the terms and conditions of network connections, network delivery prices, network investment, and network service quality have been important components of all successful electricity sector liberalization programs around the world. The benefits of a good regulatory framework include lower network service costs, improvements in service quality, investment to expand the network to support changes in supply and demand for network services, and the development of efficient network platforms to support robust competitive wholesale and retail markets. While many of the basic regulatory principles discussed in *The Economics of Regulation* still apply to these remaining regulated monopoly network segments, there have been important advances in both theory and application since those volumes were published as well.

Volume 1 of *The Economics of Regulation* focuses on the principles for pricing regulated services supplied by firms that are subject to budget or break-even constraints. While there have certainly been theoretical advances associated with the second-best pricing of services supplied by regulated monopoly firms since 1970, especially with regard to the design and welfare properties of non-linear prices to meet budget constraints, the application of these basic pricing principles at the retail level has not advanced very far. So, for example, except for large retail customers, time of day pricing and real time pricing for electricity and natural gas has not spread quickly at all, despite the fact that the information available from wholesale markets about generation and natural gas prices makes it even easier to apply these concepts today than in 1970. At the wholesale level, in those places where prices have been deregulated, the market naturally leads to the load varying prices whose basic economic principles are developed in detail in Volume 1.

The problem of designing rewards and incentives for efficient production by firms subject to cost-of-service or rate-of-return regulation is discussed briefly in Volume 1 (pages 53-54) and in (the more rarely read) Volume 2 (Chapters 2 and 3). The discussion in Volume II of *The Economics of Regulation* in particular raises the right issues:

> …the central institutional questions have to do with the nature and adequacy of the incentives and pressures that influence private management in making the critical economic decisions. (Volume II, page 47)

> …[rate of return regulation] creates strong incentives to pad their expenses." (Volume II, page 48)

It also has several insights into how incentives might be introduced into the regulatory process to improve performance:

> Freezing rates for the period of the [regulatory] lag imposes penalties for inefficiency, excessive conservatism, and wrong guesses, and offers rewards for the opposites …

From an overall economic efficiency perspective, these production cost and service quality inefficiencies are likely to be more important than are failures to adopt the most efficient (second best given budget constraints) pricing structures. This is the case because the efficiency losses from excessive costs lead to "first order" efficiency losses ("rectangles") while the pricing inefficiencies are likely to be second order (Harberger "triangles"). The historical focus on efficient price structures, rather than cost control and service quality incentives, likely flows in part from the political concerns about monopoly power, excessive prices, and price discrimination that played an important political role leading to the creation of regulated legal monopolies and oligopolies in the first place. But we must recognize as well that in the last fifteen or twenty years there have been major advances in imperfect and asymmetric information theory, and in the theories of incentive mechanisms and associated contractual arrangements generally that have now made it possible to develop relevant theories and then to apply them.

At the time *The Economics of Regulation* was written, there was relatively little formal theoretical development of the properties of alternative incentive regulation mechanisms that provide incentives to regulated firms to control costs, to offer appropriate levels of service quality, and to find it in their interest to set efficient (second-best) price structures. Absent relevant theory, it was difficult to develop applications that could be applied in the real world, though experiments with incentive regulation go back to the 19th century (Joskow, 2007). At the time *The Economics of Regulation* was published, the primary theoretical analysis that focused on the incentive properties of rate of return regulation was the Averch-Johnson (or as Kahn refers to it in his book, the Averch, Johnson, and Wellisz or A-J-W effect, to recognize the less widely cited paper by Stanislaw Wellisz that identified similar potential distortions from rate of return regulation (Averch and Johnson, 1962; Wellisz, 1963)). The A-J-W effect turns on the incentives created by a characterization of rate of return regulation that effectively reduces the regulated firm's effective cost of capital inputs (r) by creating a profit margin on increases in capital input while leaving fixed the price of other inputs ("labor" in the A-J model) since these input costs are assumed (that is, asymmetrically vis-à-vis capital costs) to be passed through dollar for dollar into regulated prices. This in turn leads a profit maximizing regulated firm subject to this type of regulation to make long run production decisions that use a higher capital/labor ratio than would be cost-minimizing given the firm's production function and true input costs. This theory ignores many attributes of real regulatory institutions and it has little if any empirical support (Joskow, 1974, 2007; Joskow and Rose, 1989), but for many years it was "the" positive theory of regulation. However, in the last fifteen or twenty years there have been significant advances in the theory of "incentive regulation" or "performance-based regulation" and these concepts are beginning to be applied in the regulation of electricity and gas transmission and distribution networks in a number of countries (Joskow 2006a, 2006b, 2007).

The rest of this paper identifies the key elements associated with the development of modern incentive regulation theory and then examines the application of alternative types of "incentive" or "performance-based" regulation of electricity distribution and

transmission network price levels, price structures and service quality. The discussion will assume that effective electricity sector restructuring and unbundling mechanisms have been put in place so that there are clearly defined distribution and transmission network entities offering unbundled delivery and network support services to market participants (as in the UK and portions of the U.S.). I will also assume that electricity networks are regulated monopolies[2] and that an independent regulator with adequate staff resources has been created to oversee the regulation of the distribution and transmission networks.

2 Theoretical considerations

The primary goal of regulation in the public interest is to stimulate the regulated firm to produce output efficiently from cost and quality (including reliability) perspectives, to price the associated services efficiently, and to achieve these goals consistent with satisfying a break-even or budget-balance constraint for the regulated firm that allows the firm to covered its costs of providing service while restraining its ability to exercise its market power to exploit consumers by charging excessive prices. Much of the older theoretical literature on optimal (first and second-best) pricing of services provided by regulated monopolies (for example, Boiteux, Steiner, Turvey) assumes implicitly that regulators are perfectly informed about the regulated firm's cost opportunities and demand patterns and can effectively enforce cost minimization on the regulated firm.[3] The literature then focuses on first and second-best pricing of the services provided by the regulated firm given defined cost functions, demand attributes and budget balance constraints (for example, Ramsey-Boiteux pricing, non-linear pricing, etc.).[4] The older literature did not focus on incentives to minimize costs or improve other dimensions of firm performance (for example, service quality attributes), aside from making the general observation that firms insulated from competition and subject to cost-based regulation were likely to be inefficient and the limited formal theoretical developments of the A-J-W effect discussed above.

In reality, regulators care (or at least should care) as well (or more) about the production efficiency and service quality implications of the regulatory mechanisms they choose. Regulators are neither completely informed nor completely uninformed about relevant cost, quality, and demand attributes faced by the regulated firm. Regulators have *imperfect* information about these firms and market attributes and the regulated firm generally has more information about these attributes than does the regulator. Furthermore, managers have discretion to make choices not only about input proportions (as in the A-J-W models) but on how hard they will work to minimize the firm's costs or in choosing the levels of service quality. Accordingly, the regulated firm may use its information advantage (*asymmetric information)* strategically to exploit the regulatory process to

[2] The economic attributes of unregulated "merchant" transmission network investment are discussed in Joskow and Tirole (2005).

[3] An exception is the extensive theoretical and limited empirical literature following Averch and Johnson (1962), and especially after Baumol and Klevorick (1970) that examines potential distortions in input proportions caused by rate-of-return constraints. The empirical foundations for these theories are discussed in Joskow and Rose (1989).

[4] Brauetigam (1989).

increase its profits or to pursue other managerial goals, to the disadvantage of consumers (Laffont and Tirole, 1993, Chapter 1).

This creates potential moral hazard (for example, too little managerial effort resulting in excessive costs) and adverse selection (for example, prices that are too high relative to production costs) problems that effective regulatory mechanism design must address. The recent theoretical literature on incentive regulation focuses on devising regulatory mechanisms to respond to these moral hazard and adverse selection problems (Laffont and Tirole, 1993; Armstrong and Sappington, 2007).

Consider a situation in which the regulator is uncertain about the firm's true underlying costs and its opportunities further to reduce costs, the regulator cannot observe the level of managerial effort expended by the firm, but the regulator can monitor accurately the firm's realized costs ex post in regulatory hearings and through audits. The regulated firm knows its true cost opportunities, its managerial effort, and the effects of managerial effort on costs. Following Laffont and Tirole (1993, pp.10-19), under these assumptions we can think of two polar case regulatory mechanisms that may be applied to a monopoly firm producing a single product with a fixed quality. The first regulatory mechanism involves setting a fixed price ex ante that the regulated firm will be permitted to charge going forward (that is, effectively forever). In a dynamic setting this is equivalent to a pricing formula that starts with a particular price and then adjusts this price for exogenous changes in input price indices and other exogenous indices of cost drivers (again, effectively forever). This type of regulatory mechanism can be characterized as a fixed price regulatory contract or, in a dynamic setting, a price cap regulatory mechanism.

Because prices are fixed with this mechanism (or vary based only on exogenous indices of cost drivers) and do not respond to changes in managerial effort or ex post cost realizations, the firm and its managers keep 100% of any cost reductions they realize by increasing effort. Accordingly, and ignoring service quality and investment considerations for now, this mechanism provides incentives to induce efficient levels of managerial effort and in turn cost reduction. This effect is a first order "rectangle" efficiency gain. However, because the regulator must ensure that any regulatory mechanism it imposes on the regulated firm meets a budget balance constraint, when the regulator is uncertain about the regulated firm's true cost opportunities she will have to set a relatively high fixed price (or dynamic price cap) 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 (efficient but high) realized costs. Accordingly, while a fixed price mechanism does well from the perspective of providing incentives to reduce costs it is potentially very poor at "rent extraction" for the benefit of consumers and society because prices may be too high relative to the firm's true cost opportunities. The social value of rent extraction depends upon the social welfare function applied to the distribution of these rents between consumers and producers (Armstrong and Sappington, 2007) or the cost of public funds in a public procurement theoretical framework (Laffont and Tirole, 1993).

At the other extreme, the regulator could implement a simplistic pure "cost of service" regulatory contract where the firm is assured that it will be compensated for all of the costs of production that it actually incurs and no more. After the firm produces, the regulator's uncertainty about whether the firm is a relatively high or a low cost opportunity firm will be resolved. And since the regulator compensates the firm only for its realized costs, there is no "rent" left to the firm or its managers in the form of excess profits. This solves the "rent extraction" or "adverse selection" problem that would arise under a fixed price

contract. However, this kind of cost of service regulatory mechanism does not provide any incentives for the management to exert optimal (indeed any) effort. Even though there are no "excess profits" left to the firm, the actual costs incurred by the firm may be inefficiently high as a result of too little managerial effort. Managers now retain 0% of any cost savings they achieve and have no incentive to exert cost-reducing effort. Accordingly, consumers may now be paying higher prices than they would have to pay if the management could be induced to exert more effort to reduce costs. 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.

Conceptually, fixed-price contracts (or price caps) are good at providing incentives for managerial efficiency and cost minimization, but bad at extracting the benefits of the lower costs for consumers. Cost of service contracts are good at aligning prices and costs but the costs will be excessive due to suboptimal managerial effort. Perhaps not surprisingly, the optimal regulatory mechanism in the presence of imperfect and asymmetric information will lie somewhere between these two extremes. It will have a form similar to a *profit sharing* contract or a *sliding scale* regulatory mechanism where the price that the regulated firm can charge is *partially* responsive to or contingent on changes in realized costs and *partially* fixed ex ante (Schmalensee, 1989; Lyon, 1996). (I should note that Volume II of *The Economics of Regulation* discusses some early profit sharing or sliding scale plans and performance benchmarking mechanisms (pp.61-63), but expresses some skepticism about the regulator's ability to apply these mechanisms effectively.) More generally, by offering the regulated firm a *menu* of cost-contingent regulatory contracts with different cost sharing provisions, the regulator can do even better than if it offers only a single profit sharing contract (Laffont and Tirole, 1993).

3 Price cap mechanisms in practice

While the theoretical literature on incentive regulation is quite rich, it still provides relatively little direct guidance for practical application in real-world circumstances. In practice, well-designed incentive regulation programs have adopted fairly simple mechanisms that reflect some of the basic theoretical principles discussed above.

A particular form of incentive regulation was introduced for the regulated segments of the privatized electric gas, telephone and water utilities in the UK, New Zealand, Australia, and portions of Latin American as well as in the regulated segments of the telecommunications industry in the U.S.[5] This mechanism chosen is the "price cap" (Beesley and Littlechild, 1989; Brennan, 1989; Armstrong, Cowan and Vickers, 1994; Isaac, 1991). Price cap regulation is a form of institutionalized regulatory lag. 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 firms supplying multiple products or different types of customers) 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."[6] Accordingly, the price p_1 in period 1 is given by:

[5] The U.S. is behind many other countries in the application of incentive regulation principles to electric distribution and transmission, though their use is slowly spreading in the U.S. beyond telecommunications.
[6] Many implementations of price cap regulation also have "z" factors. Z factors reflect cost elements that cannot be controlled by the regulated firm and are passed through in retail prices. For example, in the UK,

(1) $$p_1 = p_o (1+ RPI – x)$$

In theory, a "forever" 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) given the level of the price cap. So to speak, it kills two birds with one stone. As already noted, however, when the regulator has imperfect information about the firm's cost opportunities and must meet a budget balance constraint, pure "forever" price cap mechanisms are not optimal from the perspective of an appropriate tradeoff between efficiency incentives and rent extraction (Schmalensee, 1989) and would leave too much rent to the firm with "average" cost characteristics. Finally, any incentive regulation mechanism that provides incentives only for cost reduction also potentially creates incentives inefficiently to reduce service quality when service quality and costs are positively correlated with one another.

In practice, "forever" price caps are not typically used in the regulation of distribution and transmission network price levels. Some form of cost-based regulation is used to set an initial value for p_o. The price cap mechanism then operates for a pre-established time period (for example, five years). At the end of this period a new starting price p_o 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 process to reset or "ratchet" prices based partially on costs realized during the previous period. In addition, price caps are often only one component of a larger portfolio of incentive mechanisms that include quality of service incentives, as discussed in the next section. Finally, regulated electric distribution and transmission network firms' ability to determine the structure of prices for different types of customers or for services provided at different locations on the network under an overall revenue cap is typically limited. As a result, the applications of price caps in practice are properly thought of as cost and quality incentive mechanism not as a mechanism to induce optimal second-best pricing of various network services. So, in practice the incentive mechanisms are only targeted at one bird rather than two.

A natural question to ask about price cap mechanisms is where does "x' (and perhaps p_o) come from or, more generally, how does one choose the correct starting value for p_o and the proper dynamic price trajectory? The difficulty of answering this question in practice is one of the sources of skepticism about formal incentive mechanisms expressed in Volume II of *The Economics of Regulation.* In England and Wales and some other countries, statistical benchmarking methods have come to be used to help to determine the relative efficiency of individual firms' operating costs and service quality compared to their peers. This information can then be used as an input to setting values for both p_o and x (Jamasb and Pollitt, 2001, 2003; OFGEM, 2004a) to provide incentives for those far from the efficiency frontier to move toward it and to reward the most efficient firms in order to induce them to stay on the efficiency frontier, in a fashion that is effectively an application of yardstick regulation (Shleifer, 1985).

Although it is not discussed too much in the theoretical or empirical literature on price caps, capital-related cost are handled quite differently from operating costs in the

---

the charges distribution companies pay for connections to the transmission network are treated as pass-throughs. Changes in property tax rates are also often treated as pass-throughs.

establishment and resetting of p_o and x. The limited attention paid to capital-related costs in the academic literature on price cap regulation provides a potentially misleading picture of the challenges associated with implementing a price-cap mechanism effectively. This is the case for several reasons. First, in practice, the p_o and x values must be developed based not only on a review of the relative efficiency of each firm's operating costs, but also based on the value of the firm's current capital stock or rate base, forecasts of future capital additions required to provide target levels of service quality, and the application of depreciation rates, estimates of the cost of the firm's debt and equity capital, assumptions about the firm's dept/equity ratio, tax allowances and other variables to turn capital stocks into prices for capital services over time. The capital cost related allowances represent a large fraction of the total price (p_o) of supplying unbundled electricity network services so the choices of these parameters for defining capital user charges are very important.

Second, allowances for capital-related costs are typically established by regulators using incentive regulation mechanisms through more traditional utility planning and cost-of-service regulatory accounting methods including the specification of a rate base (or regulatory asset value), depreciation rates, debt and equity costs, debt/equity ratios, tax allowances, etc. This is the case because the kinds of statistical benchmarking techniques that have been applied to operating costs have not been developed for capital-related costs, due to significant heterogeneity between firms in terms of the age of assets, geography, service quality, lumpiness of capital investments and other considerations. Third, the efficiency properties of a regulatory mechanism that mixes competitive benchmarking with more traditional forward-looking rate of return regulation are more complex than first meets the eye (Acemoglu and Finkelstein, 2006).

In principle, operating and capital costs could be integrated and associated benchmarks determined using total factor productivity measures. This is the approach taken by the initial price caps applied to telecom companies in the U.S. by the Federal Communications Commission. In electricity, this approach has been rejected largely because of the diversity in the capital stock, much of which is several decades old, and the associated difficulties of coming up with accurate total productivity measures. The application of price caps in England and Wales and other countries in Europe that have adopted this mechanism, benchmark a firm's performance against industry specific "best practice" (production frontier analysis using data for other firms in the industry).

Thus, the implementation of price cap mechanisms is more complicated and their efficiency properties more difficult to evaluate than is often implied and places a significant information collection, auditing and analysis burden on regulators. This is precisely the source of the skepticism about formal incentive mechanisms expressed in Volume II of *The Economics of Regulation*. In practice, modern applications of incentive regulation concepts involve the application of elements of traditional cost of service regulation, yardstick regulation, and high-powered "fixed price" incentives.

The challenge of forecasting future investment needs and costs for electricity network firms has historically been a rather contentious process, sometimes yielding significant differences between what the regulated firm's claim they need and what the regulator claims they need to meet their legal responsibilities to provide safe and reliable service efficiently. There is clearly a very serious asymmetric information problem here. In the 2004 review of electricity distribution prices in the UK, the regulator adopted an innovative "menu" of sliding scale mechanisms approach to resolve the asymmetric information problem faced by the regulator as she tries to deal with differences between

the firms' claims and the regulator's consultants' claims (OFGEM, 2004b) about future capital investment requirements to meet reliability targets. The sliding scale menu allows firms to choose between getting a lower capital expenditure allowance but a higher powered incentive (and a higher expected return on investment) that allows them to retain more of the cost reduction if they can beat the target expenditure levels or a higher capital expenditure allowance combined with a lower powered sliding scale mechanism and lower expected return. (OFGEM, 2004b). This is an application of Laffont and Tirole's menu of cost-contingent contracts mechanism and provides a more effective way to deal with the imperfect and asymmetric information conditions and associated adverse selection problems than the traditional approach of offering a single regulatory contract.

An example of the use of a profit-sharing or cost-contingent form of incentive regulatory mechanism can be found in the incentive mechanism that has been applied to the costs of the transmission system operator (SO) in England and Wales (which is also the transmission owner (TO), though there are separate regulatory mechanisms for SO and TO functions). Each year forward targets are established for the costs of system balancing services and system losses (OFGEM, 2005). A sharing or sliding scale formula is specified which places the TO at risk for a fraction (for example 30%) of deviations from this benchmark (up or down) with caps on profits and losses. There is also a cap and a floor. In recent years the SO was given a menu of three alternative incentive arrangements with different sharing fractions and different caps and floors (with costs of service as a default) from which to choose. If the SO were to choose the cost-of-service default it would suggest that in constructing the menu, the regulator had underestimated the range of the SO's future cost realizations.

4 Service quality incentives

As noted earlier, any incentive regulation mechanism that provides incentives only for cost reduction also potentially creates incentives to reduce service quality when service quality and costs are positively related to one another. The higher powered are the incentives to reduce costs, the greater the incentive to reduce quality when cost and quality are correlated. Accordingly, price cap mechanisms are increasingly accompanied by a set performance standards and associated penalties and rewards for the regulated firm for falling above or below these performance norms. Similar mechanisms are used by several U.S. states and in other countries that have liberalized their electricity sectors (for example, New Zealand, Netherlands, and Argentina).

In the UK, the regulator (OFGEM) has developed several incentive mechanisms targeted at various dimensions of distribution network service quality (OFGEM, 2004b, 2004c). OFGEM uses statistical and engineering benchmarking studies and forecasts of planned maintenance outages to develop targets for the number of customer outages and the average number of minutes per outage for each distribution company.

Until recently in the UK, there was no formal incentive mechanism that applied to transmission system reliability – network failures that lead to administrative customer outages or "unsupplied energy". In 2005, a new incentive mechanism that focuses on the reliability of the transmission network as measured by the quantity of "unsupplied energy" resulting from transmission network outages went into effect (OFGEM, 2004d). NGC is assessed penalties or received rewards when outages fall outside of a "deadband" of +/- 5%

defined by the distribution of historical outage experience (and with potential adjustments for extreme weather events), using a sliding scale with a cap and a floor on the revenue impact.

5 Performance attributes

The information burden to implement incentive regulation mechanisms well is certainly no less than for traditional cost of service regulation. Incentive regulation in practice requires a good accounting system for capital and operating costs, cost reporting protocols, data collection and reporting requirements for dimensions of performance other than costs. Capital cost accounting rules are necessary, a rate base for capital must still be defined, depreciation rates specified, and an allowed rate of return on capital determined. Comprehensive "rate cases" or "price reviews" are still required to implement "simple" price cap mechanisms. Planning processes for determining needed capital additions are an important part of the process of setting total allowed revenues and associated prices going forward. Performance benchmarks must be defined and the power of the relevant incentive mechanisms determined. What distinguishes incentive regulation in practice from traditional cost of service regulation is that this information is used more effectively because it can rely on advances in incentive regulation theory to organize and apply it. Whether the extra effort is worth it depends on whether the performance improvements justify the additional effort.

Unfortunately, there are been relatively little systematic analysis of the effects of the application of incentive regulation mechanisms on the performance of electric distribution and transmission companies.[7] Improvements in labor productivity and service quality have been documented for electric distribution systems in England and Wales, Argentina, Chile, Brazil, Peru, New Zealand and other countries (Newbery and Pollitt, 1997; Rudnick and Zolezzi, 2001; Bacon and Besant-Jones, 2001; Estache and Rodriguez-Pardina, 1998; and Pollitt, 2004). However, most of these studies have focused on developing countries where the pre-reform levels of performance were especially poor prior to restructuring. Moreover, it is difficult to disentangle the effects of privatization, restructuring and incentive regulation from one another.

The most comprehensive study of the post reform performance of the regional electricity distribution companies in the UK (distribution and supply functions) has been done by Domah and Pollitt (2001). They find significant overall increases in productivity over the period 1990 to 2000 and lower real "controllable" distribution costs compared to a number of benchmarks. However, controllable costs and overall prices first rose in the early years of the reforms before falling dramatically after 1995. Moreover, the first application of price cap mechanisms to the distribution networks in 1990 was too generous (average of RPI+ 2.5%) and a lot of rent was initially left on the table for the RECs' initial owners (who cleverly soon sold out to foreign buyers). Distribution service quality in the UK, at least as measured by supply interruptions per 100 customers and average minutes of service lost per customer, has improved as well since the restructuring and privatization initiative in 1990. This suggests that incentive regulation has not led, as some had feared,

[7] There is a much more extensive body of empirical work that examines the effects of incentive regulation mechanisms, primarily price caps, on the performance of telecommunications firms. Examples are Ai and Sappington (2004), Sappington (2003), Ai, Martinez and Sappington (2005).

to deterioration in these dimensions of service quality. This is likely to have been the case because quality standards and associated mechanisms were included in the portfolio of incentive regulation mechanisms adopted in the UK.

The experience with the transmission system operator (SO) incentive mechanism in England and Wales also provides a good example of how incentive regulation can improve performance. During the first few years following the restructuring of the electricity sector in England and Wales in 1990, the SO recovered the costs of system balancing, including managing congestion and other network constraints, through a simple cost pass-through mechanism. The SO's costs escalated rapidly, growing from about $75 million per year in 1990/91 to almost $400 million per year in 1993/94. After the introduction of the SO incentive scheme in 1994, these costs fell to about $25 million in 1999/2000. OFGEM estimates that NGC's system operating costs fell by about £400 million ($600 million at current exchange rates) between 1994 and 2001. A new SO incentive scheme was introduced when NETA went into operation in early 2001. The SO's costs have fallen by nearly 20% over the three year period since the new scheme was introduced (OFGEM, 2003).

While the empirical evidence on the effects incentive regulation mechanisms applied to electric distribution and transmission system in practice is still limited, the experience so far is very encouraging.

6 Conclusion

As we look back at developments in the theory and practice of regulating firms that have been given de facto monopoly franchises since the publication of *The Economics of Regulation* in 1970 and 1971, we can come to a number of conclusions. First, the overall economic importance of getting the theory and application of "natural monopoly" regulation right has become less important over time since a smaller fraction of the economy is subject to these types of economic regulation as competition has replaced regulated monopoly in so many industries. Moreover, even in those industries where price, entry and service quality regulation continues, it is typically being applied to a smaller number of truly natural monopoly network segments of those industries as competition has been introduced into other horizontal segments that rely on the regulated network platform. The basic theory of efficient pricing (first and second best) under the assumption that the regulator is fully informed has not changed very much over the years, except perhaps for a better understanding of the properties of non-linear pricing for regulated firms subject to budget constraints. The major advances in the theory and practice of regulation have relied on formalizing the information structure that characterizes the real world. Regulators are imperfectly informed, regulated firms have better information about the cost and demand attributes they face, and regulated firms will use this information advantage to their benefit. This situation leads to adverse selection and moral hazard problems that have been incorporated into the modern theory of incentive regulation. While applying this theory in practice must confront numerous empirical challenges, the available evidence from their application to electricity distribution and transmission systems suggests that they can help to resolve what Kahn called "the central institutional question" that confronts economic regulation.

7 References

Acemoglu, D. and A. Finkelstein (2006) "Input Technology Choice in Regulated Industries: Evidence from the Health Care Sector," mimeo, MIT, Cambridge, February.

Ai, C. and D. Sappington (2005) "Reviewing the Impact of Incentive Regulation on U.S. Telephone Service Quality," *Utilities Policy*, 13: 201-210.

Ai, C., Martinez, S., and D. Sappington (2004) "Incentive Regulation and Telecommunications Service Quality," *The Journal of Regulatory Economics*, 26:263-285.

Armstrong, M., S. Cowan and J. Vickers (1994) *Regulatory Reform: Economic Analysis and British Experience*, MIT Press: Cambridge, MA.

Armstrong, M. and D. Sappington (2007) "Recent Developments in the Theory of Regulation," in M. Armstrong and R. Porter (ed.), *Handbook of Industrial Organization (Vol. III)*, Elsevier Science Publishers: Amsterdam.

Averch, H. and L.L. Johnson (1962) "Behavior of the Firm under Regulatory Constraint," *American Economic Review*, 52: 1059-69.

Bacon, R. W., and J. E. Besant-Jones (2001) "Global Electric Power Reform, Privatization and Liberalization of the Electric Power Industry in Developing Countries," *Annual Reviews of Energy and the Environment*, 26: 331–59.

Baumol, W. and A.K. Klevorick (1970) "Input Choices and Rate of Return Regulation: An Overview of the Discussion," *Bell Journal of Economics and Management Science*, 1: 169-190.

Beesley, M. and S. Littlechild (1989) "The Regulation of Privatized Monopolies in the United Kingdom," *Rand Journal of Economics*, 20: 454-472.

Braeutigam, R. (1989) "Optimal Prices for Natural Monopolies," in R. Schmalensee and R. Willig (ed.), *Handbook of Industrial Organization, Volume II*, Elsevier Science Publishers: Amsterdam.

Brennan, T. (1989) "Regulating By Capping Prices," *Journal of Regulatory Economics*, 1: 133-147.

Domah, P.D. and M.G. Pollitt (2001) "The Restructuring and Privatisation of the Regional Electricity Companies in England and Wales: A Social Cost Benefit Analysis," *Fiscal Studies*, 22:107-146.

Estache, A. and M. Rodriguez-Pardina (1998) "Light and Lightening at the End of the Public Tunnel: The Reform of the Electricity Sector in the Southern Cone," World Bank Working Paper, May.

Isaac, R.M. (1991) "Price Cap Regulation: A Case Study of Some Pitfalls of Implementation," *Journal of Regulatory Economics*, 3: 193-210.

Jasmsb, T. and M. Pollitt (2001) "Benchmarking and Regulation: International Electricity Experience, *Utilities Policy*, 9: 107-130.

Jamasb, T. and M. Pollitt (2003) "International Benchmarking and Regulation: An Application to European Electricity Distribution Utilities," *Energy Policy*, 31: 1609-1622.

Joskow, P.L. (1974) "Inflation and Environmental Concern: Change in the Process of Public Utility Price Regulation," *Journal of Law and Economics*, 17: 291-327.

Joskow, P.L. (2006a) "Incentive Regulation in Theory and Practice: Electric Distribution and Transmission," prepared for the National Bureau of Economic Research regulation project. http://econ-www.mit.edu/faculty/download_pdf.php?id=1220

Joskow, P.L. (2006b) "Incentive Regulation for Electricity Networks," *CESifo Dice Report*, 2/2006.

Joskow, P.L. (2006c) "Markets for Power in the United States: An Interim Assessment," *The Energy Journal*, 27: 1-36.

Joskow, P.L. (2007) "Regulation of Natural Monopolies," in A.M. Polinsky and S. Shavell (ed.), *Handbook of Law and Economics (Volume 2)*, Elsevier: Amsterdam.

Joskow, P.L. and N.L. Rose (1989) "The Effects of Economic Regulation," *Handbook of Industrial Organization (Volume II)*, R. Schmalensee and R Willig editors, North-Holland, Amsterdam.

Joskow, P.L. and J. Tirole (2005) "Merchant Transmission Investment," *Journal of Industrial Economics*, 53: 233-64.

Kahn, A.E. (1970, 1971) *The Economics of Regulation: Principles and Institutions*, Volume 1 (1970) and Volume 2 (1971), Wiley: NewYork.

Laffont, J-J and J. Tirole (1993) *A Theory of Incentives in Regulation and Procurement.* MIT Press: Cambridge, MA.

Laffont, J-J and J. Tirole (2000) *Competition in Telecommunication*. MIT Press: Cambridge, MA.

Lyon, T., (1996) "A Model of the Sliding Scale," *Journal of Regulatory Economics*, 9: 227-247.

Newbery, D., and M. Pollitt, (1997) "The Restructuring and Privatization of Britain's CEGB – Was it Worth It?" *Journal of Industrial Economics*, 45: 269-303.

Office of Gas and Electricity Markets (OFGEM) (2003) "NGC System Operator Incentive Scheme from April 2004, Initial Consultation Document," December 2003. London.

Office of Gas and Electricity Markets (OFGEM) (2004a) "Electricity Distribution Price Control Review," Initial Proposals, June 2004, 145/04. London.

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) "Electricity Distribution Price Control Review: Appendix – The Losses Incentive and Quality of Service, June 2004 145e/04. London.

Office of Gas and Electricity Markets (OFGEM) (2004d) "Electricity Transmission Network Reliability Incentive Scheme: Final Proposals," December, London.

Office of Gas and Electricity Markets (OFGEM) (2005) "NGC System Operator Incentive Scheme from April 2005," Final Proposals and Statutory License Consultation, March 2005 65/05. London.

Pollitt, Michael, (2004) "Electricity Reforms in Argentina: Lessons for Developing Countries," CMI Working Paper 52, Cambridge Working Papers in Economics. http://www.econ.cam.ac.uk/electricity/publications/wp/ep52.pdf

Rudnick, H., and J. Zolezzi. (2001) "Electric Sector Deregulation and Restructuring in Latin America: Lessons to be Learnt and Possible Ways Forward," in *IEEE Proceedings Generation, Transmission and Distribution*, 148: 180-84.

Sappington, D. (2003) "The Effects of Incentive Regulation on Retail Telephone Service Quality in the United States," *Review of Network Economics*, 2: 355-375.

Schmalensee, R. (1989) "Good Regulatory Regimes," *Rand Journal of Economics*, 20: 417-436.

Schleifer, Andrei (1985) "A Theory of Yardstick Competition," *Rand Journal of Economics*, 16: 319-327.

Wellisz, S. (1963) "Regulation of Natural Gas Pipeline Companies: An Economic Analysis," *Journal of Political Economy*, 71: 30-43.

Journal of Economic Perspectives—Volume 26, Number 1—Winter 2012—Pages 29–48

Creating a Smarter U.S. Electricity Grid

Paul L. Joskow

Starting in the late 1980s, federal and state governments began to restructure and deregulate some segments of the U.S. electric power industry. The basic idea was that the generation, transmission, physical distribution, and retail supply of electricity would be "unbundled" from one another. The physical distribution of electricity and the transmission of electricity would continue to be subject to state (distribution) and federal (transmission) regulation, while generation (wholesale competition) and retail supply (retail competition) would become competitive. To support this restructuring program, a number of regulatory and organizational changes were made or planned to create and manage wholesale power markets, transmission networks, and retail competition in an efficient manner.

These reforms spread quickly during the late 1990s. Then came the California Electricity Crisis (or the Western Electricity Crisis) of 2000–2001 (Joskow 2001; Borenstein 2002). The political reaction to this crisis put a virtual halt on additional states adopting restructuring and associated retail competition reforms. It also slowed efforts by the Federal Energy Regulatory Commission (FERC) to push forward its agenda to bring organized wholesale markets, integrating the efficient dispatch and pricing of generation supplied at different locations with the efficient allocation of scarce transmission capacity, to the entire country. FERC's efforts to rationalize the balkanized ownership and operation of transmission facilities by creating Regional Transmission Authorities (RTO) managed by Independent System Operators (ISOs) were also constrained. Today about one-third of the population has access to competitive retail supply alternatives, and

■ *Paul L. Joskow is President of the Alfred P. Sloan Foundation, New York City, New York. He is also the Elizabeth and James Killian Professor of Economics, Emeritus, Massachusetts Institute of Technology, Cambridge, Massachusetts. His e-mail address is ⟨joskow@sloan.org⟩.*

doi=10.1257/jep.26.1.29

about half of the generating capacity in the country is located in regions with organized competitive wholesale markets and transmission networks managed by independent system operators (Joskow 2006).

While efforts to refine the wholesale and retail competitive market reforms continue, public policy interest has now shifted to modernizing and expanding transmission and distribution networks. In particular, this paper focuses on efforts to build what policymakers call the "smart grid" by 1) stimulating investment to improve the remote monitoring and automatic and remote control of facilities on high-voltage transmission networks; 2) stimulating investment to improve the remote monitoring, two-way communications, and automatic and remote control of local distribution networks; and 3) installing "smart" metering and associated communications capabilities on customer premises so that customers can receive real-time price information and/or take advantage of opportunities to contract with their retail supplier to manage the consumer's demands remotely in response to wholesale prices and network congestion. While the smart grid is the focus of this paper, there are other important areas for modernizing and expanding transmission networks, including stimulating investment in new transmission capacity, especially "long distance" transmission facilities that span multiple states, and better integrating electricity demand into wholesale power markets.

A recent Electric Power Research Institute (2011a, p. 1-1) report uses the following definition of the smart grid:

> The term "Smart Grid" refers to the modernization of the electricity delivery system so that it monitors, protects, and automatically optimizes the operation of its interconnected elements—from the central and distributed generator through the high-voltage transmission network and the distribution system, to industrial users and building automation systems, to energy storage installations, and to end-use consumers, and their thermostats, electric vehicles, appliances, and other household devices.

Current "smart grid" initiatives are driven by a number of potential benefits. The EPRI (2011a, p. 1-1) report correctly notes: "The present electric power delivery infrastructure was not designed to meet the needs of a restructured electricity marketplace, . . . or the increased use of renewable power production." The reference to a "restructured marketplace" emphasizes that a smarter grid can facilitate wholesale and retail competition in the supply of power, as well as the need to accelerate replacement of an aging transmission and distribution infrastructure and to conserve on meter reading and other network operating costs. The reference to renewable power points out that a smart grid may be needed if solar, wind, geothermal, and other renewable energy technologies are to make a sizable contribution to national electricity needs as well as engage with demand-side issues like charging electric vehicle batteries or encouraging consumers to use electricity more efficiently.

The American Recovery and Reinvestment Act of 2009 (ARRA) provided $4.5 billion in funds for smart grid demonstration and technology deployment projects, including various analyses of consumer behavior in response to the installation of "smart meters," discussed at ⟨http://www.smartgrid.gov/federal_initiatives⟩, a website sponsored by U.S. Department of Energy. About 140 projects have been funded under these programs with about $5.5 billon of matching funds from utilities and their customers. Several states have adopted regulations that require utilities to install smart meters and make other smart grid investments, while others have started more modestly with pilot programs. The costs of these efforts are typically recovered through regulated prices for physical distribution services. The federal funds have certainly accelerated activity on smart grid projects around the country, and these financial incentives have been reinforced by state mandates and pilot programs. Since it is unlikely that federal subsidies for smart grid investments will be sustained at their recent ARRA level, the performance of the projects supported with these funds and the experience with state mandates and pilot programs will be a powerful influence on whether and how fast smart grid investments continue in the future.

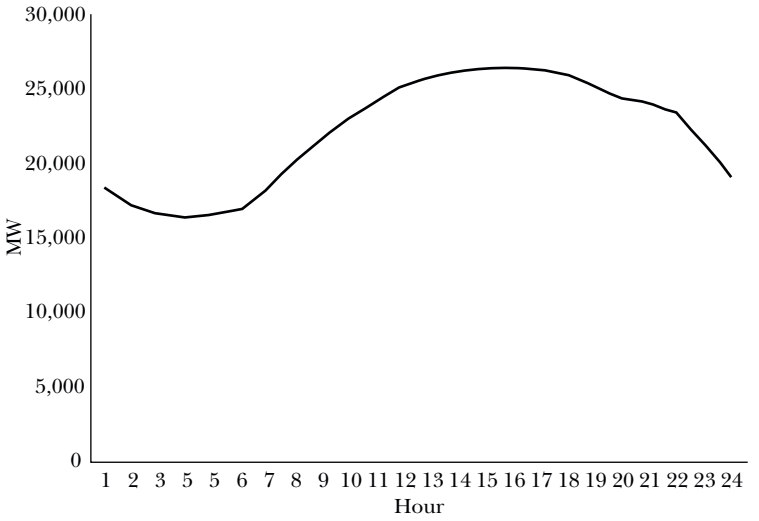
In what follows, I will examine the opportunities, challenges, and uncertainties associated with investments in "smart grid" technologies at each of the traditional components of the grid. I start by discussing some basic electricity supply and demand, pricing, and physical network attributes that are critical for understanding the opportunities and challenges associated with expanding deployment of smart grid technologies. I then discuss issues associated with the deployment of these technologies at the high voltage transmission, local distribution, and end-use metering levels. I will not discuss "behind the meter" technologies that may be installed inside of homes and businesses in response to the availability of smart grid capabilities, smart metering, and variable pricing.

Attributes of Electricity Markets

The demand for electricity varies widely from hour to hour, day to day, and month to month. Electricity demand is typically highest during the daytime hours and lowest at night. It tends to be very high on unusually hot or unusually cold days and is lowest at night on mild spring and fall days. Demand typically reaches its highest levels during only a few hours each year. There is also a minimum "base" aggregate demand that is sustained through the entire year. Figure 1 displays the levels of demand or "load" at different times of the day in New England on July 7, 2010. The peak demand is 60 percent higher than the lowest demand on that day. Figure 2 depicts the associated spot prices for electricity at each hour on that day, which I will discuss presently.

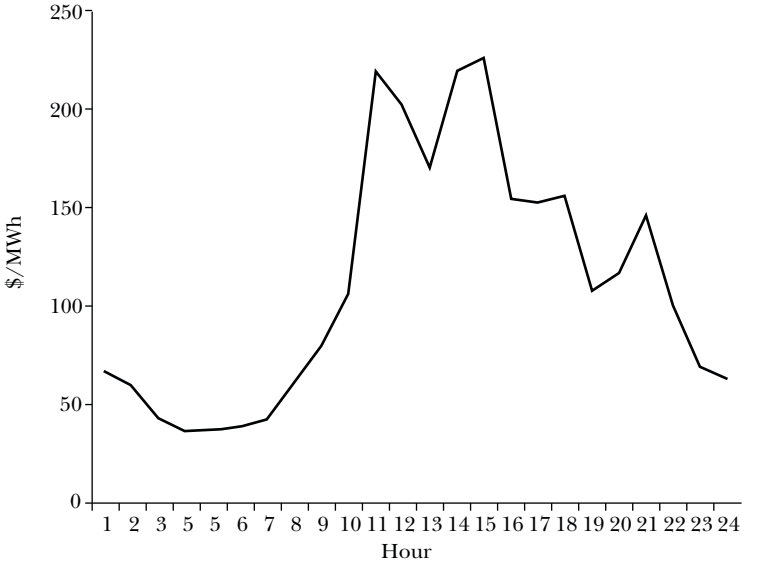
Electricity cannot be stored economically for most uses with current technologies (except in special applications where batteries, pumped storage, compressed air, and the like may be economically attractive). In electricity markets, physical

Figure 1
Real-Time Demand for Electricity, July 7, 2010
(in megawatts)



Source: Constructed from data from the New England ISO at ⟨http://www.iso-ne.com⟩.

Figure 2
Real-Time Energy Prices, July 7, 2010
(dollars per megawatt hour)



Source: Constructed from data from the New England ISO at ⟨http://www.ne-iso.com⟩.

inventories are not generally available to balance supply and demand in real time, and "stockouts" mean rolling blackouts or a larger uncontrolled system collapse (Joskow and Tirole 2007). On an electricity grid, supply and demand must be balanced continuously to maintain a variety of physical network criteria—like frequency, voltage, and capacity constraints—within narrow bounds. Electricity is the ultimate "just in time" manufacturing process, where supply must be produced to meet demand in real time.

These considerations have implications for the variations in the spot price of electricity in an unregulated wholesale electricity market and the shadow price of electricity in a traditional regulated environment that relies on an economic dispatch curve based on estimates of marginal generating costs. As demand increases, "dispatchable" generating capacity—first "base load," then "intermediate," then "peaking" capacity—with higher and higher marginal operating costs, is called to balance supply and demand (Turvey 1968; Boiteux 1964a; Joskow and Tirole 2007). As a result, wholesale market prices that reflect short-run marginal costs of generation are generally high when demand is high and low when demand is low, reflecting the marginal cost of the generation supplies needed to meet demand at different points in time. During unusually high-demand periods, supply and demand may (theoretically) be rationed on the demand side. When unexpected outages occur due to generation supply constraints or network failures, electricity consumers bear costs typically measured as the Value of Lost Load or VOLL (Stoft 2002; Joskow and Tirole 2007).

As noted, Figure 2 displays the variations in wholesale spot prices in New England associated with the variations in demand displayed in Figure 1 for a hot day in July 2010. The highest price is five times the lowest price on that day. More extreme price variability has been observed under more extreme weather conditions, though there is a $1,000 per megawatt hour cap placed on spot prices for energy in most areas ($3,000 in Texas).[1]

The prices in Figure 2 are *wholesale* spot prices. However, most *retail* residential and small commercial consumers are charged a retail price per kilowatt hour that does not vary dynamically with the time they consume electricity. As a result it does not reflect the wide variations in wholesale prices and the marginal cost of generating electricity. This is the case because traditional residential and small commercial users of electricity have meters that record only aggregate consumption between monthly or semi-monthly readings.[2] In some states, residential and small commercial consumers can opt for time-of-use meters, which charge different pre-set prices during predetermined "peak" and "off-peak" periods (for example, daytime and nighttime prices), based on averages of historical prices. While these

[1] The price caps are generally thought to be well below the Value of Lost Load in most circumstances and this raises other issues for efficient short-run and long-run performance of competitive wholesale markets (Joskow 2005).

[2] In a few cases, the largest retail consumers were billed based on prices that did vary more or less with variations in wholesale market prices (Mitchell, Manning, and Acton 1978, pp. 9–16).

time-of-use retail prices somewhat more accurately reflect variations in marginal generation costs and wholesale market prices, the relationship between actual retail prices and actual wholesale prices is necessarily very rough indeed, and penetration of time-use retail pricing has been low.

Enhancing High Voltage Transmission Systems

High voltage transmission networks are central to the operation of a modern electric power system: they make it possible to meet locationally dispersed demand with locationally dispersed generation in an efficient and reliable manner. High voltage alternating current (AC) networks are not switched networks—like a traditional railroad or telephone network—that is, power generated at point A does not flow to a specific customer located at point B. Electricity flows on an AC power network according to physical laws known as Kirchoff's laws and Ohm's law (Clayton 2001; Stoft 2002; Hogan 1992; Joskow and Tirole 2000). To drastically oversimplify, electricity produced on an AC electric power network distributes itself to follow the paths of least resistance. Transmission networks can also become congested in multiple locations, which may not lie on the "path" between buyer B and seller A, as suppliers of relatively low-cost generation seek to use the network to sell power to areas with higher cost generation. Network congestion is reflected in differences in wholesale market prices for electricity (or in shadow prices where wholesale markets with locational pricing have not been created) at different locations on the network (Hogan 1992; Joskow and Tirole 2000).

Each of the three high voltage AC networks covering the continental United States experiences significant congestion during certain hours of the year, including many "off-peak" hours, although as far I know the costs of congestion have never been quantified systematically for the entire country.[3] A natural approach to measuring the magnitude and costs of congestion is to use the differences in locational wholesale prices over time. For example, Table 1 displays the average spot wholesale prices during peak hours at different locations on the Eastern Interconnection on that same day in July 2010. It should be clear that on July 7, 2010, power was not flowing from one location to another on the Eastern Interconnection to arbitrage away large differences in wholesale spot prices; the ability of the network to transfer power from one location to another was constrained by scarce transmission capacity.

This congestion and lack of wholesale locational price arbitrage arises for three primary reasons: First, the transactions costs for moving power from the North,

[3] These costs, or at least the congestion rents, are quantified for the Regional Transmission Organizations (RTO) and Independent System Operators that have markets based on a locational marginal price market design. For example, in the PJM RTO, region, the independent market monitor estimated congestion costs to be as high as $2 billion in 2005, with substantial year-to-year variation, in his State of the Market Report for 2006 (PJM 2007, p. 39).

Table 1
Day-Ahead Peak Period Prices for Delivery July 7, 2010

Location	*$/MWh*
Boston (Massachusetts Hub)	117.75
New York City (Zone J)	138.50
Buffalo (Zone A)	79.00
Virginia (Dominion Hub)	107.75
Chicago (Illinois Hub)	68.75
Minneapolis (Minnesota Hub)	42.50
Florida	37.00

Source: Megawatt Daily, July 7, 2010, p. 2.

West, and South to New York City are high, requiring transactions with multiple Regional Transmission Organizations, Independent System Operators, and other balancing authorities with different market designs, settlement rules, and transmission service prices. Second, system operators place a very high value on reliability, which means maintaining "contingency" margins to be prepared for unanticipated events in neighboring areas which might affect their area. Third, most system operators have inadequate monitoring, communication, and control equipment on their high voltage network—an inadequate ability to "see" the state of neighboring networks—so they enforce higher contingency margins than would be necessary if they had better information and a wider span of control.

The benefit of these high contingency margins is that the U.S. electric transmission system is presently very reliable. While good comprehensive numbers are not available, it is extremely rare that retail consumers lose power because of failures of equipment or operating errors on the high voltage transmission system. EPRI (2011a, p. 2.1) estimates that U.S. power systems achieve 99.999 percent reliability at the high voltage (bulk) transmission network level and that over 90 percent of the outages experienced by retail customers are due to failures on the distribution system, not the transmission system (p. 6.1). However, when a rare major failure does occur on the high voltage transmission network, as with the 2003 Northeast Blackout when 50 million customers suffered power outages that lasted up to a couple of days, the associated costs can be high. (The 2003 Northeast Blackout was due in part to poor communications between system operators of interconnected control areas.)

It is widely accepted that there has been underinvestment in monitoring, communications, and control equipment on the high voltage transmission network to meet the needs of supporting efficient wholesale power markets, squeezing more effective capacity from existing transmission facilities, and achieving policy goals for renewable energy from grid-based wind and solar generating systems (for discussion, see EPRI, 2011a, chap. 5; New York ISO, ⟨http://www.nyiso.com/public/energy_future/issues_trends/smart_grid/index.jsp⟩; U.S. Department of Energy ⟨http://www.oe.energy.gov/⟩). EPRI (2011a, p. 5.1) recognizes that while it is hard to estimate

with precison the costs of upgrading the high voltage transmission system with this "smart" equipment, EPRI estimates that the total investment cost is $56–$64 billion. EPRI also concludes the investments in improved monitoring of high voltage transmission networks represent the most cost-effective category of smart grid investments. Investments in this category also represent about 20 percent of the total cost of EPRI's defined Smart Grid program. This is qualitatively consistent with my own assessment.

These smart grid investments at the high voltage transmission level are likely to have even higher returns as "intermittent" generating capacity like high voltage grid-connected wind and solar generating capacity grows (local photovoltaic facilities on the roofs of homes create related challenges for distribution networks—see below). High voltage grid-based wind and solar installations supply electricity intermittently. This means that their output is driven by wind speed, wind direction, cloud cover, haze, and other weather characteristics rather than by supply and demand conditions and wholesale market prices. As a result, their output typically cannot be controlled or economically dispatched by system operators based on economic criteria in the same way as traditional electricity generation technologies (Joskow 2011a, b). Since wind and sun intensity vary widely and quickly, output of intermittent generating units can vary widely from day to day, hour to hour, minute to minute, and location to location. To balance supply and demand continuously when there is significant intermittent generation on the high voltage network requires that system operators have the capability to respond very quickly to rapid changes in power flows at different locations on the network by holding more dispatchable generation in operating reserve status and having the capability to monitor and adjust the configuration of power flows on the transmission network to balance supply and demand continuously while minimizing costs.[4]

Smart grid investment on the high voltage network has only a limited ability to increase the effective capacity of transmission networks. A large increase in transmission capacity, especially if it involves accessing generating capacity at new locations remote from load centers, requires building new physical transmission capacity. However, building major new transmission lines is extremely difficult. The U.S. transmission system was not built to facilitate large movements between interconnected control areas or over long distances; rather, it was built to balance supply and demand reliably within individual utility (or holding company) service areas. While the capacity of interconnections have expanded over time, the bulk of the price differences in Table 1 are due to the fact that there is insufficient transmission capacity to move large amounts of power from, for example, Chicago to New York City. The regulatory process that determines how high voltage transmission capacity (and smart grid investments in the transmission network) is sited and paid

[4] These network issues associated with intermittent generating capacity are different from issues related to the proper comparative valuation of intermittent and dispatchable generating technologies (Joskow 2011a, b); Borenstein (2008) applies methods compatible with those in Joskow (2011a, b) to derive the cost per kilowatt hour supplied and the cost per ton of CO_2 displaced by substituting solar for fossil-fuel generation expected to result from California's rooftop solar energy subsidy program.

for in regulated transmission prices is of byzantine complexity (Joskow 2005). It is clear, however, that the combination of FERC cost allocation policies, the requirement to receive siting permits from each state in which a new transmission line is located, and not-in-my-backyard political constraints hinder efficient investment in long distance transmission lines. FERC has been trying to resolve the issue of "who pays" and "how much" for new transmission lines for years, most recently promulgating Order 1000 in July 2011. This rule has many constructive features, but it will take several years to see how and to what extent it is implemented. Nor does that rule address state siting requirements or NIMBY constraints. Congress gave the Department of Energy authority to designate National Interest Electric Transmission Corridors to respond to the diffusion of siting authority among many states, but the DOE's procedures have been rejected by the courts (Watkiss 2011). The best solution to the siting problems would be to move regional transmission planning authority from the states to FERC. However, the political barriers to such a change are enormous. Thus, underinvestment in multistate high voltage transmission facilities is likely to continue to be a problem for many years.

Automating Local Distribution Networks

The smart grid technologies being deployed on local distribution systems include enhanced remote monitoring and data acquisition of feeder loads, voltage, and disturbances; automatic switches and breakers; enhanced communications with "smart" distribution substations, transformers, and protective devices; and supporting communications infrastructure and information processing systems. Smart grid investments in local distribution networks offer a variety of potential gains: to reduce operation and maintenance costs (goodbye meter readers, manual disconnects, and responses to nonexistent network outages); to improve reliability and responses to outages; to improve power quality (for example, to eliminate very short disruptions in voltage or frequency); to integrate distributed renewable energy sources, especially solar photovoltaic systems installed at customer locations that produce power intermittently and can lead to rapid and wide variations in the (net) demand placed on the distribution network; to accommodate demands for recharging of the electric vehicle of the future; to deploy "smart meters" that can measure customers' real-time consumption and allow for dynamic pricing that reflects wholesale prices; and to expand the range of products that competing retail suppliers of electricity can offer to customers in those states that have adopted retail competition models.

The U.S. Department of Energy has supported about 70 smart grid projects involving local distribution systems on a roughly 50/50 cost sharing basis, with details available at ⟨http://www.smartgrid.gov/recovery_act/tracking_deployment/distribution⟩. However, a full transformation of local distribution systems will take many years and a lot of capital investment. Are the benefits likely to exceed the costs?

In the only comprehensive and publicly available effort at cost–benefit analysis in this area, the Electric Power Research Institute (2011a) estimates that deployment (to about 55 percent of distribution feeders) would cost between $120–$170 billion, and claims that the benefits in terms of greater reliability of the electricity supply would be about $600 billion (both in net present value).[5] Unfortunately, I found the benefit analyses to be speculative and impossible to reproduce given the information made available in EPRI's report.

According to EPRI (2011a, page 6.1), over 90 percent of the electricity supply outages experienced by retail electricity consumers occur because of failures on the local distribution network rather than the transmission network. These failures may be caused by wind and storms, tree limbs falling on overhead distribution lines, icing up of distribution equipment, overloads of the local distribution network, failures of low-voltage transformers and breakers due to age or poor maintenance, cars that crash into poles and knock down distribution equipment, flooding of underground distribution, excessive heat, natural aging, and so on. No matter how smart we make local distribution systems, power outages will arise from many of the natural causes on this list, especially in areas that rely on overhead (rather than underground) distribution lines. Using standard measurement criteria from the IEEE (Institute of Electrical and Electronics Engineers), which exclude certain planned and weather-related outages, the average residential household has (rounding to simplify the calculation) 1.5 unplanned outages per year with an average outage duration of about 100 minutes per year (Power Engineering Society, 2006). Accordingly, the average residential customer experiences about 150 minutes of unexpected outages per year or 10.5 percent of one day per year. When I compare EPRI's estimates of the benefits from greater reliability with typical estimates of Value of Lost Load—for example, $5,000 to $10,000 per megawatt hour lost—the EPRI estimates of reliability benefits appear much too high.

Very short voltage drops and electrical transients that appear almost as flickers of lights (poor "power quality")—potentially create significant problems for very sensitive digital equipment. Investments in smart distribution grid technology can reduce these transients, but at significant cost. The value of reducing these transients also varies widely among customers. Having to reset one's clock is less costly than maintaining backup facilities for a critical server or data storage system in the event of disruption or damage from a voltage spike, as a financial management firm might have to do. In crafting a response, we must address the question of whether investments to improve power quality should be made for everyone, or whether they should be made "behind the meter" by those who value power quality highly? This issue would benefit from more independent empirical evidence and analysis.

[5] More specifically, nearly half of the overall benefits ($445 billion in net present value) for EPRI's (2011a) entire smart grid program are attributed to "reliability," which appears to be shorthand for reliability and power quality. There is another benefit category called "security" (benefits of $151 billion in net present value), which seems to be a subset of "reliability." Adding these gives the total of roughly $600 billion in the text.

Of more pressing concern are the new demands that may be placed on at least some distribution systems by distributed generation, primarily solar photovoltaic systems, and by the potential future need to recharge plug-in electric vehicle batteries. Several states are promoting solar photovoltaic technology with large subsidies (Borenstein 2008). Due to the wide variability in the output of photovoltaic technologies, sometimes over short time periods (NERC 2009, pp. 27–29), and related variability of net demand (which could be net *supply* if there are enough photovoltaic facilities and it is very sunny) that consumers place on the distribution system, photovoltaic technologies will place new stresses on local distribution feeders where they are installed. Better remote real-time monitoring and remote and automatic control capabilities, data acquisition and analysis of the state of the distribution system, and automatic breakers and switches will be required to accommodate significant quantities of these resources efficiently and safely.

The rate of new photovoltaic installations will vary widely from distribution feeder to distribution feeder and from state to state because of differences in subsidy policies and the relative economic attractiveness of photovoltaic investments. This variation suggests a targeted approach to local distribution system automation: focus first on areas where distributed generation, and the related stress on specific distribution feeders, will happen sooner.

The potential future demands placed on the local electric distribution system by plug-in electric vehicles raise similar issues. In 2010, out of 11.6 million total car sales, there were at most 3,000 pure electric plug-in electric vehicles sold and about 275,000 plug-in hybrids. The future path of electric vehicle sales depends on many factors: the price of gasoline, subsidies for electric vehicles, technological change affecting battery life and costs, new emissions standards, reductions in electric vehicle costs, and consumer behavior. Forecasts of the fraction of new vehicles that will be electric plug-ins by 2035 varies from less than 10 percent to over 80 percent. The U.S. Energy Information Administration (EIA 2011a, p.72; see also EPRI 2011b, chap. 4) forecasts a market share of light-duty vehicles of only 5 percent for plug-in and all-electric vehicles in 2035 in its reference case. The National Research Council (Committee on Assessment of Resource Needs for Fuel Cell and Hydrogen Technologies, 2010, p. 2) concludes that a realistic estimate is that by 2030 about 4.5 percent of the national light duty vehicle fleet will be plug-in electrics—with a maximum possibility of about 13 percent.

Along with the number of plug-in cars, the load placed on the distribution system will depend on the attributes of the car batteries and charge-up time selected by vehicle owners. Shorter charging times at higher voltages can place very significant loads on local distribution networks even with modest electric vehicle penetration (Browermaster 2011). An interesting possibility arises here. The demand on portions of the *local distribution system* in areas where electric vehicle sales may be concentrated—say, in places like Berkeley, California or Cambridge, Massachusetts—could peak at night, even when prices in the much broader wholesale power market are low. The possibility that the marginal cost of electricity distribution service on some distribution feeders could peak at times when aggregate system demand for power

and associated power prices in the wholesale market are low suggests that more thought should be given to altering the pricing of electricity distribution service. Distribution service prices are now based on a flat per kilowatt hour rate and do not vary with the marginal costs of distribution service, which is driven largely by the peak demand on distribution feeders. Distribution service prices would more closely reflect the associated marginal cost of this service if we moved, instead, to a distribution charge based at least partially on individual customers' peak load on the local distribution system.

These considerations all lead to the conclusion that phasing in the installation of smart grid technology, targeting investment where it is likely to be most needed quickly, and collecting data using a controlled experimental framework to evaluate its costs and benefits would be a sensible strategy.

I offer one caveat to this conclusion. Many U.S. distribution systems are aging, and utilities are embarking on large distribution network replacement programs. These investments are long-lived, and it makes sense for these programs to take advantage of the most economical modern distribution technologies. In many cases, forward-looking investment optimally should deploy much more automation and communication technologies than immediately needed, even if deployment of distributed generation and electric vehicles is expected to be slow.

Smart Meters and Dynamic Pricing Incentives

It is not unusual for the incremental generating capacity held by a utility or RTO to meet the peak demand on its system during the 100 highest demand hours each year (1.1 percent of the total hours) to account for 10 to 15 percent of the generating capacity on its system. This is the direct consequence of the wide variability of demand (especially in response to extreme weather conditions), the failure of retail prices to reflect the true marginal cost of supply under these extreme conditions, and the utility practice of building enough generating capacity to meet demand even under very extreme low-probability demand states. Accordingly, using appropriate prices to provide consumers with an incentive to cut peak demand during a small number of hours can reduce generating costs significantly in the long run. Retail prices that are not tied to variations in wholesale prices inefficiently increase the level of peak demand by underpricing it, and may also discourage increased demand during off-peak hours by overpricing it.

The idea of moving from time-invariant electricity prices to "peak-load" pricing where prices are more closely tied to variations in marginal cost has been around for at least 50 years (Boiteux 1964b, c; Turvey 1968; Steiner 1957; Kahn 1970, pp. 63–123). The application of the theory in practice has lagged far behind, especially in the United States. (Mitchel, Manning, and Acton (1978) discuss early developments in other countries.) There is evidence from the well-designed experiments with time-of-use pricing of electricity in the 1970s showing consumers do respond more or less as expected to price incentives (Aigner 1985), in the sense that

higher prices lead to less consumption. The magnitude of the estimated responses varies widely though and does not reflect potential adaptations in the attributes of the appliance stock resulting from widespread deployment. Moreover, a 2010 survey indicated that only about 1 percent of residential customers were on time-of-use rates (FERC 2011, pp. 28, 99).

"Smart meters" can record *real-time* consumption of electricity. They also can have two-way communications capabilities allowing for real-time retail prices tied to variations in wholesale prices (and a number of variations on this theme), and could lead to remote control of customer demand by allowing the retail supplier or the customer to adjust appliance utilization inside the customer's home or business. For example, a customer might program the air conditioning (or have the utility program it) to turn off, at least intermittently, when electricity prices reach a certain level. Real-time pricing to reflect variations in wholesale market prices can increase efficiency, at least when it is applied to larger customers (Borenstein 2005). In addition, real-time pricing can stimulate innovations in appliances and equipment inside the home or in business to make them capable of responding more easily to changes in real-time prices and load management arrangements made with retail suppliers.

The traditional arguments for not introducing real-time pricing were: 1) the meters and billing costs would be so costly that residential and small commercial customers would not benefit from them; 2) retail consumers would not understand or effectively utilize complex rate designs; and 3) changing rate designs would lead to large redistributions of income reflecting the wide variations in consumption patterns across individuals and decades-old mechanisms for allocating costs among types of customers and within customer classes (Borenstein 2007a, b).

At least some of these arguments are increasingly being questioned, and solutions being contemplated. Metering and communications technology have moved forward with more capabilities and lower costs. Today's more advanced smart meters use two-way communication and capabilities for active real-time interaction between the distribution system and the customer: they can record consumption at least once each hour, can be turned on and off remotely, can support the introduction of dynamic retail prices that are closely tied to dynamic wholesale market prices, and can control the utilization of appliances remotely in a way that facilitates active demand-side management of the electrical grid. In addition, the information available through smart meters can inform the distribution company about variations in demand and equipment outages on the distribution grid instantly, thus creating synergies between "smart meters" and smart distribution grid investments. Variations on full real-time pricing, in particular "critical peak pricing," are easier for consumers to understand and provide much better incentives than flat rates. Nevertheless, relatively few advanced "smart meters" had been installed and used effectively in the United States, although the number is now increasing at a rapid rate (U.S. Energy Information Administration 2011b) as a result of federal subsidies and state mandates. As many as 8.7 million smart meters have now been installed at residential and small commercial locations, about 6 percent of the total—though

the definition of what counts as a smart meter varies and experience with them is limited (St. John 2009).

Analysis of the costs and benefits of large-scale deployment of smart meters must look at both changes in consumer surplus and changes in the cost of supplying electricity and metering its consumption. On the demand side, one needs to be able to measure the demand elasticities and cross-elasticities for a very diverse population of consumers who have different appliance stocks, live in homes of widely varying sizes, experience different weather conditions, face different levels and structures of incumbent electricity tariffs, have different incomes, and consume different quantities of electricity each month. An added complexity is that it would be implausible to measure *long-run* demand elasticities taking the current attributes of appliances and equipment as given, because the appliance stock and opportunities to adjust energy use will change to take advantage of the new incentives if smart meters and dynamic pricing are widely used.

On the supply side, there are questions about how much all of this whizzy smart grid technology will cost and, as always, there is the need to measure the effects on generation, distribution, and transmission costs. Measuring incremental metering costs is not easy. The many different vendors of smart meters sell meters with different functionalities and different communications methods. Moreover, buying and installing the meters is only part of the cost: communications systems must be built to integrate smart meter information with automated distribution network capabilities; a new information technology infrastructure for data acquisition, analysis, and billing created and installed; customer service personnel retrained to respond to questions about more complex rate structures; and investments made in complementary distribution system upgrades. On the other hand, smart meters should also save operating costs, primarily by reducing meter-reading costs (especially for systems that have not already installed the first-generation one-way communication meters). We also know that as a theoretical matter, setting retail prices to reflect marginal supply costs will increase overall efficiency with which electricity is consumed and supplied. But is this efficiency gain large enough to cover the additional costs of smart meters and associated information and automated distribution technology, both in the aggregate and for customers with different utilization characteristics? I do not think that this question has yet been answered satisfactorily or the public adequately convinced that the answer is likely to be affirmative.

A large number of U.S. utilities began offering time-of-use and interruptible pricing options for large commercial and industrial customers during the 1980s, either as a pilot program or as an option (for example, Barbose, Goldman, Bharvirkar, Hopper, Ting, and Neenan 2005). More recently, a number of states have introduced pilot programs for residential (household) consumers that install smart meters of various kinds, charge prices that vary with wholesale prices, and observe demand. For example, Taylor, Schwarz, and Cochell (2005) estimate hourly own- and cross-price elasticities for industrial customers on Duke Power's optional real-time rates and find large net benefits for these customers. Faruqui

and Sergici (2010) summarize the results of 15 earlier studies of various forms of dynamic pricing, including time-of-use pricing, peak pricing, and real-time pricing. Faruqui and Sergici (2011) analyze the results of a dynamic pricing study performed by Baltimore Gas & Electric using treatment and control groups drawn from a representative group of households. Wolak (2006) analyzes a peak pricing experiment in Anaheim, California. Wolak (2010) analyzes a pilot program using peak pricing in Washington, D.C. Allcott (2011) analyzes data from the Chicago Energy Smart Pricing Plan that began operating in 2003. Faruqui (2011) summarizes the reduction in peak load from 109 dynamic pricing studies, including those that use time-of-use pricing, peak pricing, and full real-time pricing, and finds that higher peak period prices always lead to a reduction in peak demand. However, the reported price responses across these studies vary by an order of magnitude, and the factors that lead to the variability of responses have been subject to very limited analysis.

Before discussing what conclusions can be drawn from this evidence, a few warnings seem appropriate. First, there is wide variation in the design of the pilot/experimental studies and the variation in prices included in them. Just looking at the magnitude of the response without more information is not adequately informative. Second, essentially all of these studies include only "volunteers," which raises the possibility that those who choose to participate may be unusually sensitive to price variation. Third, many of these pilots include a very small number of participants, and in at least one study a large fraction of those who started in the pilot dropped out before it was completed. Fourth, few of the pilot programs use full real-time pricing. A few use "critical peak pricing" mechanisms, and this may yield results similar to what we would get with full real-time pricing. For example, PG&E's voluntary tariff for customers with smart meters starts with the regular tariff price, except during "Smart High Price Periods," which are communicated to the customer in advance by telephone, Internet posting, or text messaging, and the price rises to 60 cents per kilowatt hour between 2 p.m. and 7 p.m. for a maximum of 15 days per summer season. Fifth, several of the pilots apply only one price to the treatment group, which makes it impossible to trace out the relevant demand functions without making very strong assumptions about the shape of the demand curves. Using several treatment groups requires a larger pilot study than has often been the case (see Aigner, 1985, regarding the need for multiple treatment groups). Finally, few of these studies use data on consumer responses along with electricity supply and metering cost data to perform a proper cost–benefit analysis. Indeed, I have not yet seen a recent study as well designed, or with a welfare analysis as carefully performed, as the Los Angeles experiments managed and analyzed by RAND during the 1970s (Mitchell and Acton 1980).

Despite these concerns, the available evidence does suggest a number of conclusions: First, consumers do respond to higher peak prices by reducing peak demand. Second, dynamic pricing with very high prices during critical periods generally leads to much larger prices responses than traditional time-of-use pricing with predetermined time periods and prices, which typically use

much smaller price differences. Third, wide variation in price responsiveness is observed across studies, suggesting wide underlying variation in the attributes of households and pilot study conditions. Fourth, most if not all of the price response to higher peak period prices is to reduce peak demand rather than to shift from peak to off-peak demand. For example, a common reaction is to use less lighting, air conditioning, and refrigeration when prices are high—with no offsetting increase in electricity used at other times. However, the diffusion of plug-in vehicles or other technologies where time-of-use is a more important choice variable could yield very different results.[6] Fifth, technologies and information that make it easier for consumers to respond to high price signals lead to larger responses to any given price increase, although many of the reported results do not contain adequate information to estimate demand functions or to perform proper cost–benefit analyses.

Faruqui and Wood (2011) present a well-conceived "template" for what items should be included in a comprehensive cost–benefit analysis and present simulations for four "prototype" utilities. The aggregate benefit/cost ratios vary from 1.4 to 1.9 for the four simulations. The simulations are not based on real utilities nor complete data, but the hypothetical numbers are not unreasonable and the results are suggestive. Of course, cost–benefit analysis of universal smart meter installation and real-time pricing may also find that while the benefit/cost ratio is greater than 1.0 in the aggregate, it may not be beneficial to some significant number of individual customers. Borenstein (2007b) takes the wide variation in customer utilization attributes seriously, although his focus is on larger commercial and industrial customers, not residential customers. But the heterogeneity of the effects of smart meters and real-time pricing on residential and small commercial customers is an important issue that still needs to be addressed.

Some states that have mandated the installation of smart meters for all customers have found the decision to be controversial (for newspaper accounts, see Smith, 2009; Turkel, 2011a, b; Fehrenbacher, 2010; Baker, 2010). Some consumers have reacted negatively to increases in up-front distribution costs to pay for the smart grid enhancements. First, some customers with "unfavorable" consumption patterns—weighted toward times when prices are high—may see higher bills, rather than the lower bills they are being promised, compared to their billing under flat rates (Borenstein 2007b). Second, some smart meters have been deployed too quickly and have not worked properly. Third, with all of the data that these meters can collect, privacy advocates have raised concerns about

[6] "Storage space heating" allows off-peak electricity to be used to heat up special bricks or slabs or water tanks, which are then used as a source of warmth during on-peak hours. When storage space heating was introduced in Europe during the 1960s, it was consciously designed to shift demand to off-peak periods. It did such a good job that the peak shifted from day to night in England and northern Germany and the regulated prices no longer reflected the patterns of demand and cost. Steiner (1957) and Kahn (1970) discuss this "shifting peak" case theoretically. More generally, we should be reminded that we should not take our eyes off of the long-run equilibrium which may look very different from the short-run equilibrium—especially after technological change.

what data will be made widely available and how it may be used and protected. Finally, some public utility commissions and some utilities have done a poor job educating their customers and have rolled out their smart meter installation program too quickly. There are lessons to be learned about deployment strategy from these experiences.

Conclusions

The existing electricity distribution system is very old in many areas, and investments to replace key components will have to accelerate just to maintain the reliability of the system. These replacement programs should be consistent with longer-term strategies for modernizing the distribution system. However, there is a lot of uncertainty about the size of costs and benefits, and these costs and benefits vary across distribution feeders as well as customers and regions. The rate and direction of future technological change on both sides of the meter is also uncertain. The transition to smart grid technology is going to take years, and there are sure to be notable bumps along the way.

Accordingly, it seems to me that a sensible deployment strategy is to combine a long-run plan for rolling out smart-grid investments with well-designed pilots and experiments. Using randomized trials of smart grid technology and pricing, with a robust set of treatments and the "rest of the distribution grid" as the control, would allow much more confidence in estimates of demand response, meter and grid costs, reliability and power quality benefits, and other key outcomes. For example, Faruqui's (2011b) report on the peak-period price responses for 109 pilot programs displays responses between 5 to 50 percent of peak demand. An order-of-magnitude difference in measured price responses is just not good enough to do convincing cost–benefit analyses, especially with the other issues noted above. In turn, the information that emerges from these studies could be used to make mid-course corrections in the deployment strategy. Given the large investments contemplated in smart meters and complementary investments, along with the diverse uncertainties that we now face, rushing to deploy a particular set of technologies as quickly as possible is in my view a mistake.

Despite these reservations, the country is on a path to creating smarter transmission and distribution grids. Exactly how far and how fast we go remains quite uncertain, especially as the federal subsidies enacted in 2009 for promoting the smart grid come to an end.

■ *The views expressed here are my own and do not reflect the views of the Alfred P. Sloan Foundation, MIT, Exelon Corporation, Transcanada Corporation, or any other organization with which I am affiliated. I am an outside director of Exelon Corporation and of Transcanada Corporation. My other affiliations are identified on my CV at ⟨http://econ-www.mit.edu/faculty/pjoskow/cv⟩.*

References

Aigner, Dennis. 1985. "The Residential Time-of-Use Pricing Experiments: What Have We Learned?" Chap. 1 in *Social Experimentation,* edited by Jerry A. Hausman and David A. Wise. University of Chicago Press.

Allcott, Hunt. 2011. "Rethinking Real-Time Electricity Pricing." *Resources and Energy Economics* 33(4): 820–42.

Baker, David R. 2010. "Some PG&E Customers Want Choice on Smart Meters." *San Francisco Chronicle,* August 16, page D-1. http://www.sfgate.com/cgi-bin/article.cgi?f=/c/a/2010/08/16/BUPK1EUELD.DTL.

Barbose, Galen, C. Goldman, R. Bharvirkar, N. Hopper, M. Ting, and B. Neenan. 2005. "Real Time Pricing as a Default or Optional Service for C&I Customers: A Comparative Analysis of Eight Case Studies." Working Paper LBNL-57661, Lawrence Berkeley National Laboratory.

Boiteux, Marcel. 1964a. "The Choice of Plant and Equipment for the Production of Electricity." In *Marginal Cost Pricing in Practice,* edited by J. R. Nelson, 199–214. Englewood, NJ: Prentice-Hall.

Boiteux, Marcel. 1964b. "Marginal Cost Pricing." In *Marginal Cost Pricing in Practice,* edited by J. R. Nelson, 51–58. Englewood, NJ: Prentice-Hall.

Boiteux, Marcel. 1964c. "Peak Load Pricing." *Journal of Business* 33(2): 157–79.

Borenstein, Severin. 2002. "The Trouble with Electricity Markets: Understanding California's Restructuring Disaster." *Journal of Economic Perspectives* 16(1): 191–211.

Borenstein, Severin. 2005. "The Long-Run Efficiency of Real-Time Electricity Pricing." *Energy Journal* 26(3): 93–116.

Borenstein, Severin. 2007a. "Customer Risk from Real-Time Retail Electricity Pricing: Bill Volatility and Hedgeability." *Energy Journal* 28(2): 111–30.

Borenstein, Severin. 2007b. "Wealth Transfers among Large Customers from Implementing Real-Time Retail Electricity Pricing." *Energy Journal* 28(2): 131–49.

Borenstein, Severin, and Stephen Holland. 2005. "On the Efficiency of Competitive Markets with Time-Invariant Retail Prices." *RAND Journal of Economics* 36(3): 469–93.

Browermaster, Dan. 2011. "Plug-in Vehicles and Their Impact." PowerPoint presentation. http://www.cacities.org/resource_files/29491.PGEPEVIntro(2011-01-18).pdf.

Clayton, Paul R. 2001. *Fundamentals of Electric Circuit Analysis.* Hoboken, NJ: John Wiley & Sons.

Committee on Assessment of Resource Needs for Fuel Cell and Hydrogen Technologies, National Research Council. 2010. *Transitions to Alternative Transportation Technologies—Plug-In Hybrid Electric Vehicles.* National Academies Press, Washington, D.C.

Electric Power Research Institute (EPRI). 2011a. "Estimating the Costs and Benefits of the Smart Grid: A Preliminary Estimate of the Investment Requirements and Resultant Benefits of a Fully Functioning Smart Grid." Technical Report 1022519, Electric Power Research Institute, Palo Alto, CA.

Electric Power Research Institute (EPRI). 2011b. "Transportation Electrification: A Technology Overview." Technical Report 1021334, Electric Power Research Institute, Palo Alto, CA.

Faruqui, Ahmad. 2011a. "Dynamic Pricing: The Top 10 Myths." The Brattle Group. April 7. http://www.brattle.com/_documents/UploadLibrary/Upload936.pdf.

Faruqui, Ahmad. 2011b. "109 and Counting: An Update on Dynamic Pricing Experiments." The Brattle Group, PowerPoint Presentation. July 12, 2011. Provided by the author.

Faruqui, Ahmed, and Sanem Sergici. 2010. "Household Response to Dynamic Pricing of Electricity: A Survey of 15 Experiments." *Journal of Regulatory Economics* 38(2): 193–225.

Faruqui, Ahmed, and Sanem Sergici. 2011. "Evaluation of Baltimore Gas and Electric's Smart Energy Pricing Plan." PowerPoint presentation, The Brattle Group. http://www.brattle.com/_documents/UploadLibrary/Upload940.pdf.

Faruqui, Ahmed, and Lisa Wood. 2011. "The Benefits of Smart Meters." PowerPoint presentation, Institute for Electric Efficiency.

Fehrenbacher, Katie. 2010. "PG&E's Smart Meter Report: A Case Study of Infrastructure over Customer." Gigaom, May 12. http://gigaom.com/cleantech/pges-smart-meter-report-a-case-study-of-infrastructure-over-customer/.

Hogan, William. 1992. "Contract Networks for Electric Power Transmission." *Journal of Regulatory Economics* 4(3): 211–42.

Joskow, Paul. 2001. "California's Electricity Crisis." *Oxford Review of Economic Policy* 17(3): 365–88.

Joskow, Paul. 2005. "Transmission Policy in the United States." *Utilities Policy* 13(2): 95–115.

Joskow, Paul. 2006. "Markets for Power in the United States: An Interim Assessment." *Energy Journal* 27(1): 1–36.

Joskow, Paul. 2011a. "Comparing the Costs of Intermittent and Dispatchable Generating Technologies." *American Economic Review* 101(3): 238–41.

Joskow, Paul. 2011b. "Comparing the Costs of Intermittent and Dispatchable Electricity Generating Technologies." MIT CEEPR Working Paper (revised). http://econ-www.mit.edu/files/6317.

Joskow, Paul, and Jean Tirole. 2000. "Transmission Rights and Market Power on Electric Power Networks." *RAND Journal of Economics* 31(3): 450–87.

Joskow, Paul, and Jean Tirole. 2005. "Merchant Transmission Investment." *Journal of Industrial Economics* 53(2): 233–64.

Joskow, Paul, and Jean Tirole. 2007. "Reliability and Competitive Electricity Markets." *RAND Journal of Economics* 38(1): 68–84.

Kahn, Alfred. 1970. *Economic Principles.* Vol. 1 of *The Economics of Regulation: Principles and Institutions.* New York: John Wiley & Sons.

Mitchell, Bridger, and Jan Acton. 1980. "The Effects of Time-of-Use Rates in the Los Angeles Electricity Study." RAND, N-1533-DWP/HF.

Mitchell, Bridger, Willard Manning, and Jan Acton. 1978. *Peak-load Pricing: European Lessons for U.S. Energy Policy.* Cambridge, MA: Ballinger.

North American Electric Reliability Corporation (NERC). 2009. *Accommodating High Levels of Variable Generation: Special Report.*

PJM. 2007. *2006 State of the Market Report.* March 8. http://www.pjm.com/documents/reports/state-of-market-reports/2006-state-of-market-reports.aspx.

Power Engineering Society. 2006. "IEEE Benchmarking 2005 Results." PowerPoint Presentation.

Smith, Rebecca. 2009. "Smart Meter, Dumb Idea." *Wall Street Journal,* April 27, p. R5. http://online.wsj.com/article/SB124050416142448555.html.

Steiner, Peter O. 1957. "Peak Loads and Efficient Pricing." *Quarterly Journal of Economics* 71(4): 585–610.

St. John, Jeff. 2009. "8.3M Smart Meters and Counting in U.S." GreenTechGrid. July 17. http://www.greentechmedia.com/articles/read/8.3m-smart-meters-and-counting-in-united-states/.

Stoft, Stephen. 2002. *Power System Economics.* IEEE Press.

Taylor, Thomas N., Peter M. Schwartz, and James E. Cochell. 2005. "24/7 Hourly Response to Electricity Real-Time Pricing with up to Eight Summers of Experience." *Journal of Regulatory Economics* 27(3): 235–62.

Turkel, Tux. 2011a. "CMP: Smart Meter Bills Come with Huge Costs." *Portland Press Herald,* April 5, 2011. http://www.pressherald.com/news/cmp-smart-meter-bills-come-with-huge-costs_2011-04-05.html.

Turkel, Tux. 2011b. "PUC Allows 'Smart Meter' Opt-outs." *Portland Press Herald,* May 18, 2011. http://www.pressherald.com/news/puc-allows-smart-meter-opt-outs_2011-05-18.html.

Turvey, Ralph. 1968. "Peak Load Pricing." *Journal of Political Economy* 76(1): 101–13.

U.S. Federal Energy Regulatory Commission (FERC). 2011. *2010 Assessment of Demand Response and Advanced Metering.* Staff Report, February. http://www.ferc.gov/legal/staff-reports/2010-dr-report.pdf.

U.S. Energy Information Administration (EIA). 2010. *Annual Energy Review 2009.* Report No. DOE/EIA-0384(2009). Washington, DC: EIA.

U.S. Energy Information Administration (EIA). 2011a. *Annual Energy Outlook 2011,* Report No. DOE/EIA-0383(2011). Washington, DC: EIA.

U.S. Energy Information Administration (EIA). 2011b. "Advanced Electric Meter Installations Rising in Homes and Businesses." March 15. http://www.eia.gov/todayinenergy/detail.cfm?id=510.

Watkiss, Dan. 2011. "Divided Court Disconnects DOE Transmission Corridors." Bracewell & Giuliani Energy Legal Blog. February 8. http://www.energylegalblog.com/archives/2011/02/08/3483.

Wolak, Frank. 2006. "Residential Customer Response to Real-Time Pricing: The Anaheim Critical-Peak Pricing Experiment." CSEM Working Paper 151, Center for the Study of Energy Markets, University of California Energy Institute, Berkeley. http://www.ucei.berkeley.edu/PDF/csemwp151.pdf.

Wolak, Frank. 2010. "An Experimental Comparison of Critical Peak and Hourly Pricing: The PowerCents DC Program." Preliminary draft, March.

JOBNAME: No Job Name PAGE: 1 SESS: 11 OUTPUT: Wed Mar 31 13:37:57 2010
/v2451/blackwell/journals/ecaf_v30_i2/02ecaf_2000

PROMOTING INNOVATION IN THE ELECTRICITY INDUSTRY[1]

L. Lynne Kiesling

Smart metering can bring significant benefits to the electricity markets by allowing customers to reduce demand or increase supply when generation capacity is temporarily scarce. To reap the full efficiency and environmental benefits of this technology, regulators must allow price volatility and free entry into the market. The efficiency gains are enormous as both demand and supply will be affected by both temporary and longer-lasting price changes. Experiments have shown the value of this approach.

Keywords: Utility regulation, electricity markets, smart metering.

Technology and institutional design

How does technological change affect the role of economic regulation in retail electricity markets? I will offer some observations and research from a US perspective that I believe generalise to other settings as well. I also hope to challenge some current thinking about the potential retail value propositions involving electricity, and the role of regulation in retail markets in the presence of technological change. In particular, digital communication technology has made rich, vibrant, customer-oriented innovation possible. Some examples of such consumer-focused innovation include appliances and devices that can respond autonomously to changes in electricity prices; electric vehicles that can be programmed to charge when electricity prices fall below a trigger price set by the owner and serve as an energy storage device as well as a vehicle; or home energy management systems that enable the homeowner to set appliances to power down if the electricity they are using would be bought on the wholesale market during an hour when renewable power was not available (the green–grey portfolio mix). The technologies to bring these products and services to retail customers exist: the regulatory institutions that enable it do not.

These topics fall under the general category of smart grid. The buzz around the idea of smart grid is palpable: old companies such as General Electric and new companies such as Google are changing their business models to incorporate more smart grid activities and products, entrepreneurs are exploring new products and services at unprecedented rates, and federal legislation supports smart grid investments and has directed $40 billion of taxpayer funding to smart grid investments.

Defining 'smart grid' requires thinking about the subject from two different directions simultaneously – its technologies and its functionalities. Technologically, a smart grid is a digital communication overlay and integration into the electric power network. This communication technology includes:

- digital switching networks;
- remote sensing and monitoring in wires and in transformers;
- fault detection in wires and in transformers;
- devices for automated fault repair; and
- intelligent end-use devices in homes, stores, office buildings, garages and factories.

These various smart grid technologies enable a variety of functionalities in the electric power network, such as:

- transactive co-ordination of the system (many of the following functionalities contribute to this co-ordination);
- distributed resource interconnection, including renewable generation and electric vehicles;
- the ability of a resource/agent to be either a producer or consumer of electricity, or both;

© 2010 The Author. Journal compilation © Institute of Economic Affairs 2010. Published by Blackwell Publishing, Oxford

JOBNAME: No Job Name PAGE: 2 SESS: 11 OUTPUT: Wed Mar 31 13:37:57 2010
/v2451/blackwell/journals/ecaf_v30_i2/02ecaf_2000

- demand response to dynamic pricing, and automation of that response;
- the ability of an agent to program end-use devices to respond autonomously to price signals; and
- distribution system automation by the wires company, leading to better service reliability.

The integration of these technologies into the electric power network will embed distributed intelligence in the physical systems that the network comprises. When I refer to the grid or the electric power network I am including distributed human agents (and their private knowledge–preferences–intelligence) in the definition of the network, not just the physical assets.

The potential ways that smart grid capabilities can create value are large, and they transcend the traditional utility-provided, generic electricity generation and delivery value proposition. By enabling better, and more decentralised, co-ordination of electricity supply and demand, smart grid functionalities contribute to the optimisation of resource use in the entire electricity system. This optimisation has both economic (cost reduction) and environmental (reduced resource use, reduced emissions) implications. Note, though, that the connection between distributed intelligence and decentralised co-ordination allows these economic and environmental benefits to align and converge. One example of this convergence is how dynamic retail pricing induces consumers to shift consumption away from expensive peak hours, which leads to a reduced need for expensive infrastructure investment that is built to meet peaks and then sits idle for 95% of the year: avoiding that investment saves costs and saves resources.

Paradigm shift

These dramatic and exciting technological changes are on the verge of creating a paradigm shift in this industry. Historically vertically integrated and regulated, the electric power industry in the USA was designed for centralised physical and economic control, both for operational reasons and for economic efficiency and equity reasons. Reliability and system balance have always been the paramount policy objective, and from an economic perspective, the principal illuminating regulation has been a concept of 'the public interest' used to control the exercise of market power. Under the public interest theory of regulation, creating monopoly electric utilities and regulating the rates of return they earned minimised the deadweight loss associated with monopoly, while at the same time eliminating the competition that drove market prices below long-run average cost under free entry. This argument is based on neoclassical natural monopoly theory, in which the cost structure of the firms in the industry relative to the size of the market would ultimately lead to having a single firm operate in that market.

These regulatory and legal concepts are generally static, and do not adapt well to unforeseen and changing conditions, including economic dynamism and technological change. Consumer-focused, end-use innovation creates the opportunity to access consumer intelligence at the edge of the electric power network, enabling distributed, individual agents to co-ordinate their plans and actions transactively. Through that process we achieve what is known in complexity science as emergent order through decentralised co-ordination. However, regulatory institutions that are designed for centralised control, and the economic interests of incumbents who profit from that state of affairs, present barriers to achieving this decentralised co-ordination.

Moreover, regulatory policy objectives in electricity are increasingly in tension. Reliability of service and of the physical network is paramount, and rightly so. Another high-priority objective for the past century has been affordability and predictability, in the form of low, stable retail prices that insure all retail customers against price volatility. Intended as a bulwark against monopoly, this policy also stifles both the communication of cost information to consumers and the communication of consumer preferences back all the way through the value chain to generators. Consequently, to the extent that consumer preferences and production costs vary from average benefit and average cost over time, these regulated, averaged prices lead to inefficient outcomes with distorted resource allocation.[2]

Over the past 20 years, though, the complexity of our policy objectives has increased as our interest has grown in reducing the environmental impacts of energy use, and in decreasing energy use and improving energy efficiency. Low, fixed (and still regulated in most of the USA) retail prices are in direct tension with consumer empowerment to reduce the environmental impacts of energy use. Fixed, averaged, regulated retail prices send no signal, and give no incentive, about the relative economic or environmental value of changing consumption at a particular time. The current context of increasing capital costs and supply constraints compounds that tension.

These observations prompt the question: are low, stable retail prices that do not reflect underlying cost fluctuations actually in the public interest, particularly in light of innovations in end-use digital technology? Is our historical policy focus on low, stable prices still serving the public interest?

A different way to frame this question is to note that technological innovation and institutional design are symbiotic, and that when regulatory institutions do not evolve along with technology they can become maladaptive and counterproductive to our broader social goals of economic efficiency and equality of opportunity. The interaction between technological change and retail regulatory and market design is substantial, and can either be positive or negative for consumers, for entrepreneurs, and for incumbents.

The GridWise Olympic Peninsula case study

The GridWise Olympic Peninsula project is a concrete research project that illustrates these ideas. The Olympic Peninsula is located in Washington, in the northwest corner of the USA, and residents in this area use electricity primarily for winter heating. Population growth in the area has created a constrained electricity distribution network, with expectations of further future constraints. The traditional regulated utility response to demand growth is, naturally, to increase capital investment in generation and wires capacity, because under

© 2010 The Author. Journal compilation © Institute of Economic Affairs 2010. Published by Blackwell Publishing, Oxford

JOBNAME: No Job Name PAGE: 3 SESS: 11 OUTPUT: Wed Mar 31 13:37:57 2010
/v2451/blackwell/journals/ecaf_v30_i2/02ecaf_2000

rate-of-return regulation the firm's profits are a function of its 'rate base', its capital stock. Two challenges to this traditional response presented themselves – population and demand are growing because people move to the Olympic Peninsula for its natural beauty, and do not want it despoiled by more generating and distribution capacity, and the physical terrain of the area would make such infrastructure construction extremely costly. Thus, the proposal came from researchers at the Pacific Northwest National Laboratory to look at digital end-use technology and retail pricing as ways to increase capacity utilisation with the existing network assets.[3]

Homeowners participating in this project received a transactive, price-responsive communicating thermostat and water heater, and they could choose from among three different retail contracts:

- *Fixed price:* the price per kilowatt-hour did not change over time, and reflected the insurance premium for the insurance that the consumer received against price volatility;
- *Time of use (TOU):* this peak/off-peak price structure included a higher price in peak hours (6am–9am and 6pm–9pm) and a lower price during all other hours; and
- *Real-time price (RTP):* the energy portion of the consumer's price fluctuated to reflect real-time changes in the wholesale electricity price (as approximated in this project by the Dow Jones Mid-Columbia wholesale price). This price could change as often as every five minutes.

They also had in-home visual displays that made changing settings easy – if you can programme a digital video recorder you can control your own electricity use according to your preferences using this technology. The experiment also included a control group with technology that did not participate in the retail choice treatment.

The most novel institutional design innovation in the project was the design of a double auction for the RTP contract customers. A double-auction market design has active buyers and sellers simultaneously submitting bids and offers, which creates an information-rich environment that converges to equilibrium more quickly than other market designs. In this double-auction retail market, the household's transactive appliances submitted bids, while suppliers submitted offers to supply. This market cleared every five minutes for a year, and is the first use of a retail, real-time double-auction market.

In brief, the results of this year-long project were customer savings, increased distribution capacity utilisation and high reliability. On average, customers saved 10% on their bills, peak demand fell by 15% and overall electricity consumption fell, resulting in multiple types of conservation.

The RTP customers saved the most money, and the TOU customers reduced their consumption the most. This project provides just one example of how retail choice and product differentiation can be welfare-enhancing for individual consumers and also achieve system-related policy objectives. Other test beds and projects in the USA are finding corroborating results, although none has yet implemented a real-time double-auction retail market design (but some are working towards it).

Note also that extrapolating the results of this project nationally would suggest substantial avoided investment over 20 years. This avoided investment would mean lower production costs (other things equal) and long-term environmental benefits due to reduced resources being used to meet electricity demand.

This project is a concrete example of my main points:

- The symbiosis of technology and institutions is crucial for achieving the greatest potential value for electricity consumers from technological change (transactive technology + portfolio of retail products).
- Technology and market design can combine to transform a complex system into a complex *adaptive* system, harnessing edge intelligence to create economic and environmental value by empowering residential retail consumers to control and manage their own electricity consumption decisions and use technology to automate their responses to dynamic prices.
- Technology and market design can lead to emergent order through decentralised co-ordination rather than the hierarchical physical and economic control on which this industry has relied for the past century.

Consumer-focused smart grid innovation

Digital communication capabilities in the network, including in end-use devices, provide increasingly feature-rich, mobile ways to create consumer awareness about electricity consumption, electricity expenditure and the environmental impact of that consumption. They also provide ways to change electricity consumption, either manually or automatically, in the home or remotely. For example, a home can have a home area network (HAN) that connects its appliances, its heating and cooling, its water heater, its laundry, its entertainment (stereo, TV, DVR, game console) and its lighting into one communication network, accessible either through a computer screen in the home or a web-based portal that can be accessed via a computer or a web-enabled mobile device. Through this communication interface, the customer's electricity retailer can communicate real-time information about the quantity of electricity consumed, the price the consumer is paying, and even the type of generation resources being used to generate the power being consumed. The retailer can also communicate price signals to the customer, and the customer can program the different devices in the HAN to change their settings in response to price changes – e.g. if the price increases from nine cents to 12 cents, reduce the temperature in the water heater by five degrees, and increase the thermostat air conditioner setting by five degrees. Moreover, the consumer can have remote web access to the HAN, and can change settings, monitor energy consumption and analyse data on the home's electricity consumption.

Say, for example, you are on the train to work and you get an SMS notification that due to unexpected weather, there will be a higher-than-normal electricity price between 9am and 10am. You may have already programmed your devices to respond to price signals, but what if the price is high enough that you want to change your settings? You can log in to your

© 2010 The Author. Journal compilation © Institute of Economic Affairs 2010. Published by Blackwell Publishing, Oxford

JOBNAME: No Job Name PAGE: 4 SESS: 11 OUTPUT: Wed Mar 31 13:37:57 2010
/v2451/blackwell/journals/ecaf_v30_i2/02ecaf_2000

HAN from your mobile device, or from your computer at work, and change the device settings in the home through the web portal.

Such functionality requires intelligent end-use devices, which are increasingly feasible and cost-effective as the costs of information technology fall. Intelligent devices include thermostat, water heater, television and so on that have digital communication capability. Intelligent devices can have their settings changed remotely and can be programmed to respond to data, including price signals.

Furthermore, if the home has distributed generation installed, such as solar photovoltaic rooftop panels, the customer can program the network to reduce electricity use once the home's consumption reaches the generation capacity of the solar resource, thereby reducing the use of energy overall and reducing the use of fossil-fuel-generated power if the marginal generation resource at that time is coal or natural gas (of course, with retail choice, the customer could choose a 100% renewable energy contract if desired, which would alleviate the green/grey mix consideration). These digital communication technologies enable new value creation, reduction in environmental impact, and decentralised co-ordination in the electricity industry precisely because they make more of the network, and more of the participants in the network, transactive.

One of the hallmarks of smart grid technology is how it enables and reduces the cost of two-way communication. In electricity as in other industries, digital communication technology makes it possible and easy to have two-way communication and to use that communication capability to automate individual actions. As we have seen throughout society, as communication technology has proliferated it makes it easier and cheaper to engage in transactions. The implications of this for the electric power network are a smart grid that is a rich transactional environment, a market platform, a network connecting producers and consumers who contract and negotiate their mutual exchange of value (product, service) for value (payment). *A smart grid is a transactive grid.*

Take a non-electricity example – personal banking. Two innovations have transformed personal banking: the ATM and the internet. The digital communication technology that the internet comprises enables us to do our banking online instead of going to a branch or doing bank-by-phone. The transactions in which we engage with our bank are thus easier, quicker and cheaper for us. Furthermore, we can use this technology to automate our actions, such as scheduling recurring bill payments, or establishing trigger rules by which we receive alerts about our account status or activity. Digital communication technology enables us to bank any time, from anywhere. The value creation due to this transactive capability has been enormous, and has largely been in the form of consumer surplus/benefit (with some increases in retail banking profits before the current recession).

Digital communication technology lowers transaction costs. Transaction costs reduce the extent to which private parties engage in mutually beneficial exchange, thus as the internet has grown and our communication capabilities have expanded, our transactiveness has also increased dramatically. Banking is just one example; others include online shopping and eBay.

The implications of this transactive capability have been enormous: it has reshaped markets, firms and consumer expectations. Markets are increasingly global and competitive and we can engage in transactions with people on the other side of the world. But the most important implication of transactiveness for the electric power industry is the effect on firms. High transaction costs provide one reason for vertical integration, and have contributed to the vertically-integrated firms that have been the producers in this industry for a century – where transaction costs are high, firms make their transactions internal instead of through markets. But the transaction cost reductions arising from digital communication technology shift the margin at which it is profitable to organise transactions within a firm instead of through markets. If it is now cheaper to transact in markets, then transactive activity should shift from within-firm to between-firm, and the boundaries of firms should change. Vertically-integrated firms exist for several reasons (not the least of which in electricity is historical path dependence), and transaction costs provide only one reason, but lower transaction costs will contribute quickly to the increasingly stand-alone capability of both the generation and the retail portions of the electricity supply chain.

From the consumer's perspective, the implications of a transactive smart grid are profound. Take the personal banking experience and the GridWise Olympic Peninsula project example, and imagine what that kind of transactive capability would be like with respect to our energy use. Online home energy management, remote access, the ability to automate our electricity consumption decisions, the array of new products and services that could make use of this transactive functionality. Large industrial and commercial consumers already have such capability, but as technology prices have fallen and entrepreneurs have developed new products and services, a transactive smart grid brings this functionality into the home, creating a large range of value potential for consumers, producers and entrepreneurs.

A smart grid is a transactive grid

The reverse is also true. If a communication-rich electric power network does not take advantage of this transactive capability, we forsake all of this value creation. We still get value from the engineering-related optimisation of power flows, of fault detection and repair, and of distribution automation. But the engineering-related possibilities of a smart grid are only the tip of the iceberg: they are tweaks and improvements on the physical management of a closed-loop system. The true, meaningful, resilient and long-lived value proposition in smart grid is in enabling the multitudes of diverse, distributed, heterogeneous agents in the electric power network to exchange with each other for mutual benefit; in enabling a neighbourhood to form a microgrid and exchange among themselves; in enabling me to make a choice of whether to pay my employer for allowing me to charge my plug-in vehicle, or whether to sell my employer some of my stored electricity in the battery of my plug-in vehicle; and in allowing consumers to choose dynamic pricing and empowering them to use the technology and the price signals to control and manage their

© 2010 The Author. Journal compilation © Institute of Economic Affairs 2010. Published by Blackwell Publishing, Oxford

JOBNAME: No Job Name PAGE: 5 SESS: 11 OUTPUT: Wed Mar 31 13:37:57 2010
/v2451/blackwell/journals/ecaf_v30_i2/02ecaf_2000

own electricity use. In other words, if a grid is not transactive, it is not a smart grid.

But what are the likely outcomes if we invest in such smart grid technology, but retain our historic method of regulating retail prices?

Smart technology, dumb pricing

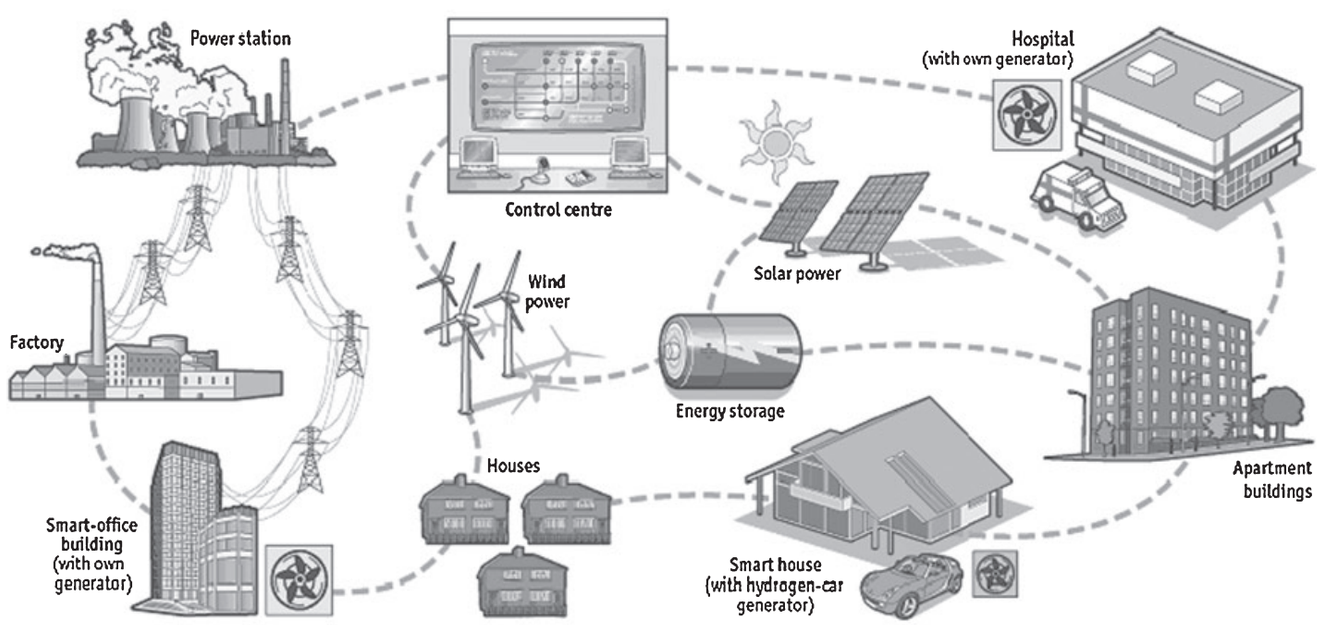
The GridWise Olympic Peninsula project and others suggest that smart technology and dumb pricing will nullify most of the potential consumer and system benefits of smart grid technology. Achieving the potential value creation from transactive end-use technology also requires enabling consumers to choose how much price volatility they are willing to accept, knowing that they have technology to manage their price responsiveness autonomously. At a minimum, transactive technologies require dynamic retail pricing if these innovations are to create value for consumers. Without dynamic pricing, the power system will fail to deliver efficiency and value to consumers. The 'one size fits all' of regulated and fixed retail rates is obsolete because of technological, institutional, regulatory and cultural changes that have created a diversity of products and services that the electricity industry can profitably sell to consumers. Dynamic pricing is necessary to maximise the value of technological innovation and other market reforms that characterise the most valuable, flexible and resilient power system; dynamic pricing also is, in and of itself, a valuable step in producing efficient and fair electricity markets.

Dynamic pricing, product differentiation and complementary technologies are the foundation of achieving decentralised co-ordination in the electric power industry. They bring timely information to consumers and enable them to participate in retail market processes; they also enable retailers to discover and satisfy the heterogeneous preferences of consumers, all of whom have private knowledge that is unavailable to firms and regulators in the absence of such market processes. Institutions that facilitate this discovery through dynamic pricing and technology are crucial for achieving decentralised co-ordination. Thus retail restructuring that allows dynamic pricing and product differentiation, that does not stifle the adoption of digital technology, and that reduces retail entry barriers is necessary if this value-creating decentralised co-ordination is to happen.

More generally, transactive end-use technologies make retail competition more feasible and beneficial. The inverse is also true: competitive retail electricity markets will be a platform for unleashing entrepreneurial creativity to enable consumers to get the most out of these technologies, and future ones that we cannot yet even imagine. The vibrant innovation in consumer electronics in the past 15 years illustrates this technology-institutions symbiosis that leads to consumer benefits and economic growth.

The present and future of the electricity system

These technological and institutional innovations would have implications beyond the individual consumer. Consider the electric power network as a system of interconnected individuals and physical assets. Historically, the technology, economics and business model in the industry has been a linear value chain with substantial homogeneity among the agents in the network, and with the delivery of value being uni-directional – the sale of generated electricity to end-use consumers. As Figure 1 suggests, technological



Sources: *The Economist*; ABB

Figure 1: The future smart grid value network

© 2010 The Author. Journal compilation © Institute of Economic Affairs 2010. Published by Blackwell Publishing, Oxford

JOBNAME: No Job Name PAGE: 6 SESS: 11 OUTPUT: Wed Mar 31 13:37:57 2010
/v2451/blackwell/journals/ecaf_v30_i2/02ecaf_2000

innovation is going to transform this traditional, linear value proposition.

Technological innovations in both supply and demand will transform the system into a value network, not a chain. The variety of resources and types of 'loads' increase, and now many agents have the technology that enables them to be a producer and consumer or a buyer and seller. These changes increase the complexity of the system. In such a complex system, adaptation to unknown and changing conditions is crucial for reliability and is also a way that consumers can benefit beyond the traditional benefits of generic electricity consumption. For example, home-level distributed generation and storage may improve the reliability of service for that homeowner, even if their normal consumption exceeds the capacity of their on-site generation. It also makes the network as a whole more resilient, because if an unanticipated shortage arises (such as an unplanned generator outage), a price signal would ripple through to the consumer indicating a higher price. This may induce a shift more use of the person's own generation or a reduction in demand by automatic changes in appliance settings. The individual could also sell electricity they generate to other consumers in the market. Similarly, if a homeowner has an electric vehicle, it is also an electricity storage device, which gives the owner the potential to be both a buyer and a seller of electricity. The home management system could be programmed and the vehicle used to sell stored electricity into the market, based on trigger prices that are set to reflect the owner's preferences.

The most effective and efficient system for enabling these adaptations is to use the technology's transactive capabilities to respond to price signals – in other words, enable these heterogeneous agents with their own private knowledge about their preferences and costs to co-ordinate using market processes.

Some policy implications

One of the largest challenges facing the electric industry and policy-makers today is the mismatch between regulatory and technological inertia within the industry and the vibrant, thriving economic and social dynamism that technological change has helped to create in nearly all other aspects of human society. Is it possible for regulatory institutions to be less subject to inertia and to be more adaptive to unknown and changing conditions such as innovation, changes in fuel costs and changes in environmental policy? Are regulators willing to enable the paradigm shift from control to co-ordination? Doing so would require regulators to shift their emphasis from retail rate structures and evaluations of cost recovery and towards retail choice, retail competition, market design and reducing transaction costs. It will also require reframing the idea of 'the public interest' away from its current narrow focus on low, stable retail prices to include environmental benefits and other more general concepts of consumer choice, product availability and consumer empowerment. Such an evolution of regulatory policy will mean incorporating more theories and practices from competition policy and an increased focus on information provision and consumer protection instead of retail rate determination.

One of the most effective institutional changes to enable decentralised co-ordination is to open retail electricity markets to competitive entry. Removing retail entry barriers and enabling retail competition would facilitate the promulgation of dynamic pricing options and product differentiation that could include green power and priority insurance, among other things. The decentralised and distributed network of retailers and customers can also contribute to grid resiliency and flexibility through more active participation in the market. Even if the magnitude of the shift in consumption is small, the effect may be large because of the non-linear relationship between peak load reductions and network reliability. A small load reduction at just the right time can keep the system from hitting capacity, can increase capacity utilisation and, in the long run, can reduce required investment in peak resources that sit idle much of the year. Letting go of some of the centralised economic and physical control would increase the resiliency of the system.

Bundling of retail electricity services with other services also has the potential to create value for consumers by making it easy and convenient for them to save money on their electricity bill in combination with other home services they consume, such as home security or home entertainment. For example, in regulated, vertically-integrated states in the USA a company such as ADT home security cannot enter the retail market and offer a bundled energy–home security service; in restructured US states (except for Texas) they have no profit driver to induce them to enter because of default service entry barriers. Reducing the regulatory barriers to enable us to learn if customers value such bundles would be a good thing, particularly for consumers and entrepreneurs.

By establishing preconditions for markets to function and creating an institutional environment in which they thrive, regulation will adapt to change because markets are complex adaptive systems that achieve ordered outcomes through decentralised co-ordination. By allowing markets to function, regulation will also benefit consumers by delivering differentiated products and services at different price points: note also that competition-facilitating regulation also enables entrepreneurial producers to profit from meeting the needs of consumers (who have diverse preferences and diffuse private knowledge). Market processes are positive-sum interactions in ways that traditional regulation cannot anticipate or duplicate.

Technology, automation, regulation and consumer empowerment

Economic growth and technological change have brought the electric industry and its regulation to a crossroads. Technological change from outside the industry has prompted changes in both regulatory institutions and business models, leading to the incremental disaggregation of the vertically-integrated firm in some regions of the USA and not in others. Simultaneously, increasing use of market transactions within this vertical value chain provides further strains on the existing institutional environment, both in the USA and in other countries as we see both incremental market

© 2010 The Author. Journal compilation © Institute of Economic Affairs 2010. Published by Blackwell Publishing, Oxford

JOBNAME: No Job Name PAGE: 7 SESS: 11 OUTPUT: Wed Mar 31 13:37:57 2010
/v2451/blackwell/journals/ecaf_v30_i2/02ecaf_2000

liberalisation in the USA and the EU and privatisation of the industry in many other countries.

The transactive technology now exists to empower consumers to control and manage their own energy use as they see fit and to automate their choices. These technologies reduce the transaction costs of responding to dynamic price signals. The broader consequence is that transactive technologies enable us to overcome the knowledge problem, by making that 'edge intelligence' in consumer preferences active. This use of market processes to co-ordinate the choices of diffuse private agents is the hallmark of economic dynamism and of efficiency in a complex adaptive system. However, technology alone cannot accomplish this as long as regulatory barriers exist that prevent consumers from choosing products and services that have dynamic pricing. Technology and institutions are symbiotic.

And on the social value of autonomous, transactive technology, I will give the last word to Alfred North Whitehead: 'Civilization advances by extending the number of important operations which we can perform without thinking of them.'

1. This article is an extended version of a Beesley lecture delivered on 1 October 2009, and draws extensively on L. Lynne Kiesling (2008).
2. Kiesling, ibid., Chapter 2.
3. This project was funded through the US Department of Energy's GridWise programme, and project partners included IBM, Whirlpool and the local distribution utilities. For further information see Hammerstrom *et al.* (2007) and Chassin and Kiesling (2008).

References

Chassin, D. and L. Kiesling (2008) 'Decentralized Coordination through Digital Technology, Dynamic Pricing, and Customer-driven Control: The GridWise Testbed Demonstration Project', *Electricity Journal*, 21 (October), 51–59.

Hammerstrom, D. *et al.* (2007) *Pacific Northwest GridWise Testbed Demonstration Projects. Part I: The Olympic Peninsula Project*, Report to the US Department of Energy, Technical Report No. PNNLSA-50009, Pacific Northwest National Laboratory.

Kiesling, L. (2008) *Deregulation, Innovation and Market Liberalization, Electricity Regulation in a Continually Evolving Environment*, London: Routledge.

L. Lynne Kiesling is Senior Lecturer at the Department of Economics and Kellogg School of Management, Northwestern University, Evanston, Illinois, USA (lkiesling@northwestern.edu).

© 2010 The Author. Journal compilation © Institute of Economic Affairs 2010. Published by Blackwell Publishing, Oxford

JOBNAME: No Job Name PAGE: 8 SESS: 11 OUTPUT: Wed Mar 31 13:37:57 2010
/v2451/blackwell/journals/ecaf_v30_i2/02ecaf_2000

SNP Best-set Typesetter Ltd.	
Journal Code: ECAF	Proofreader: Elsie
Article No: 2000	Delivery date: 30 March 2010
Page Extent: 7	

Decentralized Coordination through Digital Technology, Dynamic Pricing, and Customer-Driven Control: The GridWise Testbed Demonstration Project

David P. Chassin and Lynne Kiesling

Available online 17 October 2008.

The project highlights the idea that technology-enabled decentralized coordination can achieve the same, or better, economic and reliability benefits when compared to utility-focused centralized physical and economic control. Among the design's unique features was a retail double auction with five-minute market-clearing intervals that included residential customers as direct, active market participants.

Article Outline

I. Introduction
II. Structure of the Project
III. Results of the Project
IV. Conclusion
Vitae

I. Introduction

Current forecasts of both electricity demand and generation costs indicate that electricity prices are likely to rise over the next several years.[1] Meeting rising electricity demand using the traditional method – by building more generation, transmission, and distribution capacity – could cost up to $2000/kW,[2] and those costs are likely to increase due to rising construction costs.[3] Between rising costs and difficulties siting new plants and wires, the traditional approach is likely to be more costly and difficult to achieve than a new approach: using distributed digital end-use technology and dynamic pricing to enable consumers to control and manage their own electricity consumption.

Dynamic pricing can increase system reliability and capacity utilization, thereby delaying or avoiding infrastructure investments while saving consumers money in the process. Extensive research shows that residential customers can and do respond to dynamic pricing, and enabling end-use technologies increase the customers' ability to respond and the magnitude of the effects.[4] This article describes some of the results of the GridWise® Olympic Peninsula Testbed Demonstration, which looked specifically at the interaction of retail choice and enabling technologies. Rather than using traditional utility-centric centralized control to achieve reliable service, this demonstration employed a combination of smart technology and dynamic pricing to enable consumer-centric, decentralized coordination that achieved enhanced reliability, increased capacity utilization, and higher customer satisfaction.

The Olympic Peninsula Project looked at how consumers, especially residential customers, responded to real-time energy pricing information, and how consumers adjusted their energy consumption based on price changes. The project gave participants a choice to purchase electric power service through a variety of contract types; these contracts ranged from a fixed-type contract similar to those that have prevailed for residential customers over the past century, to real-time prices that could change every five minutes.

One important hypothesis tested was whether customers' responses to price changes were a function of the technology available to them. The project installed automated control technologies to allow industrial, municipal, and residential customers to automate their choice to change their energy consumption in response to price changes. For example, they could reduce their electricity consumption during times of peak demand or when prices were high and they had automated demand response. Residential customers also had smart appliances (including thermostats, water heaters, and clothes dryers) that could respond automatically to price signals. They could choose their balance between comfort and economy, and could customize and override their initial settings.

Extensive research shows that residential customers can and do respond to dynamic pricing.

This year-long project demonstrated that this combination of distributed communication and control technology and dynamic pricing can achieve a variety of beneficial outcomes. Consumers saved money on their energy bills, and the amount they saved varied depending on the type of contract they chose. On average, consumers saved 10 percent on their electricity bills.

The decentralized coordination made possible through distributed technology and dynamic pricing also reduced peak demand by approximately 15 percent on average over the year of the project. Moreover, the price-induced demand response combined with distributed generation to reduce peak demand consistently over a course of several days during the project, sometimes by as much as 50 percent. This approach cost less than alternative methods of providing peak generation and increasing distribution capacity, and we suggest the approach might further be used to mitigate intermittency of renewable resources.

Two of the most meaningful implications of these results relate to investment and the barriers to achieving these benefits through the extension of such technologies and pricing. First, the project demonstrated that this decentralized coordination can alleviate the need to build expensive new infrastructure to address constraints on the distribution or transmission system during times of peak demand. If all customers were engaged in reducing peak demand, as in this project, peak electricity prices would be substantially lower, and construction of about $70 billion (over 20 years) of new generation, transmission, and distribution system infrastructure could be avoided, with those savings passed along to consumers.[5] Not only does that reduced investment requirement reduce long-term costs to consumers relative to what they would have been otherwise, it also reduces overall resource use and idle capacity, which has positive long-term environmental benefits. Second, regulatory institutions matter. The traditional cost-based structure of regulation, and the reticence of regulators to enable consumers to choose dynamic pricing, are the most substantial barriers to the wide-scale adoption of these feasible technologies, which enable reliability and resilience through decentralized coordination.

II. Structure of the Project

The Olympic Peninsula Project was a demonstration project, led by the Pacific Northwest National Laboratory (PNNL), testing a mixed residential, commercial, and industrial power distribution utility network with highly distributed intelligence and market-based dynamic pricing. Washington's Olympic Peninsula is an area of great scenic beauty, with population centers concentrated on the northern edge. The peninsula's radial electricity distribution network is connected to the rest of the network through a single distribution substation. While the peninsula is experiencing economic growth and a resulting increase in electricity demand, the natural beauty of the area and other environmental concerns mean that the residents want to

explore options other than building generation capacity on the peninsula or building additional transmission capacity.

The traditional cost-based structure of regulation, and the reticence of regulators to enable consumers to choose dynamic pricing, are the most substantial barriers.

Consequently, the Bonneville Power Administration (BPA) initiated an effort to address the transmission constraint through a so-called non-wires solution, among others. Siting a testbed where a real need for alternative supply solutions is already apparent amplifies the prospect that any demonstrated benefits may be clearly recognized and rapidly adopted. These considerations provided a strong incentive for selecting the Olympic Peninsula's distribution system as a prime project site, where GridWise technologies could address a present need and be unambiguously demonstrated.

Thus this project tested the combination of enabling technologies and market-based dynamic pricing to investigate the effects of dynamic pricing and enabling technology on utilization of existing capacity, deferral of capital investment, and the ability of distributed demand-side and supply-side resources to create system reliability. Two issues were of primary interest in this project: (1) what dynamic pricing contracts are attractive to consumers, and how does enabling technology affect that choice? (2) to what extent will consumers choose to automate energy use decisions?

The project examined the impact of different pricing contracts and automation technologies on 116 broadband-enabled households with electric heating/cooling for the year April 2006 through March 2007. Of these, 112 remained in the project for the duration of the study. Each household received a two-way programmable communicating thermostat (PCT) with a visual user interface that allowed the consumer to program the thermostat for the home, and specifically to program it to respond to price signals if desired. Some households also received dryers equipped with a Grid Friendly™ appliance (GFA) controller chip developed at PNNL, which enabled the appliance to receive price signals and be programmed to respond automatically to those price signals. Consumers could control the sensitivity of the appliance through the PCT settings.

These households also participated in a market field experiment involving dynamic pricing. While they continued to purchase energy from their local utility at a fixed price, they also received a cash account with a pre-determined balance. The received the balance remaining each

quarter and the account was replenished based on their historical energy consumption. The energy use decisions they made would determine how much was deducted from their cash account, and they were able to keep any difference as profit. The worst a household could do was a zero balance, so they were no worse off than if they had not participated in the experiment. At any time, customers could log in to a secure Web site to see their current energy use and how effective their energy use strategies were.

Upon signing up for the project the households received extensive information and education about the technologies available to them and the kinds of energy use strategies made possible by these technologies. They were then asked to choose a retail pricing contract from three options: a fixed-price contract (based on BPA's price forecast), a time-of-use (TOU) contract with a variable critical-peak pricing (CPP) component that could be called in periods of tight capacity, or a real-time price (RTP) contract that would reflect a retail-level market-clearing price in five-minute intervals.

The RTP was determined using a uniform price double auction, in which buyers (households, commercial, and industrial) submit bids and sellers (wholesale and retail-level distributed generation) submit offers simultaneously. The digital technology in the household enabled residential customers to participate actively in such frequent markets because they could automate the bidding of their demand functions into the market. This project was the first in which a double auction retail market design was tested in electric power, and the use of a retail double auction with residential customers in the market is one of the unique features of this market design.

Households participated in a market field experiment involving dynamic pricing.

Using ranked contract preferences, the households were then divided fairly evenly among the three contracts types and a control group that received the enabling technologies and would have their energy use monitored, but did not participate in the dynamic pricing market experiment. All but 11 percent of households not placed in the control group received either their first or second choice (49 percent and 16 percent, respectively); interestingly, nearly 90 percent of the households ranked RTP as their first or second choice. This result counters the accepted wisdom that residential customers want only reliable service at low, stable prices, but may be enhanced by an early-adopter effect.

Of the 116 households, 30 were in the fixed-price contract, 30 were in the RTP contract, 31 were in the TOU contract, and 25 were in the control group that received the digital technology but did not participate in the market experiment.

The control group participants were not charged for their energy consumption. Fixed-price group participants were charged 8.100 ¢/kWh. The TOU participants were charged under two different rate structures depending on the season. During the period fall, winter, and spring season (Oct. 1–July 24) the off-peak (9:00 AM–5:59 PM and 9:00 PM–5:59 AM) price was 4.119 ¢/kWh and the on-peak (6:00 AM–8:59 AM and 6:00 PM–8:59 PM) price was 12.150 ¢/kWh. During the summer period (July 25–Sept. 30) the off-peak (9:00 PM–2:59 PM) price was 5.000 ¢/kWh and the on-peak (3:00 PM–8:59 PM) price was 13.500 ¢/kWh. A single CPP event was called Nov. 1 from 2:00 AM to 6:00 AM with a price of 35.000 ¢/kWh. The RTP participants were charged the price of energy as cleared every five minutes by a retail-level market.

The system was operated with different constraints on the feeder at different times of year. From April 1 to Sept. 22, the feeder capacity was set to 1,500 kW and the Mid-Columbia wholesale price of power reported by Dow Jones was bid at the level. From Sept. 22 to Dec. 8, the feeder capacity was reduced to 500 kW, and from Dec. 8 to March 31, it was increased to 750 kW. Figure 1 represents how the active RTP households and the DG resources could interact to determine the market-clearing price in five-minute intervals.



Full-size image (19K)

Figure 1. Representative Supply and Demand in Five-Minute RTP Market

View Within Article

III. Results of the Project

We focus on some of the most important economic results of this project: household energy consumption, prices paid, household savings, and changes in overall load duration. Table 1

presents the average hourly household energy consumption by contract group. The average household in the TOU contract group consumed the least electricity per hour (1.420 kW), followed by the average fixed-price customer (1.790 kW), the average RTP customer (2.100 kW), and finally the control group (2.116 kW).

Table 1.

Mean and Standard Deviation of Hourly Household Energy Use by Group.

Group	Mean (kW)	Standard Deviation	Number of Observations
Control	2.116	1.25	8759
Fixed price	1.790	0.84	8759
TOU	1.420	0.77	8759
RTP	2.100	1.00	8759

View Within Article

These consumption patterns differ statistically from each other (based on nonparametric Kruskal–Walls tests of medians across groups). Thus, we found that the type of dynamic pricing contract shapes individual behavior. Furthermore, note that the incentives inherent in different forms of pricing led to different average consumption beyond just having the technology, as was the case for the control group. This result suggests that simply the transparency and information provided by the technology does not necessarily reduce electricity consumption as effectively as the combination of the technology and the dynamic pricing with its embedded economic incentives.

The consumption data presented in Table 1 suggest that the TOU contract households consumed less energy than the other customers. After controlling for price response, weather effects, and weekend days, the TOU group's overall energy consumption was 20 percent lower than the fixed-price group's.[6] This result indicates that the TOU (with occasional critical peaks) pricing induced the greatest overall energy conservation and reduction in electricity use.

Table 2 reports the average hourly price per MWh by contract group. This was computed as a blended average by dividing the total energy consumed by the total payments made for each

contract group. In the case of the control group, this quantity could not be computed because they did not pay for energy used.

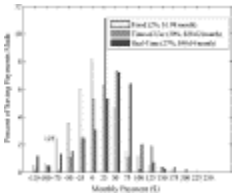
Table 2.

Mean and Standard Deviation of Hourly Average Price/MWh by Group (dollars).

Group	Mean ($/MWh)	Standard Deviation	Number of Observations
Control	n/a	n/a	n/a
Fixed price	81.000	0.000	8759
TOU	63.271	35.904	8759
RTP	49.198	47.462	8759

View Within Article

The low average price for those on the RTP contract indicates that the RTP customers used their automation and control capabilities to shift their use to less expensive times. The customer savings achieved corroborate this observation. Figure 2 shows average household savings by contract group.



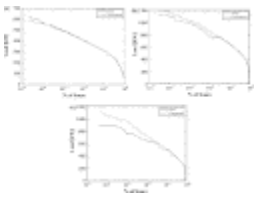
Full-size image (28K)

Figure 2. Monthly Savings Estimate by Contract Group

View Within Article

Participants in the fixed-price contract received about 2 percent savings relative to the control group; the TOU group saved 30 percent and the RTP group saved 27 percent. Note the difference in the distribution of the savings across the three groups. The RTP savings are skewed substantially to the right of the other two groups. This distribution of RTP savings indicates the

significantly greater savings earned by the RTP customers who selected the most economical appliance settings, relative to those who selected more comfort and did not earn such savings. In terms of peak demand reduction, the RTP group saw peak consumption decreases of between 15 and 17 percent relative to what the peak would have been in the absence of the dynamic pricing. We also compared the actual energy use to the counterfactual energy use, i.e., the amount that would have been consumed at the average price in that market period instead of the market-clearing price as determined by the double auction. Figure 3 shows the actual and the counterfactual load duration curves (graphed logarithmically) divided among three system condition categories: when the distribution feeder was unconstrained, moderately constrained, and severely constrained. In essence, a load duration curve shows the distribution of consumption over time; if consumption were distributed uniformly, the load duration curve would be a straight line, and capacity utilization or load factor would be the same at all times. Flattening the load duration curve, which indicates shifting some peak demand to non-peak hours, improves capacity utilization and reduces the need to invest in additional capacity, for a given level of demand. The peak load reduction of the RTP group is seen at the top left corner, where the actual curve is substantially below the counterfactual curve. Note Figure 3(c) in particular, which shows how extensively the RTP market and demand response automation reduced demand relative to the level of demand without the combination of the RTP market and the distributed residential automation technology. A 15 to 17 percent reduction is substantial, and is similar in magnitude to the reductions seen in other projects, such as the California Statewide Pricing Pilot.[7]



Full-size image (33K)

Figure 3. Actual and Counterfactual Load Duration Curves for (a) Unconstrained, (b) Moderately Constrained and (c) Very Constrained Systems

View Within Article

In addition to the reduction in peak demand in the RTP group, consumers also had a shifted load shape as a result of the dynamic pricing and the automated technologies that responded directly to market price signals. Figure 4 shows the actual and counterfactual thermostat loads for thermostatically controlled space conditioning of RTP contract homes during the most-constrained and least-constrained periods on the distribution feeder. Because all participant bids for RTP contracts were recorded when the market cleared every five minutes, and the bid price formula based on the thermostat status is reversible given the information gathered during the project, both the actual and the counterfactual energy could be computed for each market period.


Full-size image (44K)

Figure 4. Diurnal RTP Group Heating End-Use Load (a) Constrained Feeder, High Demand, (b) Unconstrained Feeder, Low Demand

View Within Article

The RTP induced an interesting shift in this automated consumption in both constrained and unconstrained feeder conditions. When demand was high and the feeder was constrained, the shift of demand from peak to off-peak was large, induced by the differential between peak and off-peak market-clearing prices. On unconstrained feeder days, however, the moderation of price volatility meant that the thermostats were sensitive to smaller diurnal price variations. While the transactive control strategy did not explicitly forecast future prices, the diurnal nature of the price movements themselves effectively induced opportunistic pre-heating or pre-cooling. The use of pre-heating/pre-cooling is generally viewed as an essential mechanism to mitigate the effect of load curtailment rebound phenomena. Effective pre-use strategies can be very difficult to engineer, and it is encouraging to see that market-based strategies are at least as effective as administered ones. Figure 4 shows the diurnal load duration curves for the RTP group during the period of constrained feeder (4a) and unconstrained feeder (4b).

Finally, the project's participants were very satisfied with the technology and the pricing with which they experimented during the project. Final project participant surveys indicate that 80 percent of participants were either very satisfied (51 percent) or somewhat satisfied (29 percent) with the end-use technology, and that 82 percent were either extremely likely (48 percent) or very likely (34 percent) to participate in a program like this one if it were offered again.[8]

IV. Conclusion

The GridWise Olympic Peninsula Testbed Demonstration is the most significant demonstration to date that dynamic pricing and digital end-use technology combine to empower consumers to manage their own electricity use and save money in a user-friendly way. Moreover, these individual choices in aggregate led both to immediate and sustained reductions in peak demand. One implication of these results is that dynamic pricing and enabling end-use technology can improve capacity utilization and forestall or eliminate the need to invest in costly and unpopular generation, transmission, and distribution assets.

A unique feature of the tested market design was a retail double auction with five-minute market-clearing intervals that included residential customers as direct, active market participants. The positive individual and system results suggest that further implementation of such a retail market design would benefit residential customers, including those who do not choose a RTP contract.

The Olympic Peninsula Project also demonstrated how new business models and regulatory approaches can overcome institutional barriers. Dynamic pricing and retail customer choice offer both individual and system benefits, but substantial barriers have precluded them from evolving for residential customers. Two of the largest social and cultural barriers are risk aversion on the part of utilities, who are heavily invested in their existing business models based on rate recovery through physical (generation, transmission, and distribution) assets, and risk aversion on the part of regulators, who have traditionally believed that the public interest and consumer protection are best served by fixed, regulated retail prices. This project has demonstrated how changing those beliefs, and the policies arising from those beliefs, can benefit consumers, utilities, and regulators.

In brief, this project highlights the idea that technology-enabled decentralized coordination can achieve the same, or better, economic and reliability benefits when compared to utility-focused

centralized physical and economic control. The dramatic transformation of digital technology in the past few decades has decreased transaction costs and increased the extent of feasible decentralized coordination in this industry, just as we have seen occur in so many other industries. Institutions, which structure and shape the contexts in which such processes occur, provide a means for creating this decentralized coordination. In particular, regulatory institutions affect whether or not this coordination can occur. Thus, effective regulation should not focus on allocation, but rather on decentralized coordination and how to bring it about. A focus on decentralized coordination means a focus on market processes, which are adaptive institutions that evolve along with technological change.

[1] Federal Energy Regulatory Commission, *Increasing Costs in Electric Markets*, June 19, 2008, at 2, available at http://www.ferc.gov/legal/staff-reports/06-19-08-cost-electric.pdf.

[2] D. Hammerstrom *et al. Pacific Northwest GridWiseTM Testbed Demonstration Projects, Part I: The Olympic Peninsula Project*, 2007, at 1.1, available at http://gridwise.pnl.gov/docs/op_project_final_report_pnnl17167.pdf.

[3] FERC, *supra* note 1, at 6.

[4] Lynne Kiesling, Deregulation, Innovation and Market Liberalization: Electricity Regulation in a Continually Evolving Environment (Routledge: 2008), Ch. 4.

[5] W.S. Baer *et al.*, *Estimating the Benefits of the GridWise Initiative*, RAND, Santa Monica, CA, May 2004, at 28.

[6] Kiesling, *supra* note 4, Ch. 4 Appendix.

[7] Charles River Associates, *Impact Evaluation of the California Statewide Pricing Pilot*, Oakland, CA, 2005.

[8] Hammerstrom *et al.*, *supra* note 2, at A.11.

Vitae

David P. Chassin is staff scientist with the Energy Science & Technology Directorate at Pacific Northwest National Laboratory where he has worked since 1992. He was Vice-President of Development for Image Systems Technology from 1987 to 1992, where he pioneered a hybrid raster/vector computer-aided design (CAD) technology called CAD Overlay. He led the development of building energy simulation and diagnostic systems, including Softdesk Energy

and DOE's Whole Building Diagnostician. His recent research focuses on emerging theories of complexity as they relate to high-performance large-scale simulation and modeling. He contributes to the Western Electricity Coordinating Council's (WECC) Load Modeling Task Force and is a member of the WECC's Market Integration Committee. He received his B.S. in Building Science from Rensselaer Polytechnic Institute in Troy, New York.

Lynne Kiesling is a Senior Lecturer in the Department of Economics at Northwestern University, and in the Social Enterprise at Kellogg (SEEK) program in the Kellogg School of Management at Northwestern University. At Northwestern she is also a Faculty Member in the Northwestern Institute on Complex Systems (NICO) and a Faculty Affiliate in the Center for the Study of Industrial Organization (CSIO). She is also currently a member of the GridWise Architecture Council, a group of 13 experts volunteering their time to articulate the guiding principles for an intelligent, transactive energy system of the future, and to guide and promote measures to transform the nation's electricity system into a more reliable, affordable, secure network in which users collaborate with suppliers in an information- and value-rich market environment.

Alfred Kahn As A Case Study of A Political Entrepreneur:

An Essay in Honor of His 90th Birthday

PHILIP J. WEISER *

Professor of Law and Telecommunications, University of Colorado

Abstract

The success of airline deregulation challenged the claims of public choice theory, which asserts that regulation serves the purposes of the regulated firms themselves. One prominent explanation for airline deregulation is that "political entrepreneurs" can, under certain circumstances, challenge the status quo. Nonetheless, commentators have failed to examine how Fred Kahn fits the model of a political entrepreneur. By examining Kahn's success as a political entrepreneur, this Essay highlights how commentators and policymakers can gain important insights into how to spur regulatory reform and oversee regulatory reform efforts like that necessary to modernize our system of spectrum management.

1 Introduction

On many accounts, Alfred Kahn is the seminal figure in regulation in the United States during the late 20th and early-21st century (McCraw, 1984). As a regulator, as an academic, and as an advocate, Fred Kahn has been preeminent in each of these roles, and shaped the way that institutions behave, academics study and, most importantly, regulators regulate. Not only did he write the book (or a two-volume treatise, as it were) on the *Economics of Regulation* (Kahn, 1970), he also wrote and executed the play book on how to conceive of and manage a major deregulatory initiative as a "political entrepreneur." In so doing, he underscored that ideas and effective public administration can make a difference—proving wrong those skeptics who doubted that regulatory reform could ever be accomplished over the wishes of vested interests who benefit from the status quo.[1]

Kahn's leadership as a political entrepreneur is a topic that commentators generally gloss over, offering only a sentence or two to note that Kahn fits the prototype (Peltzman et al., 1989). The concept of a political entrepreneur continues to be developed and is currently used to cover a range of actors, including consumer advocates like Ralph Nader (who famously pushed for auto safety regulations), politicians like Ed Muskie (who spearheaded the drive for environmental protection), and, in a few cases, regulators like

* Contact Author. University of Colorado Law School. 404 Wolf Law Building, 401 UCB, Boulder, CO 80309-0401. E-mail: phil.weiser@colorado.edu

[1] Notably, I do not take a position on the costs or benefits of airline deregulation, but rather, I highlight that the very success of the initiative is worth studying as a model for effectuating regulatory reform.

Electronic copy available at: http://ssrn.com/abstract=1263148

Fred Kahn (Pozen, 2008).[2] In most cases, discussions of political entrepreneurship tend to be more abstract and not grounded in the practical experience of policymakers like Fred Kahn.

This Essay emphasizes the importance of examining case studies in political entrepreneurship. In particular, as to Fred Kahn, a case study of his leadership underscores the importance of developing a public case for regulatory reform, stretching the legal authority of a regulatory agency to pursue reform objectives, and seeking to empower entities that would benefit from reform. Although this Essay does not seek to undertake an exhaustive study of how Kahn embodied these qualities in his leadership of the Civil Aeronautics Board (CAB), Part 2 explains the political entrepreneur concept and highlights how Kahn's leadership at the CAB provides an exemplary case of political entrepreneurship. Building on this explanation, Part 3 suggests some relevant strategies for political entrepreneurship in the area of spectrum policy reform. Finally, Part 4 offers a short conclusion, calling for a greater investigation of Kahn's model of political entrepreneurship and an increased focus on the role of such leadership in advancing the cause of regulatory reform.

2 The Path of Airline Deregulation and Kahn's Political Entrepreneurship

The concept of a political entrepreneur takes, as a starting point, the premise that individual and collective initiative can influence policy outcomes. On more economically determined views of history, such individual initiative is irrelevant. For strict adherents to public choice theory, for example, it is an article of faith that powerful rent-seeking forces use regulation to protect concentrated interests over diffuse ones, thereby making swimming against that tide both foolhardy and irrational. But such a public choice perspective, as the last thirty years of deregulation have shown, is over-determined and needs to be tempered by, among other things, the political entrepreneur concept.

Scholars are still developing the concept of "political entrepreneurship." Over the last three decades, scholars have discussed the concept or variants of it (such as "policy entrepreneurs"), but it remains an under-developed concept.[3] On many accounts, James Q. Wilson (1975) often receives the credit for the concept, developing it as part of a four part typology (Wilson, 1980).[4] Notably, Wilson recognized the possibility of "entrepreneurial politics" whereby a political entrepreneur, such as Ralph Nader, can overcome public choice forces and mobilize diffuse interests to support regulation over the opposition of

[2] For an even broader conception of the term, see Kingdon (1995, p.204), where he explains that "[political] entrepreneurs are found at many locations; they might be elected officials, career civil servants, lobbyists, academics, or journalists. No one type of participant dominates the pool of entrepreneurs."

[3] As Pozen (2008, p.301) put it, "the phrase 'policy entrepreneur' is [often] invoked without any explanation or references, presumably on the belief that the concept is either self-explanatory or so well known as to make citation pedantic. Not infrequently, authors will use 'policy' interchangeably with 'public,' 'political,' 'bureaucratic,' 'administrative,' or some other variant." Michael Mintrom (2000, p.ix) related, in a similar vein, that although they "have been mentioned in the literature on policymaking, policy entrepreneurs remain ghost-like figures, in need of clearer definition."

[4] Peter Schuck (1981) credits Wilson with developing the concept, but others developed it around the same time (see Walker, 1974).

Electronic copy available at: http://ssrn.com/abstract=1263148

concentrated interests (i.e., the regulated sector).[5] In short, a quintessential political entrepreneur, like Fred Kahn at the CAB, demonstrates that ideas matter in by politics by implementing an effective political strategy to spur policy reform.[6]

For my purposes, there are three essential elements—all embodied by Kahn's leadership at the CAB—that define the role of a political entrepreneur in spurring policy reform. As a starting point, the critical strategic goal of any political entrepreneur is to develop a sufficient level of public awareness of and support for a policy objective. In Kahn's case, his ability to make the case for reform publicly and with intellectual credibility was essential to his success—not unlike the ability of an actual inventor to make the case for commercializing her invention. In tandem with the ability to make the case for deregulation to the public, Kahn also pursued the objective of eroding the airline industry's commitment to the legacy regulatory regime by both undermining the manner in which it protected established incumbents and bolstering the strength of those interests that would benefit from deregulation. Roughly analogous to the role of market entrepreneurs, the challenge of articulating the case for reform parallels the need to make the case for funding a business venture whereas the second two objectives present execution challenges that, if managed effectively, demonstrate that the venture can and should succeed.

Kahn's success in pursuing these three goals resulted from a unique set of skills and abilities that he brought to the effort. First, he understood and was an evangelist for airline deregulation. Like a private sector entrepreneur, a political entrepreneur need not be the inventor of a policy strategy, but it is essential that she fully understands it and can credibly explain it. Second, Kahn's level of commitment was singular and determined. In this respect, he understood and embodied Max Weber's (1946, p. 128) description of politics as the "strong and slow boring of hard boards. It takes both passion and perspective."

In 1977, when Kahn was appointed Chairman of the Civil Aeronautics Board (CAB), the incumbent airlines constituted a cartel that the CAB, in effect, protected by insisting on uniform pricing (thereby preventing firms from detecting and engaging in price-cutting) and preventing new entry (thereby preventing them from maverick firms). By pursuing a deregulatory initiative against the wishes of the established firms, Kahn challenged the premises of public choice theory, which predicted that any such initiative was doomed to fail. According to public choice theory and, in particular, the allied concept of capture theory, concentrated interests (like the established airline companies) benefit from and dictate the course of regulation to protect their economic rents at the expense of diffuse interests (namely, average consumers).[7]

---

[5] Wilson's (1980) typology provides for three other categories of politics. On his view, public choice politics can be divided into two parts—interest group politics, whereby narrowly dispersed costs and benefits lead to bargaining among affected interests; and client politics, whereby the benefits are concentrated and costs are dispersed. Wilson also highlights that some laws, such as the Social Security Act and the Sherman Act, apply across a wide array of affected parties and are the result of majoritarian politics.

[6] Other commentators have offered a number of different definitions of the concept. Russell Hardin (1982, p.35), for example, explains that "political entrepreneurs are people who, for their own career reasons, find it in their private interest to work to provide collective benefits to relevant groups." Sam Peltzman (1989, p.51), by contrast, focuses on the function of political entrepreneurs, defining them as persons who discover "how to take advantage of the fundamental instability of majority rule within the constraints imposed by the institutional arrangements designed to induce stability."

[7] See Levine (2006, p.270), who related that "airline deregulation was a policy that principally benefited poorly organized consumers and was adopted over the opposition of a relatively small and very well organized group of regulated airlines."

By late 1970s, public choice theory had discredited the public interest theory of regulation. Its popularity undoubtedly reflected the views among policy observers that it accurately described the ways in which many regulatory agencies operated (Schuck, 1981). Reflecting this perspective, Migue (1977, p.213) suggested that "[i]t seems fair to say that among economists the most widely accepted theory of government regulation is that, as a rule, regulation is acquired by the industry regulated and is designed and operated primarily for its benefit." George Stigler (1971, p 5), who was a pioneer in public choice theory, stated similarly that "regulation is acquired by the industry and is designed and operated primarily for its benefit." By successfully overseeing the deregulation of the airlines, Kahn's leadership challenged the premises and predictive power of this theory.[8] Indeed, as Peter Schuck (1981) put it, the success of airline deregulation undermined the stronger versions of public choice theory such that no refinements could rehabilitate it.

In recounting the events that led to the deregulation of the airline industry, it is clear that a number of forces and actors came together to spur the enactment of the Airline Deregulation Act of 1978 (McCraw, 1984). On all accounts, however, Fred Kahn plays a starring role (McCraw, 1984). For future would-be political entrepreneurs, Kahn's leadership strategies and tactics thus bear close analysis.

The appointment of a world class economist to the CAB defied the tradition of appointing individuals from political backgrounds to regulatory agencies.[9] In the wake of Watergate, President Carter took seriously his commitment to restore confidence in government and thus broke from precedent by seeking out individuals with intellectual fortitude, expertise, and independence from traditional regulatory politics.[10] Fred Kahn fit this bill perfectly—with acclaimed experience as a (if not the) leading economist of regulation and as a respected regulator at the New York State Public Service Commission—and he approached his role at the Civil Aeronautics Board with the spirit of a true reformer. Moreover, as an outsider to the airline industry, he viewed both the traditional regulatory agenda and the traditional regulatory processes with skepticism and suspicion. In so doing, he astonished airline executives by making no effort to learn the details of their industry, famously remarking that "I really don't know one plane from the other. To me they are all marginal costs with wings (McCraw, 1984, p.224)."

As to Kahn's substantive agenda, he sought to undermine the traditional premises of airline regulation that competition was destructive and should be avoided through command and control regulation. Rather, he sought to act based on the economic analysis he set forth in his masterful *Economics of Regulation* (Kahn, 1970).[11] In the second

[8] An early contribution to public choice theory came in Mancur Olson, The Logic Of Collective Action: Public Goods And The Theory Of Groups 1-4 (Harvard University Press 1965). In the early 1970s, a number of seminal articles developed the concept. See George J. Stigler, The Theory of Economic Regulation, 2 Bell J. Econ. & Mgmt. Sci. 3 (1971); see also Richard A. Posner, Taxation By Regulation, 2 Bell J. Econ. & Mgmt. Sci. 22 (1971); see also Richard A. Posner, Theories of Economic Regulation, 5 Bell J. Econ. & Mgmt Sci. 335 (1974).

[9] See J. Landis, Report on Regulatory Agencies to the President-Elect (1960) for a description of the type of individuals appointed to regulatory agencies.

[10] When appointing others to the agency, Carter followed Kahn's guidance to avoid appointing a politician and to select individuals with "a firm, independent professional background and qualifications, and is in a position to make strong, intellectually-grounded professional judgments." (McCraw, 1984, p.289)

[11] See also Michael E. Levine, Airline Competition in Deregulated Markets: Theory, Firm Strategy, and Public Policy, 4 Yale J. Reg. 393, 394 (1987) for a description of how "by the mid-1970's it was probably fair to say that no impartial academic observer of any standing doubted that the airline business, if unregulated, would reach something that more or less resembled a competitive equilibrium."

volume of that treatise, Kahn (1971) cited experiments in airline competition undertaken at the state level that underscored the opportunity for discount airline operators to attract new customers and fly planes at much higher load factors than the industry average.

Immediately upon arriving at the CAB, Kahn sought to revolutionize the ways that the agency operated. For starters, he required all communication and orders to be written in understandable prose. In terms of process, Kahn criticized what he viewed as an intellectually bankrupt means of doing business—deciding issues in secret, without deliberation, and asking lawyers to develop the necessary justification for a pre-determined result.[12] In terms of agency structure, he also reorganized its operations, increased the role of professional economists, and appointed very talented deregulatory minded officials. Like at the New York State Public Service Commission, his leadership at the CAB was not hierarchical, involved recruiting and empowering talented staff, and included efforts to persuade existing professional staff of the merits of his views. In so doing, Kahn realized that the traditional substantive failings of the CAB went hand-in-hand with addressing its procedural flaws and legacy ways of doing business—by restricting competition and awarding benefits to a precious few, it was natural for the agency to operate along political lines as opposed to judicial ones.[13]

At the level of strategy, whether conscious or not, Kahn's overarching objective was to help foster a political environment conducive to regulatory reform. As Michael Levine (2006) describes it, the central challenge that Kahn faced was to overcome the presence of "slack" enjoyed by the legacy regulatory regime. On Levine's (2006) account, slack is the dynamic that creates political cover for regulatory beneficiaries to continue enjoying the benefits accorded to them while the public and their elected representatives are unable to appreciate the economic consequences of regulatory policy decisions.[14] In an alternative conception of slack, Professor Croley (1998, p.23) explains that "principal-agent 'slack'" is what enables "legislators [or regulators] to 'shirk'—to sacrifice their constituencies' collective interests for their own private goals and interests." On both explanations, slack represents the lack of awareness and attention to substance of the relevant regulatory policies and thus protects the beneficiaries of regulation and stands in the way of regulatory reform.

The reduction of slack is not necessarily something that a single political entrepreneur can accomplish. Rather, as Levine (2006, p.272) puts it: "Regulators can increase or minimize the transaction and information costs involved in monitoring and changing policy, and actors like media, academics, policy entrepreneurs, lobbyists, public interest groups, and politicians looking for an issue can increase or lower these costs through their own behavior." In the case of airline regulation, Kahn benefitted from and entered the

---

[12] In a famous speech to the American Bar Association, Kahn related that the traditional regulatory process involved:

> [A] lawyer on the General Counsel's staff, amply supplied with blank legal tablets and a generous selection of clichés—some, like "beyond-area benefits," "route strengthening" or "subsidy need reduction," tried and true, others the desperate product of a feverish imagination—would construct a work of fiction that would then be published as the Board's opinion.

(McCraw, 1984, p.286). This observation anticipates a later scathing critique of FCC regulation by Judge Richard Posner—"the [FCC's] opinion, like a Persian cat with its fur shaved, is alarmingly pale and thin." Schurz Comm, Inc. v. FCC, 982 F.2d 1043, 1050 (7th Cir. 1992).

[13] Kahn is quoted as stating that "[t]he dispensation of favors to a selected few is a political act, not a judicial one (McCraw, 1984, p.286)."

[14] Levine (2006, p.273 n.9) credits Kalt and Zupan for developing the concept.

surge of congressional interest in the topic that, in turn, "resulted in intense media coverage of the issue from 1976-1978 (Levine and Forrence, 1990, p.195)." Nonetheless, even if Kahn, acting alone, might have failed to spur regulatory reform, it is clear that his actions at the head of the CAB proved to be the lynchpin for regulatory reform. In particular, three distinct aspects of his leadership helped to catalyze the historic changes that occurred under his watch.

First and foremost, Kahn embraced the role of public advocate for regulatory reform. To be sure, he was fortunate to operate in a climate where other actors were interested in raising awareness about the benefits of deregulation, but his leadership—described by Levine (2006, p.277) as "intellectually powerful, attractive, and mediagenic"—was invaluable. In particular, his credibility stemmed from a widespread confidence that he was not trying to advance any self-serving political or parochial agenda. To that end, his prior academic writings—notably, his calls for airline deregulation in his Economics of Regulation—made him as bullet proof as a Washington inside player can be. And Kahn was not afraid to step into the line of fire, either, as he "testified before Congress, lobbied the White House, engaged the industry in debate, and was very accessible to the media (Levine, 2006, p.277)." Thus, a considerable part of Kahn's success as a political entrepreneur stemmed from his ability to translate his sophisticated academic understanding of regulation in terms that overcame the general bias that regulatory policy issues are too complex, esoteric, or boring to sustain public attention and concern.[15] Praising his public relations skills and ability to come up with the right one-liner, Thomas McCraw (1984, p.223) called him the "model of the modern media general."

A second component of Kahn's political entrepreneurship was his willingness to use the CAB's legal authority to move the industry toward deregulation. Notably, the agency took a number of steps to loosen its control over the industry and encourage entry,[16] thereby both facilitating experiments that made the case for further deregulation, undercut the industry's firm and united stance against deregulation, and made the ultimate move to complete deregulation less drastic for the industry players. Notably, under Kahn's leadership, the agency began, for the first time, to approve and promote the use of discount airfares, which facilitated price competition and improved dramatically the load factors on airplanes. (McGraw, 1984, p.275-77).[17] Indeed, when Congress enacted its legislation deregulating the industry, it did so recognizing that the law might actually be able to constrain the ongoing deregulatory zeal Kahn had brought to the CAB.[18] For Kahn, a willingness to use novel strategies was hardly new; at the New York Public Service Commission, he did just that, pioneering what might be called a "negotiated rulemaking," where he told the parties "I hope we can proceed as cooperative seekers of the truth, rather than as adversaries seeking an advantage over one another." (McCraw, 1984, p.251).

[15] As Levine and Forrence (1990, p.190) describe, "Most regulatory issues are highly specialized and do not generate much general attention. Complexity and lack of salience make efforts to overcome slack on regulatory issues—especially issues of economic regulation—particularly costly."
[16] See, for such an example, Oakland Serv. Case, 78 C.A.B. 593 (1978) (final order).
[17] In fairness to Kahn's predecessor, John Robson, he did oversee the approval of "supersaver" fares on an experimental basis, but it was Kahn who pushed to allow such fares more broadly and on a permanent basis.
[18] As the conference report stated, "Congress expects the deregulation of the aviation industry to move in accordance with this legislation and not in accordance with the, perhaps, differing concepts of some members of the CAB." H.R. REP. NO. 95-1779, at 56 (1978), as reprinted in 1978 U.S.C.C.A.N. 3773, 3775.

The third and final component of Kahn's strategy was to identify and bolster the interests and forces vested in deregulation. As Kahn (1990, p.331) explained, one of his goals at the CAB was to bolster the "vested interests in deregulation" on the view that the successful upstarts "having now achieved freedom from regulation, will not readily surrender it." In general, such interests are less well organized than the established interests that benefit from the status quo, meaning that they are uniquely apt to be activated by a political entrepreneur. Again, Kahn had tested this approach out in New York State, where he enlisted a number of groups (including less likely ones like the Environmental Defense Fund as well as a number of utilities) to support his initiative to reform utility rate structures (McCraw, 1984).[19] In the case of airline deregulation, it was his effort to empower and highlight upstarts like Southwest Airlines as well as to work with United Airlines, which had begun to doubt the wisdom of the legacy regulatory regime, that helped to lay the groundwork for regulatory reform.

3 Applying the Political Entrepreneurship Model to Spectrum Regulation

In many respects, airline regulation and spectrum regulation share some basic characteristics. On some accounts, both regulatory regimes emerged from an effort to protect established interests—the major airlines and the broadcasters—and limited output by restricting the use of the resource in question (Hazlett, 2001 and Keyes, 1951). In both cases, moreover, early academic criticism—famously, Ronald Coase (1959) in the spectrum context—called for regulatory reform that went unheeded (Keyes, 1949).[20] In the case of spectrum regulation, the movement toward regulatory reform came later, was halting, and remains incomplete. Indeed, Tom Hazlett has criticized the FCC's failure to escape the public choice trap of protecting incumbents by suggesting its initials can best be understood as "Forever Captured by Corporations" (Clark, 2005 and Hazlett, 2001). Part of the challenge facing spectrum reform, like the challenge faced by Kahn at the CAB, is that the FCC itself is not institutionally designed to, among other things, manage spectrum rights in a deregulated, property rights-like manner and thus invites the type of public choice behavior decried by Hazlett.[21]

In the case of spectrum reform, the biggest single force driving regulatory reform and challenging the status quo was the introduction of spectrum auctions. Before 1993, Congress called for the regulation of spectrum as a public trust whereby the Federal Communications Commission would hold comparative hearings to determine what

[19] "As McCraw (1984, p.255) reported, Kahn "[w]ork[ed] assiduously over a two year period . . . [to] shape[] a political consensus among unusual bedfellows: environmentalists, electric utilities, consumer groups, business firms, and residential customers."

[20] For an interesting study of the progress (and relative lack of progress) in the spectrum reform context, see the special issue of the *Journal of Law and Economics*, Vol. 41, No. 2, Part 2 (Oct. 1998), which emerged from a conference on "The Law and Economics of Property Rights to Radio Spectrum," sponsored by the Program on Telecommunications Policy, Institute of Governmental Affairs, University of California, Davis.

[21] The FCC's institutional limitations do not only restrict its ability to oversee property rights in spectrum. It also, as I have explained elsewhere, restricts its effectiveness in facilitating the use of unlicensed (or commons) spectrum. See Philip J. Weiser & Dale N. Hatfield, Policing the Spectrum Commons, 74 Ford. L. Rev. 663 (2005).

individual or firm merited a license to use the radio spectrum.[22] In 1993, Congress authorized the FCC to use auctions to assign spectrum licenses and, after the FCC held its first such auction (to assign new blocks of spectrum to be used for second generation (i.e., digital) cellular telephony), it generated billions of dollars in revenue for the Treasury.[23] Congress, recognizing easy money when it saw it, passed a new law requiring that all future assignments of the right to use radio spectrum be determined by an auction that would generate funds for the treasury.[24]

The radio spectrum remains, on almost any account, vastly underutilized. Thus, at the same time when entrepreneurs and users of all stripes are interested in gaining access to spectrum—and many established firms recently paid, in total, $19 billion for licenses that will become available after the digital TV transition—there is precious little spectrum available. This state of affairs reflects, among other things, that much of the wireless spectrum was allocated to specified uses well before the introduction of auctions and the FCC has historically done very little to ensure that spectrum is being used at all. To its credit, the FCC has authorized the leasing of spectrum licenses.[25] To date at least, a secondary market in spectrum has developed relatively slowly for bands in which it is authorized and, for many bands of spectrum, spectrum leasing is not allowed at all.

In 2002, the FCC, under Chairman Powell's leadership, sought to reform spectrum policy and, among other things, take aim at the agency's highly conservative and relatively undefined approach to spectrum property rights.[26] Under the agency's legacy model of regulation, licensees are protected against interference through a set of before-the-fact protective regulations that minimize the possibilities of interference. This means, among other things, that parties seeking additional flexibility must ask the agency to grant them rights to use its spectrum more flexibly—even if it highly confident (and would put its money on it, say, in a performance bond) that the different use would create a negligible, if any, amount of additional interference. During his tenure, Powell did not see any of these initiatives to completion and his successor declined to spearhead them, closing the relevant proceedings without attempting to address the underlying issue.

Powell's spectrum reform effort is notable both for its creative thinking about the need to define property rights in spectrum as well as the lack of any effort to re-think current spectrum allocation decisions. Consider, for example, that Powell did not attempt to develop a framework for enabling UHF TV broadcast licensees to allow other users (such as wireless broadband providers) to gain access to such spectrum.[27] Similarly, Powell did

[22] This system, as public choice theory would predict, rewarded politically connected firms and made entry more difficult. Among the famous tales of public choice behavior were Lyndon Johnson's ownership of radio and TV stations in Texas (Lady Bird Johnson was awarded the licenses). For a discussion of this episode, and how spectrum licenses were historically managed, see Nuechterlein & Weiser (2005).

[23] For the auction revenues collected see CAPCP, FCC Spectrum Auction Data, available at http://econ.la.psu.edu/CAPCP/FCC/index.html.

[24] Omnibus Budget Reconciliation Act of 1993, Pub. L. No. 103-66, § 6002 (1993).

[25] Promoting Efficient Use of Spectrum Through Elimination of Barriers to the Development of Secondary Markets, Second Report & Order, Order on Reconsideration, & Second Further Notice of Proposed Rulemaking, 19 FCC Rcd. 17,503, 17,505 (2004).

[26] The centrepiece of Powell's reform program was set out in an interagency task force report. See FCC, Spectrum Policy Task Force Report, ET Dkt. No. 02-135 (2002), available at http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-228542A1.pdf.

[27] To be fair, Powell's focus on facilitating the digital TV transition may well have limited his willingness to push on this point. For a discussion of a proposal aimed at addressing this very issue, see Philip J. Weiser,

not press—in large part because the FCC lacks jurisdiction over the issue—the question of whether the federal government is effectively using its spectrum.[28]

Future Commissions will confront the opportunity to advance the cause of spectrum reform along a number of possible dimensions. First, as noted above, there needs to be greater scrutiny of whether legacy allocations still make sense in light of the current demand for spectrum for emerging uses, such as wireless broadband. Second, the legacy approach for defining rights to use spectrum is, on all accounts, highly conservative and prevents spectrum from being used more intensively.[29] Third, there is a huge opportunity to welcome and encourage the development of new technologies, such as "smart radio" technologies, that can provide for more efficient uses of spectrum. Finally, there needs to be greater transparency of what spectrum licenses are out there and how they are being used.

One obvious question is how a future political entrepreneur at the FCC could follow Kahn's example to spearhead successful spectrum policy reform. The threshold and most fundamental challenge that a political entrepreneur would face in spurring spectrum policy reform along the lines discussed above is that the issue has proven politically intractable and incomprehensible to many legislators, let alone, the public at large. The formidable nature of this challenge has led some commentators, such as Jim Speta (2005), to doubt whether spectrum reform is readily achievable.

A political entrepreneur interested in developing the public case for spectrum policy reform confronts two notable challenges. First, many academics continue to debate different strategies (notably, whether to emphasize property rights or unlicensed spectrum), the appropriate tactics necessary to advance a particular strategy (say, whether to use courts or an administrative agency to oversee the use of spectrum property rights), and the proper role of spectrum allocations (what amount of spectrum should continue to be dedicated to over-the-air broadcast, for example).[30] Second, the entire vocabulary and discourse around spectrum policy is arcane and intimidating to outsiders. Thus, this state of affairs is a product of the culture of telecommunications regulation more broadly where, as Nicholas Lemann (2002) once explained, it is "an insider's game, less because the players are secretive than because the public and the press—encouraged by the players, who speak in jargon—can't get themselves interested."

The Untapped Promise of Wireless Spectrum, The Hamilton Project (2008), available at http://www.brookings.edu/~/media/Files/rc/papers/2008/07_wireless_weiser/07_wireless_weiser.pdf.

[28] In a later initiative, the Office of Management and Budget did just that, mandating that government users create and use a "shadow price" that the captures the value inherent in their use of spectrum with the aim of dissuading them of the fallacy that their use of spectrum is free. *See* Office of Management and Budget, "Preparation, Submission, and Execution of the Budget Section 33." Circular No. A-11 (2007), available at http://www.whitehouse.gov/omb/circulars/a11/current_year/a_11_2007.pdf. As of this writing, however, it remains to be seen how effectively this requirement will be enforced.

[29] For a critique of this system, see Philip J. Weiser & Dale Hatfield, Spectrum Policy Reform and the Next Frontier of Property Rights, 15 Geo. Mason L. Rev. 549 (2008). For suggestions of the role that the FCC should play with respect to unlicensed spectrum, see Philip J. Weiser & Dale N. Hatfield, Policing the Spectrum Commons, 74 Ford. L. Rev. 663 (2005).

[30] Jim Speta (2005 p. 185 n. 12 & 186) discusses the literature of different strategies for spectrum reform. For a sense of the debate on appropriate ways to implement spectrum property rights, compare Philip J. Weiser & Dale Hatfield, Spectrum Policy Reform and the Next Frontier of Property Rights, 15 Geo. Mason L. Rev. 549 (2008) with Thomas W. Hazlett, A Law & Economics Approach to Spectrum Property Rights: A Response to Professors Weiser & Hatfield, 15 Geo. Mason L. Rev. 975 (2008).

For those not well versed in the jargon of spectrum policy, it is difficult to appreciate how the current regulatory discourse clouds the relevant policy stakes of spectrum regulation. As J.H. Snider (2007) highlights in his article, *The Art of Spectrum Lobbying*, there is an enormous amount of information related to spectrum licenses—including the basic fact about who holds what licenses—that is not publicly available or, when theoretically available, is discussed in terms that are incomprehensive to anyone other than those very close observers of spectrum policy. The net result of this state of affairs is that the intellectual barriers to entry for understanding spectrum policy are seen as great and the benefits of trying to overcome them are seen as not worth the candle.

Assuming that a political entrepreneur can develop a compelling and popularly accessible rationale for spectrum reform, Kahn's example suggests a few additional prerequisites for the successful implementation of spectrum policy reform. First, it will be important that congressional leadership—like the kind demonstrated by Senator Kennedy with respect to airline deregulation[31]—emerges to champion the issue. Any such effort will need to pick up on the efforts of others to craft a simple narrative—lower prices or more advanced services for consumers, for example—that can be easily understood by policymakers and the public. Second, a political entrepreneur will need to look for ways to use the FCC's existing authority to pursue reform. Consider, for example, that the FCC has considerable legal latitude to reorient some of its own approaches for defining and enforcing property rights in spectrum.[32] Third, a political entrepreneur will need to develop support from those likely to benefit from future reform in this area.

Even if a future President taps an FCC Chair who pursues spectrum policy reform with the same level of zeal, political deftness, and intellectual honesty exhibited by Kahn in the case of airline deregulation, she will not succeed unless the overall political climate is receptive to reform. The public's level of interest in an otherwise arcane topic depends and openness to reform depends not merely on the development of an accessible narrative, but also a sense of crisis that renders the narrative compelling and important. With respect to airline deregulation, for example, the presence of rampant inflation provided an impetus for examining regulatory reform strategies that promised to lower the prices of services available to consumers.

The final ingredient of implementing spectrum reform also draws inspiration from Kahn's leadership at the CAB. In particular, if such an opportunity arises, a critical challenge—not unlike Kahn's challenge at the CAB—will be to reform how the FCC operates (i.e., its "spectrum culture"). At present, the agency focuses almost entirely on avoiding any potential for interference between users of spectrum, generally relying on the *ex ante* strategy of instituting protective measures. In so doing, it almost always avoids evaluating the actual uses of spectrum *ex post*. In this respect, spectrum regulation still largely reflects the command-and-control legacy of an earlier era and has yet to transition fully to a new model of regulation that would make use of adjudication proceedings that evaluate whether interference occurs in practice (and not merely in theory).

---

[31] For a discussion of the importance of the Kennedy hearings, see John W. Kingdon, A Model of Agenda Setting, With Applications, 2001 L. Rev. Mich. St. 331.

[32] For one such example, consider the FCC's reform of its rules governing broadband over powerline systems, where it instituted an after-the-fact assessment strategy in lieu of the traditional use of protective rules. *See* Amendment of Part 15 Regarding New Requirements and Measurement Guidelines for Access Broadband Over Power Line Systems, *Report & Order*, 19 FCC Rcd 21,265, 21,291-96 (2004).

4 Conclusion

In this Essay, I have sought to highlight that regulators have a lot to learn from Fred Kahn. As a political entrepreneur, he informs both the theory and practice of regulation like few others. As a scholar, he has called for evidence-based policy analysis that, on account of his public service, he has demonstrated can be put into practice. In so doing, he has challenged public choice theory's strongest claims and has highlighted that calls for regulatory reform are not merely tilting at windmills. Despite these accomplishments, scholars have yet to pay sufficient attention to different case studies of political entrepreneurship and Fred Kahn's leadership in particular. By sketching out how regulatory reform in the airline industry reflects the success of Kahn (and others) as a political entrepreneur, this Essay underscores that such studies are worth pursuing.

In closing, I must also acknowledge a quality of Fred's that enabled him to succeed and make him such a special role model—his concern with and engagement with others. Fred is a social person and his interest in a younger generation of regulators and scholars has benefitted numerous individuals (myself among them). For all of those who have gotten to know Fred, his striking combination of a willingness to learn from others, true humility, and superb intellect reminds us all not to take ourselves too seriously.

It is all too easy to despair about the state of our politics in general and regulatory reform efforts in particular. But Fred reminds us that such challenges are soluble and that we should undermine the power of ideas and determined leadership. He also leaves us all with the hopeful proposition that we can contribute as much intellectually at any age as he does at 90!

5 References

Clark, D. (2005). "Industry Experts Disagree on Best Path to Improve FCC," TechnologyDaily. March 24. http://www.nationaljournal.com/pubs/techdaily/pmedition/2005/tp050324.htm

Coase, R.H. (1959). "The Federal Communications Commission," *Journal of Law and Economics*, 2: 1-40.

Croley, S.P. (1998). "Theories of Regulation: Incorporating the Administrative Process," *Columbia Law Review*, 98:1-168.

Hardin, R. (1982). Collective Action. RFF Press: Washington.

Hazlett, T.W. (2001). "The Wireless Craze, The Unlimited Bandwidth Myth, The Spectrum Auction Faux Pas, and the Punchline to Ronald Coase's Big Joke: An Essay on Airwave Allocation Policy," *Harvard Journal of Law and Technology*, 14: 335-567.

Kahn, A.E. (1970). "The Economics of Regulation: Principles and Institutions," John Wiley & Sons, Inc., New York.

Kahn, A.E. (1971). "The Economics of Regulation Volume 2: Institutional Issues," John Wiley & Sons, Inc., New York.

Keyes, L.S. (1949). "National Policy Toward Commercial Aviation – Some Basic Problems," *Journal of Air Law and Commerce*, 16: 280-97.

Keyes, L.S. (1951). Federal Control of Entry into Air Transportation. Harvard University Press: Cambridge.

Kingdon, J. (1995). Agendas, Alternatives, and Public Policies, Harper Collins: New York.

Lemann, N. (2002). "The Chairman," *The New Yorker*. October 7.

Levine, M.E. and J.L. Forrence. (1990). "Regulatory Capture, Public Interest, and the Public Agenda Toward a Synthesis," *Journal of Law, Economy and Organization*, 6: 167-198.

Levine, M.E. (2006), "Why Weren't The Airlines Reregulated?" *Yale Journal on Regulation*, 23: 269-297.

McCraw, T. (1984). Prophets of Regulation. Belkap Press: Cambridge.

Migue, J.L. (1977). "Controls Versus Subsidies in the Economic Theory of Regulation," *Journal of Law and Economics*, 20: 213-21.

Mintrom, M. (2000). Policy Entrepreneurs and School Choice. Georgetown University Press: Washington, D.C.

Nuechterlein, J. and P.J. Weiser. (2005). Digital Crossroads. MIT Press: Cambridge.

Peltzman, S., M.E. Levine and R.G. Noll (1989). "Economic Theory of Regulation After a Decade of Deregulation," *Brookings Papers on Economic Activity: Microeconomics*, 1989: 1-59.

Pozen, D.E. (2008). "We Are All Entrepreneurs Now," *Wake Forest Law Review*, 43: 283-340.

Schuck, P.H. (1981). "The Politics of Regulation," *Yale Law Journal*, 90: 702-725.

Speta., J.B. (2005). "Making spectrum Reform Thinkable," *Journal of Telecommunication and High Technology Law*, 4: 183-215.

Snider, J.H. (2007), "The Art of Spectrum Lobbying," *New America Foundation Working Paper*, available at http://www.newamerica.net/files/WorkingPaper19_SpectrumGiveaway_Snider.pdf.

Stigler, G. J. (1971) "The Theory of Economic Regulation," *The Bell Journal of Economics and Management Science*, 2: 3-21.

Walker, J.L. (1974), "Performance Gaps, Policy Research, and Political Entrepreneurs: Toward a Theory of agenda Setting," *Policy Study Journal*, 3:112-116.

Weber, M. (1946). Politics as a Vocation. In H.H. Gerth & C. Wright Mills (eds), From Max Weber: Essays in Sociology. Oxford University Press: Oxford.

Wilson, J.Q. (1975). The Politics of Regulation. In J.W. McKie (ed), Social Responsibility and the Business Predicament. Brookings Institution Press: Washington.

Wilson, J.Q. (1980). The Politics of Regulation. Basic Books: New York.

6 Acknowledgments

Thanks to Brad Bernthal, Andrew Crain, Ray Gifford, Ellen Goodman, Jon Nuechterlein, Adam Peters, Jim Speta, Tim Tardiff, Heidi Wald, Dennis Weisman, and an anonymous referee for helpful comments and encouragement and to Dan McCormick for outstanding research support.

Journal of Economic Literature
Vol. XLIV (June 2006), pp. 325–366

Regulation, Competition, and Liberalization

Mark Armstrong and David E. M. Sappington*

In many countries throughout the world, regulators are struggling to determine whether and how to introduce competition into regulated industries. This essay examines the complexities involved in the liberalization process. While stressing the importance of case-specific analyses, this essay distinguishes liberalization policies that generally are procompetitive from corresponding anticompetitive liberalization policies.

1. *Introduction*

Economists have developed an extensive set of principles for regulating a monopoly supplier. The benefits of unfettered, pervasive competition are also well documented and well understood. However, our understanding of the precise conditions under which regulated monopoly supply is preferable to unregulated competition is limited. Furthermore, we know relatively little about optimal liberalization policies—the policies that govern the transition to competitive market conditions—in cases where competition is deemed superior to monopoly.

The purpose of this essay is to explore these two issues, both of which are of substantial practical importance throughout the world. These issues are particularly relevant in key network industries (such as the telecommunications, natural gas, electricity, transport, and water industries) where scale economies can render production by many firms uneconomic, but where some competition may be useful to help discipline incumbent suppliers of key services. Our analysis of the choice between regulated monopoly and unregulated competition, like our analysis of the design of liberalization policies, emphasizes the problems that imperfect information and imperfect institutions pose for the design of industry policy.

This essay has three main parts: (1) section 2 reviews some recent experience with liberalization policy; (2) sections 3 through 5 consider the choice between unregulated competition and regulated monopoly; and (3) sections 6 and 7 consider the design of liberalization policies in settings where competition is preferred to monopoly.

Section 2 summarizes selected experiences with liberalization in three network industries where liberalization has garnered significant attention in recent years:

* Armstrong: University College London. Sappington: University of Florida. We are grateful to Sanford Berg, Severin Borenstein, Cory Davidson, Roger Gordon, Antonio Estache, Jerry Hausman, Mark Jamison, Paul Joskow, Mircea Marcu, John McMillan, Paul Sotkiewicz, John Vickers, Catherine Waddams, Helen Weeds, and anonymous referees for helpful comments and observations.

telecommunications, natural gas, and electricity. The experiences in Chile's telecommunications industry, the United Kingdom's natural gas industry, and the state of California's electricity industry have varied greatly, ranging from substantial success in Chile to substantial failure in California.

Section 3 begins the discussion of the choice between unregulated competition and regulated monopoly by analyzing simple formal models of these two forms of industrial organization. The models emphasize the role of imperfect information. Section 4 considers additional factors that affect the choice between regulation and competition, including the resources available to the regulator, the regulator's independence and autonomy, the need to ensure ubiquitous service at affordable prices, and the importance of investment and innovation. Section 5 notes that, even when monopoly supply is the preferred mode of industry operation, some of the benefits of competition may be secured by allowing potential operators to bid for the right to serve as a regulated monopoly supplier.

Section 6 begins the discussion of liberalization policies in settings where the benefits of competition outweigh its costs. The potential merits and risks of direct entry assistance are considered first. Liberalization policies that are likely to reduce the intensity of long-term industry competition (and therefore generally are not recommended) are reviewed next. These policies include providing temporary monopolies or oligopolies, excluding foreign investors, specifying market share targets for industry participants, restricting incumbent suppliers asymmetrically, and affording competitors unduly favorable long-term access to the incumbent's infrastructure. Section 7 discusses liberalization policies that typically will enhance the intensity of long-term industry competition (and therefore are recommended). These policies include reducing the costs that customers incur when they switch suppliers; rebalancing the prices charged by the incumbent supplier to better reflect its operating costs and otherwise redesigning the regulations imposed on the incumbent supplier to account explicitly for emerging competition; privatizing state-owned enterprises; establishing appropriate access prices; and increasing monitoring, data reporting, and antitrust scrutiny, at least during the early stages of liberalization.

Two central themes emerge in this essay. First, even the comparatively simple choice between regulated monopoly and unregulated competition can be intricate and complex in practice. Second, the decision to introduce competition into an industry is only the beginning of a journey down a long and winding road that can present many obstacles and detours. Furthermore, the best route from monopoly to competition can differ substantially in different settings. Therefore, there is no single set of directions that can guide the challenging journey from monopoly to competition in all settings.

Even though detailed, comprehensive directions typically are not available, some broad conclusions (summarized in section 8) that can serve as useful guide posts can be drawn. These broad conclusions include the following five. First, the greatest potential gains from competition tend to arise when (1) industry scale economies are limited relative to consumer demand; (2) the industry regulator has limited information, limited resources, and limited instruments with which to craft policy; (3) the regulator's commitment powers are limited; and (4) subsidization of the consumption of some of the dominant supplier's services either is not critical or can be achieved by means other than through distortions in the supplier's price structure. Second, there are a wide variety of liberalization policies, and the merits of the different policies vary considerably. Therefore, it generally is more appropriate to inquire about the benefits and costs of specific liberalization policies than to ask whether liberalization per se is desirable or undesirable. Third, liberalization policies

that primarily aid some competitors and handicap others on an ongoing basis can hinder the development of vigorous long-term competition. Fourth, liberalization policies that remove entry barriers and empower consumers to discipline industry suppliers typically are the best means by which to foster vigorous long-term competition. Fifth, well-designed liberalization policies can facilitate a transition from stringent, detailed regulatory control to less intrusive antitrust oversight. However, during the transition process, heightened regulatory and antitrust scrutiny may be required.

2. *Recent Experience with Liberalization*

To help ground the ensuing discussion of the principles of liberalization policy, we begin by reviewing briefly the recent experience with liberalization in Chile's telecommunications industry, the United Kingdom's natural gas industry, and California's electricity industry.

2.1 *Liberalization in Chile's Telecommunications Industry*[1]

The General Law of Telecommunications opened Chile's telecommunications industry to competition in 1982. This law generally did not restrict the number of licenses that would be granted to deliver telecommunications services in Chile. Importantly, the licenses were for *nonexclusive* provision of wireline and wireless services, both local and long distance. The primary obligation imposed on all telecommunications operators in Chile was to connect their networks in compliance with specified technical requirements. Other terms and conditions of interconnection typically were left to negotiation among the operators.

In 1982, local telecommunications services were supplied almost exclusively by Compeñía de Teléfonos Chile (CTC), while long distance telecommunications services were supplied almost exclusively by Entel. Not surprisingly, these two state-owned enterprises were not anxious to connect their networks with the networks of emerging rivals. Few interconnection agreements were signed and competition was limited until Chile implemented a dispute resolution process that ensured the timely execution of interconnection agreements. Long distance competition also was limited until an equal access requirement was imposed in 1994. The requirement stipulated that every time a customer placed a long distance call, she had to explicitly designate (via dialing a special code) a long distance company to carry the call. The special code for each long distance carrier—incumbent and entrant alike—had the same number of digits. Most importantly, a call was not automatically carried by Entel if the caller did not specify an alternative carrier.

Once new suppliers were afforded substantial opportunity to provide telecommunications services on terms comparable to those faced by incumbent suppliers, competition began to flourish in Chile. The number of fixed lines more than tripled (from roughly one million to more than three million) between 1992 and 2000, and the extensive waiting list for (fixed) wireline telephone service that had inconvenienced Chile's citizens for so many years quickly disappeared. The number of mobile telephone subscribers in Chile increased nearly tenfold (from roughly 36,000 to more than 3.4 million) during the 1991–2000 period and nearly doubled again (to more than 6.7 million) between 2000 and 2003. Prices for most telecommunications services in Chile also have declined substantially since 1990.

By 2003, nearly one-fourth of the fixed telecommunications lines in Chile were supplied by CTC's competitors rather than by the incumbent supplier. Competitors

[1] The following discussion is drawn primarily from Ahmed Galal (1996), Pablo T. Spiller and Carlo G. Cardilli (1997), David M. Newbery (1999, pp. 123–25), Cecile Aubert and Jean-Jacques Laffont (2002, pp. 39–41), Laffont (2005, pp. 214–16), and Ricardo D. Paredes (2005).

secured especially large market shares in Chile's urban regions. This pattern of entry likely reflects in part the lower unit costs of serving urban areas given their relatively high population densities. This entry pattern also may reflect regulated prices of basic telephone service that are further above costs in urban regions than in rural regions. The regulated prices of local telephone services in rural regions often do not reflect fully the relatively high unit costs of serving these regions. Instead, prices in rural regions tend to be set closer to (and sometimes below) cost while prices in urban regions often are set substantially in excess of cost. Such pricing patterns are designed to ensure that basic telephone service is affordable to all citizens. These patterns may be less pronounced in Chile than in some countries because a separate fund has been established to subsidize the purchase of basic telephone service in rural regions and low-income urban regions. Nevertheless, in jurisdictions where CTC is deemed to be a dominant supplier of local telephone services, it is required to set the same price for basic services across substantial geographic regions. Some suggest that this restriction on CTC's ability to target price reductions to those regions in which competition is most intense may be one cause of the substantial market share that competitive suppliers of local telecommunications services have been able to secure in Chile.[2]

2.2 *Liberalization in the United Kingdom's Natural Gas Industry*[3]

Natural gas was one of the principal network industries in the United Kingdom targeted for privatization under Margaret Thatcher's program of industrial reform. Natural gas was first extracted from the U.K. Continental Shelf in 1965 and British Gas (BG) was formed as a state-owned company in 1972. BG held a legal monopoly over the sale of gas to consumers and a legal monopsony over the purchase of gas from producers in the U.K. fields. Since gas exploration and production involve substantial sunk costs, BG's monopsony power necessitated the use of long-term (e.g., twenty-five year) purchase contracts to limit expropriation of the substantial investments made by gas producers.

An initial attempt to liberalize the gas sector occurred in 1982 when entrants were authorized to employ BG's pipelines (at terms negotiated with BG) to supply gas to final customers. In addition, BG's legal monopsony was removed. Although these steps were intended to facilitate competition in the gas industry, competition did not emerge immediately.

BG was privatized as a vertically integrated monopoly in 1986.[4] At this time, consumers were divided into two categories for regulatory purposes: (1) the tariff market, consisting of those consumers who purchased less than 25,000 therms of gas per year, and (2) the contract market, consisting of the remaining higher-volume consumers. BG continued to enjoy a legal monopoly in the tariff market after privatization. BG's prices in this sector were regulated: the annual increase in the average price per therm was limited to $RPI + Y - X$, where RPI is the inflation rate, Y is a measure of the increase in the price BG pays for gas, and X is a productivity improvement factor.[5] X was set at 2 percent in 1986, but raised to 5 percent in 1992.[6]

[2] See Paredes (2005), for example.

[3] The material is drawn from Mark Armstrong, Simon Cowan, and John Vickers (1994, chapter 8) and Newbery (1999, chapter 8).

[4] Alternatively, BG could have been separated into a national pipeline business and several regional suppliers before privatization, an approach subsequently pursued in the U.K. electricity industry.

[5] This Y factor enabled BG to pass through its costs of purchasing gas to consumers, and thereby limited BG's incentive to secure low gas prices. Subsequent regulation permitted BG to pass through only a more exogenous index of gas costs.

[6] The 2 percent value for X was widely believed to underestimate the realistic potential for productivity gains. Of course, lenient price regulations enhance revenues from privatization.

Competition in the contract market was potentially open to competition and BG's prices in this market were not regulated. Although it was legal, competition did not emerge in this market until 1990. The delay reflected in part limited regulation of the prices BG charged retail competitors for access to its pipelines. BG also practiced price discrimination in the contract market, charging higher prices for gas to those industrial users who had no reasonable alternative energy source (e.g., electricity).

To combat these perceived problems, new regulations were imposed in 1988. The regulations required BG to (1) publish a tariff on which any customer in the contract market could purchase gas (in an attempt to limit price discrimination); (2) publish a corresponding tariff of network access charges (although these charges remained otherwise unregulated); and (3) purchase no more than 90 percent of the gas in any newly discovered gas field. This final requirement essentially forced some limited entry into the contract market. By 1990, though, BG supplied 93 percent of the gas sold in this market.

Limited competition in the contract market promoted three further regulatory reforms in the early 1990s. First, the monopoly threshold in the tariff sector was reduced from 25,000 therms to 2,500 therms, thereby permitting smaller customers to purchase gas from competitors. Second, the prices charged for access to BG's pipelines came under regulatory control.[7] Third, BG was required to reduce to 40 percent (by volume) its share of sales to the contract market by 1995. At the same time, new gas fields were being opened, and entrants were able to buy supplies from these new fields.[8]

Although regulation did not force BG to separate its pipeline and gas supply operations, intense regulatory scrutiny and onerous separate accounting requirements led BG to undertake such vertical divestiture voluntarily. Furthermore, the monopoly franchise threshold of 2,500 therms was removed entirely in 1998, so all consumers were permitted to purchase gas from suppliers other than BG. Competition intensified, and all retail price controls were removed in 2002. Only access charges continue to be regulated. Many consumers continue to purchase natural gas from BG even though BG charges more than some competitors for what is largely a homogeneous product. This fact may suggest that customer switching costs or other causes of customer inertia are important in this industry.

2.3 *Liberalization in California's Electricity Industry*[9]

Although the state of California's experience with liberalization in its electricity industry is quite recent, the experience is already legendary. In 1996, California enacted legislation that introduced five primary changes in the state's electricity sector. First, electricity generation and wholesale prices for electricity were deregulated. Second, as they were required to do, the three incumbent (vertically integrated) suppliers of electricity sold a sizable portion of their generation capacity, focusing instead on the transmission and delivery of electricity. By 1999, the incumbent suppliers had sold to five independent suppliers generation capacity that produced roughly one-third of

[7] These access charges were distance related. Because BG's retail prices did not vary across geographic regions, this pricing policy enabled entrants to secure a higher profit margin when serving customers located close to the gas landing point.

[8] As a result, BG was left with long-term obligations to purchase gas that it no longer needed to serve its own consumers. When BG was privatized, investors were promised that BG would be able to supply approximately two-thirds of the demand for natural gas by volume until 2009. As regulation and liberalization developed, this promise was not honored.

[9] The following discussion is based primarily on Severin Borenstein (2002).

the state's electricity consumption. Third, retail competition was introduced as retail customers were permitted to purchase electricity from firms other than the incumbent suppliers. Fourth, a ceiling was imposed on the retail price that incumbent suppliers charged for electricity. Fifth, the three incumbent suppliers were afforded strong financial incentives to buy and sell electricity through the California Power Exchange (Cal PX) during its first four years of operation. Cal PX was established in 1998 to run the day-ahead market for electricity in the state of California. Although the incumbent suppliers' participation in Cal PX helped to ensure substantial short-term supply of and demand for electricity, the incumbents' corresponding limited use of long-term contracts for the purchase or sale of electricity ultimately proved to be detrimental.

In the summer of 2000, unusually high temperatures and very dry conditions produced both a substantial increase in the demand for electricity and a significant reduction in the available supply of electricity from hydroelectric generating units in the state. When less efficient generating units were dispatched to meet the high demand for electricity, the wholesale price of electricity soared well above the ceiling that had been imposed on retail electricity prices. Consequently, incumbent suppliers incurred severe financial losses. The higher wholesale prices were not sufficient to augment the supply of electricity greatly due to the long lead time necessary to construct new generating capacity and due to the particularly high costs faced by established producers of peak load capacity. These high costs reflected less efficient operating technologies, particularly high wholesale prices for natural gas, and expensive pollution permits.[10]

[10] It has also been suggested that some of the few large suppliers of electricity may have exacerbated the problem by intentionally reducing their supply of electricity, thereby raising the market-clearing price of electricity and, hence, the profits of electricity suppliers. (See Paul L. Joskow and Edward Kahn 2002, for example.)

In 2001, the state of California adopted drastic measures to quell the crisis that had developed in its electricity industry. The state established itself as the wholesale buyer of power for the incumbent suppliers in response to the reluctance of other parties (including wholesale suppliers of electricity) to deal with the financially distressed incumbents. The state also raised substantially the prices incumbent suppliers could charge for electricity, especially to large retail customers. In addition, to prevent these customers from securing electricity from alternative suppliers at lower prices, the state terminated retail competition. In essence, California's experiment with deregulation and liberalization had ended.

3. *Regulated Monopoly and Unregulated Competition*

In an economic paradise, where a regulator is omniscient, benevolent, and able to fulfill any promise he makes, competition cannot improve upon regulated monopoly. In such a paradise, the regulator will ensure the firm produces the ideal range of services at the lowest possible cost and will set welfare-maximizing prices for these services. Consequently, industry performance would not improve if an additional firm operated in this setting.

Of course, the real world differs markedly from this paradise. In practice, regulators invariably lack important information about the markets they oversee and so will not be able to direct and control perfectly the activities of a monopoly producer. Because of its daily operation in the industry and its direct contact with consumers, the regulated firm will be better informed than the regulator about the demand for the regulated services it supplies, the minimum possible current cost of delivering the services, and the potential for less costly future provision. This information asymmetry generally gives rise to an unavoidable trade-off between

rent and efficiency: the firm can be motivated to operate efficiently but only if it is awarded substantial rent for doing so. In particular, the firm will operate at minimum cost and attempt to satisfy the needs and desires of customers only if it is awarded the full surplus that its activities generate.[11] However, such a generous award to the regulated firm typically will provide it with significant rent and thereby reduce the net benefits enjoyed by consumers. To limit the rent that accrues to the regulated firm, some inefficiency typically is tolerated.

To examine the optimal resolution of this trade-off and to examine the impact of limited information on the choice between regulated monopoly and unregulated competition, consider the following simple model.[12]

3.1 *A Simple Model of Regulated Monopoly*

Suppose that when regulated monopoly is implemented the regulator faithfully pursues the social goal of maximizing the expected value of $V + \alpha U$, where V denotes the surplus enjoyed by consumers, U is the firm's rent, and $\alpha \in [0,1]$ is a parameter. Because $\alpha \leq 1$, society values consumer welfare at least as highly as the welfare of shareholders, perhaps because consumers are less wealthy than shareholders or because many shareholders reside in other jurisdictions.

A transfer payment T from consumers to the firm entails a reduction of $[1 + \lambda]T$ in the surplus enjoyed by consumers. The parameter $\lambda \geq 0$ represents the social cost of public funds. This cost arises from the distortions created by the taxes imposed on consumers/taxpayers to raise the funds.[13] Notice that because public funds entail unit cost $1 + \lambda$, each dollar of public funds secured by taxing the profit of the regulated firm can be employed to increase consumer/taxpayer welfare by $1 + \lambda$ dollars.

The monopoly supplies a single product at regulated unit price $p \geq 0$. The demand curve for this product, $q(p)$, is common knowledge. The regulator sets both the unit price, p, for the regulated product and a transfer payment, T, from consumers to the regulated firm. The firm is obligated to serve all customer demand at the established price.

The firm incurs a fixed cost of operation, F, and a constant marginal cost of production, c. For simplicity, the firm's marginal cost can take on one of two values, c_L or c_H. Let $\Delta \equiv c_H - c_L > 0$ denote the difference between the high and the low marginal cost. The firm knows from the outset of its interaction with the regulator whether its marginal cost is high or low. The regulator does not share this information and never observes the firm's marginal cost directly. The regulator perceives the two possible cost realizations to be equally likely. For simplicity, the firm's fixed cost of operation, F, is assumed to be common knowledge. The firm seeks to maximize its rent, U, which is the sum of its profit, $\pi \equiv q(p)[p - c] - F$, and the transfer payment, T, it receives from the regulator.

[11] Martin Loeb and Wesley A. Magat (1979) provide a formal statement and proof of this conclusion. William W. Sharkey (1979) critiques Loeb and Magat's analysis. Jorg Finsinger and Ingo Vogelsang (1981, 1982) and David E. M. Sappington and David S. Sibley (1988) suggest dynamic modifications of the Loeb and Magat policy that afford less rent to the regulated firm and/or can be implemented with less precise information about the surplus generated by the firm's activities.

[12] This model is drawn from the large literature that examines the design of regulatory policy when the regulated firm is better informed about its environment than the regulator. Reviews of this literature include Bernard Caillaud, Roger Guesnerie, Patrick Rey, and Jean Tirole (1988), David P. Baron (1989), Laffont and Tirole (1993), Laffont (1994), and Armstrong and Sappington (forthcoming). Joskow and Richard Schmalensee (1986), Armstrong, Cowan, and Vickers (1994), Glenn Blackmon (1994), Sappington (1994), Robert Mansell and Jeffrey Church (1995), Sappington and Dennis L. Weisman (1996), and Joskow (2005), among others, provide less technical discussions of incentive regulation. Michael A. Crew and Paul R. Kleindorfer (2002) and Vogelsang (2002), for example, provide critiques of this literature.

[13] Laffont (2005, pp. 1–2) reports that for each dollar of tax revenue the government collects, citizens in a developed country bear a cost of approximately \$1.30 (so $\lambda = 0.3$). The corresponding cost typically is substantially greater in developing countries.

Suppose the regulator announces that he will set unit price p_i and deliver transfer payment T_i to the firm when the firm reports its marginal cost to be c_i, for $i \equiv L,H$. When the firm with cost c_i chooses the (p_i,T_i) option, its rent will be $U_i \equiv q(p_i)[p_i - c_i] - F + T_i$. The revelation principle ensures there is no loss of generality in examining regulatory policies that induce the firm to report its marginal cost truthfully.[14] Therefore, social welfare when the firm's marginal cost is c_i is:

$$v(p_i) - [1 + \lambda]T_i + \alpha[q(p_i)[p_i - c_i] - F + T_i] = w_i(p_i) - [1 + \lambda - \alpha]U_i,$$

where $w_i(p_i) \equiv v(p_i) + [1 + \lambda][q(p_i)[p_i - c_i] - F]$, and where $v(p)$ denotes consumer surplus when price p is established. ($v(\cdot)$ is a convex function of p.)

If the firm's realized marginal cost were observed publicly, the regulator would implement the (Ramsey) price that maximizes $w_i(\cdot)$ when cost c_i is realized, for $i \equiv L,H$. These full-information prices will be marginal-cost prices ($p_i = c_i$) when λ is zero because transfers from consumers to the firm entail no direct social costs in this case. In contrast, when λ is large, payments to consumers financed by the firm's profit are highly valued and the full-information prices will approximate the prices chosen by an unregulated profit-maximizing monopolist.

When the firm's marginal cost is not observed publicly, it must be the case that $q(p_L)[p_L - c_L] - F + T_L \geq q(p_H)[p_H - c_L] - F + T_H$ or, equivalently,

$$U_L \geq U_H + \Delta q(p_H), \tag{1}$$

to ensure the firm truthfully reports its low marginal cost of production. To ensure the firm finds it profitable to operate when its marginal cost is high, it must be the case that $U_H \geq 0$. Because social welfare declines as the firm's equilibrium rent increases (when $\alpha < 1$ or $\lambda > 0$), U_H is optimally held to zero. To limit the firm's equilibrium rent, constraint (1) also will hold as an equality under the optimal regulatory policy, so $U_L = \Delta q(p_H)$. Therefore, total expected welfare if price p_i is set when marginal cost c_i is realized (for $i \equiv L,H$) will be:

$$\tfrac{1}{2}[w_L(p_L) - [1 + \lambda - \alpha]\Delta q(p_H)] + \tfrac{1}{2}\, w_H(p_H) - [1 + \lambda]G, \tag{2}$$

where G is a (fixed) cost of regulation that is financed with public funds. This cost might include the salary of the regulator and his staff, for example, and all other costs associated with acquiring essential information about the regulated industry.[15] Differentiating expression (2) with respect to p_H reveals that, when the regulator cannot observe the firm's marginal cost, he is able to achieve the level of expected welfare that he could achieve if the firm's costs were observable, but the high marginal cost was $\hat{c}_H$, where:

$$\hat{c}_H \equiv c_H + [1 - \frac{\alpha}{1 + \lambda}]\Delta > c_H. \tag{3}$$

Therefore, the optimal price when the high marginal cost is realized is the full-information (Ramsey) price corresponding to the inflated cost $\hat{c}_H$. The inflated price when c_H is realized reduces the number of units of output on which the firm can exercise its cost advantage when c_L is realized. Therefore, the increase in p_H and the associated reduction in T_H limit the rent that accrues to the firm when its marginal cost is c_L. Because the firm has no incentive to understate its production cost, there is no value to distorting the firm's activities when it reports its marginal cost to be c_L. Consequently, the optimal price when the low marginal cost is realized will be the full-information price corresponding to the firm's actual cost (c_L).[16]

[14] See Roger B. Myerson (1979), for example.

[15] G might also include the surplus lost when regulation retards industry innovation. Charles Jackson, Tracey Kelly, and Jeffrey Rohlfs (1991) and Jerry A. Hausman (1997) estimate that delays in licensing wireless telecommunications providers in the United States reduced consumer surplus by billions of dollars.

[16] This fact is evident from expression (2).

Three features of the optimal regulatory policy in this simple setting characterize optimal regulatory policy in more general settings where the monopoly supplier is better informed than the regulator about key features of the regulated industry. First, the firm generally commands rent from its superior information. Second, to limit this rent and thereby generate greater surplus for consumers, the regulator will design a menu of options from which the firm can make a binding choice. A well-structured menu of options can induce the firm to employ its superior knowledge to secure outcomes that are better for both the firm and consumers in more favorable environments (e.g., when the firm has lower operating costs). Third, the optimal regulatory policy generally induces inefficient performance to limit the firm's rent.

3.2 *A Simple Model of Unregulated Competition*

Now consider the following simple model of unregulated competition. Suppose two firms produce a homogeneous product and engage in Bertrand price competition. Each firm knows its own constant marginal cost of production and its rival's marginal cost, $c \in \{c_L, c_H\}$, when it sets its price. No transfer payments to or from the firms in the industry are possible in this setting. In particular, the firms' profits cannot be appropriated by the government to reduce the general tax burden elsewhere. Therefore, the social cost of public funds λ plays no role in this setting and social welfare is $v(p) + \alpha\pi$, where π is industry profit. Both firms find it profitable to operate in the industry.[17]

Each firm has the low marginal cost, c_L, with probability 1/2. The firms' costs may be correlated. Let $\rho \in [1/2,1]$ represent the probability that the two firms have the same cost.[18] The firms' costs are perfectly correlated when $\rho = 1$. Their costs are uncorrelated when $\rho = 1/2$. Bertrand competition ensures the equilibrium price will be c_H except when both firms have low cost, which occurs with probability $\rho/2$. A firm's operating profit is zero unless it has the low marginal cost while its rival has the high marginal cost, in which case the firm's profit is $\Delta q(c_H)$.[19] A firm realizes this positive profit with probability $[1 - \rho]/2$. Consequently, expected industry profit in this duopoly setting is $[1 - \rho]\Delta q(c_H)$, which declines as the firms' costs become more highly correlated. Notice that the probability that the industry supplier has low marginal cost is $(1 - \rho/2)$. This probability decreases as ρ increases. In contrast, the probability that the industry price will be c_L is $\rho/2$, which increases as ρ increases.

Ignoring the firms' fixed costs of production (F) for now, social welfare in this unregulated duopoly setting is:

$$(4)\ \frac{\rho}{2}v(c_L) + [1 - \frac{\rho}{2}]v(c_H) + \alpha[1 - \rho]\Delta q(c_H).$$

3.3 *Comparing Regulated Monopoly and Unregulated Competition*

Regulated monopoly offers four potential advantages over unregulated competition in this simple setting: (1) industry prices can be controlled directly; (2) transfer payments can be made to the firm to provide desired incentives; (3) the firm's profit can be taxed to generate public funds, thereby reducing the deadweight losses associated with other sources of public funds; and (4) duplicative fixed costs of production can be avoided because there is only one industry supplier.

[17] Section 6 considers the possibility that both firms may not find it profitable to operate in the industry. See Emmanuelle Auriol and Laffont (1992) and Michael H. Riordan (1996), for example, for related models in which the duopoly is regulated.

[18] The probability that both firms have high cost is $\rho/2$, the probability that both firms have low cost is $\rho/2$, and the probability that a given firm has low cost while its rival has high cost is $[1 - \rho]/2$.

[19] The profit-maximizing price for a firm with the low marginal cost is assumed to exceed c_H. Consequently, when only one firm has the low marginal cost, it will serve the entire market demand at price c_H in equilibrium.

Unregulated competition has three corresponding potential advantages: (1) the likelihood that the industry producer has the low marginal cost is higher than under monopoly because even if one firm fails to secure the low cost, its rival may do so; (2) the presence of a rival with correlated costs reduces the information advantage of the industry producer; and (3) any direct, operational costs of regulation (e.g., the salaries of regulators and their staff) are avoided. The first of these potential advantages of unregulated duopoly is referred to as the *sampling benefit of competition*. The second potential advantage will be referred to as the *rent-reducing benefit of competition*.

To compare the performance of regulated monopoly and unregulated competition in this setting, initially suppose $\lambda = 0$, $F = 0$, and $G = 0$, so there is no social cost of public funds, no fixed cost of production, and no direct cost of regulation. Expressions (2) and (3) reveal that the maximum expected welfare under monopoly regulation in this case is:

$$(5) \qquad \tfrac{1}{2}v(c_L) + \tfrac{1}{2}v(c_H + [1 - \alpha]\Delta).$$

A comparison of expressions (4) and (5) provides four conclusions regarding the relative performance of regulated monopoly and unregulated duopoly.

First, unregulated duopoly delivers a higher level of expected social welfare than does regulated monopoly when the duopolists' costs are perfectly correlated (so $\rho = 1$).[20] When costs are perfectly correlated, the industry producer never has a cost advantage over its rival and so commands no rent in the duopoly setting. Furthermore, competition drives the industry price to the level of realized marginal cost. Therefore, the ideal (full-information) outcome is achieved under duopoly but not under monopoly, where regulated prices diverge from marginal cost in order to limit the rent the monopolist commands from its privileged knowledge of costs. In this case, then, unregulated duopoly is preferred to regulated monopoly even though the former offers no sampling benefit. The benefits of competition arise entirely from rent reduction in this case.

Second, when demand is perfectly inelastic, unregulated duopoly produces a higher level of expected welfare than does regulated monopoly.[21] When demand is perfectly inelastic, price distortions do not affect output levels and, therefore, do not serve to limit rent. Consequently, only the probability of obtaining a low-cost supplier affects expected welfare, and this probability is higher under duopoly than under monopoly because of the sampling benefit of competition.

Third, regulated monopoly will generate a higher level of expected welfare than unregulated duopoly when demand is sufficiently elastic (and $\rho < 1$). When demand is very elastic, prices that do not track costs closely entail substantial losses in surplus. Prices track costs more closely under regulated monopoly than under unregulated duopoly.[22]

Fourth, unregulated duopoly outperforms regulated monopoly when the difference between the high and the low marginal cost (Δ) is sufficiently close to zero. Monopoly rent and duopoly profit are both negligible in this case, and so the choice between monopoly and duopoly depends upon which regime produces the low marginal cost more frequently. The sampling benefit of competition ensures the duopoly regime does so

[20] If $\alpha = 1$, regulated monopoly and unregulated duopoly provide the same level of expected social welfare.

[21] When demand is perfectly inelastic (so $q(p) \equiv 1$, for example), expression (4) is weakly greater than expression (5) whenever $\rho/2 \geq \alpha[\rho - \frac{1}{2}]$, which is always the case.

[22] The convexity of $v(\cdot)$ implies the expression in (5) is at least $[v(c_L) + v(c_H) - [1 - \alpha]\Delta q(c_H)]/2$. Therefore, the difference between expressions (5) and (4) is at least $[1 - \rho][v(c_L) - v(c_H)]/2 - \Delta q(c_H)[(1 - \alpha)/2 + \alpha(1 - \rho)]$. Because this expression is nondecreasing in α, it is at least $[1 - \rho][v(c_L) - v(c_H)]/2 - \Delta q(c_H)/2$. This term is positive when demand is sufficiently elastic.

whenever costs are not perfectly correlated (so $\rho < 1$).[23]

More generally, if the fixed costs of operation (F) are sufficiently large, regulated monopoly will outperform unregulated duopoly in the simple model analyzed here because monopoly avoids the duplication of fixed costs.[24] This conclusion is a corollary of the more general observation that monopoly supply will minimize industry costs if prevailing scale economies are pronounced relative to industry demand.[25] Scale economies often are pronounced in network industries, where substantial physical infrastructure (e.g., a gas, water, or electricity distribution system or a telecommunications network) must be deployed in order to deliver service to customers.

When the social cost of funds (λ) is considered, regulated monopoly offers an additional advantage over unregulated duopoly. The monopolist's rent can be taxed to fund desirable social projects, thereby reducing the need to employ other (potentially more costly) means to raise revenue.

4. *Additional Considerations*

The simple models considered in section 3 abstracted from several important institutional factors that can affect the optimal choice between regulated monopoly and unregulated competition. These institutional factors include (1) the resource constraints the regulator faces; (2) the potential role of regulation in pursuing distributional objectives; (3) the instruments available to the regulator; (4) the prevailing degree of regulatory independence and accountability; (5) the ownership structure of the incumbent industry producers; and (6) the importance of industry investment and innovation. These factors are now considered in turn.[26]

4.1 *Resource Constraints*

Although the regulator was not omniscient in the models of section 3, he had considerable knowledge of the regulated industry. In practice, a regulator's information can be far more limited.[27] A regulator's difficult task of overseeing and directing the activities of a monopoly supplier can become nearly impossible when the regulator's information and expertise are severely limited and when he lacks the physical and financial resources to overcome these limitations. Consequently, allowing competition to replace regulatory oversight as the primary means of motivating and disciplining the incumbent supplier can be advantageous when the efficacy of regulatory oversight is severely compromised by limited regulatory resources.[28]

[23] The benefits of unregulated competition may be less pronounced if industry competition is less pronounced. For example, Cournot competition may better describe industry interaction than Bertrand competition. Alternatively, industry producers might collude in setting prices. (Recall that limited industry competition may have contributed to the substantial increase in the wholesale price of electricity in California in 2000.)

[24] In contrast, if the fixed cost of regulation, G, is sufficiently large, unregulated duopoly will outperform regulated monopoly because the former avoids this cost, by assumption.

[25] In developing countries where consumers have limited income, the corresponding limited demand for key services can result in higher average production costs when scale economies are present.

[26] For further discussion of the effects of institutions on regulatory policy, see Brian Levy and Spiller (1994, 1996), Newbery (1999), Roger G. Noll (2000), J. Luis Guasch (2004), Ioannis N. Kessides (2004), Mark A. Jamison, Lynne Holt, and Sanford V. Berg (2005), and Laffont (2005), among others. We focus on the effects, rather than the origins, of a country's institutions. Daron Acemoglu, Simon Johnson, and James Robinson (2005) trace a country's institutions to its colonial origins. In countries where poor living conditions rendered settlement unattractive, for example, little effort was devoted to developing the country's institutions. Such effort was more prominent in countries where long-term settlement was more attractive and more widespread.

[27] The theory of optimal regulation when the firm is privately informed about several aspects of its operation awaits further development. Jean-Charles Rochet and Lars Stole (2003) provide a survey of the theory of multidimensional screening.

[28] Jon Stern (2000) documents the limited regulatory resources that are available in many countries. Stern suggests that resource sharing among regulatory agencies can help to mitigate partially the effects of severe shortages of critical regulatory resources in some settings. Noll (2000) also emphasizes the merits of sharing regulatory resources.

The resources of other oversight agencies also influence the relative merits of monopoly supply and liberalization. For example, competition may be less likely to impose meaningful discipline on a dominant incumbent supplier if the antitrust agency that oversees industry competition has limited expertise and meager physical and financial resources. Consequently, both the relative and absolute level of resources available to oversight agencies influence the relative merits of monopoly and liberalization.

4.2 *Income Redistribution and Universal Service Goals*

The analysis in section 3 abstracted from any differences in the cost of serving different customers and did not model explicitly differences in the wealth levels of different consumers. Such differences can raise concerns about *universal service*—the ubiquitous delivery of essential services at affordable rates. When a government has limited ability to redistribute income directly (perhaps because income taxes are widely evaded, for example), the regulated prices of essential services can constitute an important means to promote universal service.[29] To illustrate, suppose a country wishes to ensure that all its citizens have access to clean water or to basic communications services at low prices. Because prices tend to reflect costs under unfettered competition, individuals that are particularly costly to serve (because of their geographic location, for example) may face unduly high prices for key services under unregulated competition. In contrast, a regulated monopolist will agree to serve high-cost (e.g., rural) customers at relatively low prices if it is permitted to offset the associated financial losses by charging prices sufficiently above the costs of serving low-cost (e.g., urban) customers. Consequently, universal service (and other distributional) concerns can cause regulated monopoly to be preferred to unregulated competition.

4.3 *Available Instruments*

The policy instruments available to oversight agencies also influence the relative merits of monopoly and competition. To illustrate, return to the simple setting considered in section 3, where the regulator was able to set prices and deliver transfer payments. In practice, regulators are not always able to deliver transfers to the firms they regulate. If the regulator lacked this ability in the setting analyzed in section 3.1 and wished to ensure the monopolist never terminated its operations, the regulator could do no better than to set a single price equal to the high marginal cost, $p = c_H$ (assuming fixed costs are zero). This policy would generate expected welfare

$$v(c_H) + \tfrac{1}{2}\alpha\Delta q(c_H). \tag{6}$$

It is apparent that the level of expected welfare in expression (6) is less than the corresponding level in expression (4).[30] Consequently, unregulated duopoly is always preferred to this restricted form of monopoly regulation in the setting considered in section 3.1.

More generally, if a regulator has limited ability to reward a monopoly supplier for superior performance and penalize the firm for inferior performance, the regulator may be unable to motivate the firm to serve consumers well. This is the case regardless of how well the regulator understands the firm's capabilities and consumers' preferences. Similarly, if the regulator is not

[29] The public finance literature (e.g., Anthony B. Atkinson and Joseph E. Stiglitz 1976) notes that when a nonlinear income tax is feasible, prices of goods and services often should reflect production costs. Laffont and Tirole (2000, section 6.2) consider the implications of this conclusion for pricing policy in regulated industries. Robin Burgess and Stern (1993) review the theory and the practice of tax policy in countries with limited wealth.

[30] The difference between expressions (4) and (6) is $[v(c_L) - v(c_H)]\rho/2 + \alpha\Delta q(c_H)[\frac{1}{2} - \rho]$. The convexity of $v(\cdot)$ implies this expression is at least $[\rho/2 - \alpha(\rho - \frac{1}{2})]\Delta q(c_H)$. This expression is nonnegative because it is a decreasing function of ρ and is nonnegative at $\rho = 1$.

authorized to compel the firm to report data on its operations, the regulator will find it difficult, if not impossible, to make informed policy decisions, even if the regulator has ample resources and ability to analyze and interpret data. Consequently, limited powers to reward or penalize the firm or to compel data reporting can render monopoly regulation ineffective, and so can increase the relative merits of unregulated competition.

A regulator's powers to control the activities of new suppliers also can affect the relative merits of monopoly and competition. To illustrate, suppose the regulator cannot impose any regulations on new competitors. In particular, suppose a regulator cannot limit the range of services that competitors offer or tax any of these services. In this case, even though liberalization may help to motivate the incumbent supplier to reduce its operating costs, it may allow competitors to engage in cream-skimming. Cream-skimming is the act of serving the most profitable (e.g., urban, business telecommunications) customers and leaving the incumbent supplier to serve the less profitable (e.g., rural, residential telecommunications) customers. Cream-skimming can limit the ability of the incumbent supplier to finance particularly low prices on some services with substantial profit earned on other services and thereby undermine socially desirable pricing structures.[31] Therefore, even though managed competition (which entails the regulation of both incumbent suppliers and new entrants) might be preferable to monopoly, monopoly may be preferable to unfettered competition when an ineffective tax system requires that universal service be pursued through industry prices.[32]

[31] See Nicolas Curien, Bruno Jullien, and Rey (1998) and Laffont and Tirole (1990a), for example.

[32] Of course, a country with limited ability to collect taxes also may have limited ability to enforce desired prices. The relative strengths of the country's institutions are important to consider in such cases. A country may be better able to measure electricity consumption and enforce associated charges than to prevent citizens from hiding wealth in order to evade income taxes, for example.

4.4 *Private versus State Ownership*

The extent of government ownership of the dominant incumbent supplier also can affect both the merits and the most appropriate form of liberalization. A firm that is largely owned by the government can be less responsive to the oversight and demands of shareholders than are privately owned firms.[33] Furthermore, state-owned enterprises (SOEs) can face softer budget constraints than their privately owned counterparts in the sense that the government may be more willing to tolerate losses by the firm and to finance the firm's continued operation despite poor historic financial performance. Because of the corresponding diminished incentive to minimize production costs,[34] an SOE may operate with higher costs than a privately owned monopoly.[35] Therefore, liberalization may have greater potential to reduce industry costs when the incumbent monopoly supplier is owned primarily by the government than when it is owned primarily by private investors.[36]

4.5 *Regulatory Independence and Accountability*

The simple models considered in section 3 presumed the regulator faithfully pursued the social objective. In practice, a "captured"

[33] For instance, Michael I. Cragg and I. Alexander Dyck (2003) find that the financial compensation of managers is not as closely linked to the performance of the firm in state-owned enterprises as it is in privately owned firms.

[34] See János Kornai, Eric Maskin, and Gerald Roland (2003).

[35] William L. Megginson, Robert C. Nash, and Matthias van Randenburgh (1994), Juliet D'Souza and Megginson (1999), Rafael La Porta and Florencio López-de-Silanes (1999), Kathryn L. Dewenter and Paul H. Malatesta (2001), Megginson and Jeffry M. Netter (2001), Scott J. Wallsten (2001), and Simeon Djankov and Peter Murrell (2002), among others, present (sometimes mixed) evidence that the efficiency of SOEs increases after privatization, particularly in the presence of substantial industry competition and independent regulators.

[36] Of course, in practice, this greater potential need not always translate into more pronounced reductions in industry costs. Fumitoshi Mizutani and Shuji Uranishi (2003), for example, report that liberalization in Japan's postal industry did not substantially increase the productivity of the SOE.

regulator may not do so. A regulator is said to be captured by the firm he regulates when the regulator generally implements policies that further the interests of the firm at the expense of the broader social interest.[37] Many factors increase the likelihood of regulatory capture. For example, a regulatory agency with limited expertise and resources may be forced to rely heavily on the advice and information supplied by the firm when formulating policy. Alternatively, the firm may routinely offer attractive employment opportunities to regulators who have proved to be cooperative, and the country's laws may not preclude such offers.[38] Also, the firm may provide a sizable portion of the regulator's ongoing budget and may have some discretion over the timing and magnitude of its contributions to the regulator's budget. When factors like these lead to a high likelihood of regulatory capture, the entry of additional competitors may be the best way to impose meaningful discipline on the incumbent supplier and otherwise ensure long-term gains for consumers.[39]

The importance of regulatory independence from short-term popular opinion also is apparent. Producers in network industries typically incur substantial sunk (nonrecoverable) costs. These producers also often deliver services to a large portion of the population, and so the prices of these services are of substantial public concern. Together, these two elements create substantial risk of expropriation by well-meaning but short-sighted regulators. In response to public pressure, regulators may reduce prices as far as possible toward variable production costs. As long as regulated prices allow the firm to recover its variable production costs, the firm will prefer to continue to produce than to terminate its operations. Thus, in the short run, the regulator gains by securing low prices and ongoing production of key services. Although such a policy may provide short-term gains, it can have substantial long-term costs. Potential and actual producers will realize they are unlikely to recover any sunk costs they incur. Consequently, they will be reluctant to incur such costs, and so existing network infrastructure will be permitted to decay and new network infrastructure will not be built. Thus, if regulators are to design and implement policies that best serve the long-term interests of consumers, they must be able to develop policy credibility by resisting short-term pressures to renege on long-term promises.[40]

A regulator's commitment powers can be enhanced by a variety of factors, including strong legal institutions and a long tenure. Strong legal institutions can thwart attempts by other government agencies to intervene in the day-to-day operations of the regulatory agency and can thereby enhance a regulator's commitment powers by reducing the likelihood that the terms of announced regulatory policies will be changed.[41] In particular, strong legal institutions can enforce long-term

[37] See George J. Stigler (1971), Richard A. Posner (1974), Gary S. Becker (1983, 1985), Laffont and Tirole (1993, chapter 11), and James F. Dewey (2000), for example, for models of capture and the strategic choice of expenditures to influence regulators.

[38] See Yeon-Koo Che (1995) and David J. Salant (1995), for example.

[39] Bengt Holmstrom and Paul Milgrom (1991), Holmstrom (1999), and Mathias Dewatripont, Ian Jewett, and Tirole (1999) examine the effects of long-term career concerns on short-term behavior. Such concerns can, for example, induce a regulator to focus primarily on performance dimensions that are readily observed (e.g., short-term price reductions) and less on performance dimensions that are more difficult for potential employers to observe in a timely fashion (e.g., prospects for viable long-term industry competition).

[40] See Oliver Williamson (1975) for a pioneering treatment of the problem and Newbery (1999) for an extensive discussion of the problem of regulatory commitment. Some authors (e.g., Witold J. Henisz 2000) have developed indices to measure expropriation risk in different countries. Jamison, Holt, and Berg (2005) review these indices.

[41] Wallsten (2001) finds that the privatization of state-owned telecommunications providers in Latin America and Africa is associated with improved industry performance when the industry regulator is independent (in the sense of not being directly under the control of a government ministry), but not otherwise. Geoff Edwards and Leonard Waverman (2006) find that regulatory independence is associated with lower charges for access to the network infrastructure of state-owned incumbent suppliers of telecommunications services.

contracts between the regulator and the firm, and can prevent a regulator from changing the terms of announced policies (including promised returns on investment) in response to pressure from other government agencies or the citizenry at large. When legal institutions are weak, pressure groups may anticipate substantial benefits from convincing the regulator to renege on the promises he has made to the firm.[42] Thus, the independence of a country's judicial system and the ability of a country to enforce the terms of legal contracts can affect the optimal design of industry policy.

A regulator may also be better able to pursue policies (such as delivering promised rewards to the regulated firm) that promote the long-term, rather than the short-term, interests of consumers if his tenure as regulator is relatively long. Long-lived regulators with sufficient concern about future industry outcomes will realize that short-term expropriation of an incumbent producer's investment will discourage future investment by the same firm or its successor.[43] In addition, a regulator may feel less pressure to pander to popular opinion if he is not elected by direct vote of the citizenry.[44] Thus, a setting where the regulator serves a fairly long term and faces reappointment by the government may provide an appropriate trade-off between independence and accountability.[45]

There is an additional benefit of a reasonable degree of regulatory independence from direct intervention by other government agencies. A government that is unable to intervene in the short-term operations of a regulatory agency is compelled to state as fully and clearly as possible the mission of the regulatory agency. Failure to do so will afford the regulatory agency the opportunity to pursue its own mission and goals, rather than those of the government more broadly.[46] A coherent and transparent statement of the principles that will guide industry policy provides greater certainty for industry participants, which can encourage investment and facilitate long-term planning.[47]

When regulatory capture is likely, explicit restrictions on the regulator's autonomy and commitment powers may be desirable in order to limit the regulator's ability to pursue interests other than the long-term social interest.[48] For example, key powers might be dispersed among multiple regulators or divided between a regulator and other government agencies.[49] Alternatively or in addition, the regulator's discretion in formulating

[42] Kessides (2004, p. 102) describes settings where governments have changed the terms of legal contracts.

[43] See Newbery (1999), Salant and Glenn A. Woroch (1992), and Paul Levine, Stern, and Francesc Trillas (2005) for formal analyses of this effect. Levine, Stern, and Trillas (2005) also emphasize the potential value of delegating authority to an independent regulator with substantial concern for the welfare of the firms he regulates.

[44] Timothy Besley and Stephen Coate (2003) employ data from the United States to examine the effects of the method by which state regulators are selected on performance in the state's electric power industry. The authors find lower prices and reduced investment (as proxied by the frequency of power outages) in states where regulators are elected than in states where regulators are appointed.

[45] See Maskin and Tirole (2004) for a formal analysis of this trade-off.

[46] In this respect, regulatory independence can deliver benefits similar to those provided by the privatization of a state-owned enterprise. Privatization increases the cost to the government of intervening in the day-to-day operations of the firm, and thereby renders more credible the government's promise to refrain from such intervention (Sappington and Stiglitz 1987). Regulatory independence enhances the commitment powers of regulators in similar fashion, once the mission and goals of the regulatory agency are stated clearly.

[47] Spiller and Vogelsang (1997) and Tonci Bakovic, Bernard Tenenbaum, and Fiona Woolf (2003) emphasize the value of coupling substantial regulatory independence with a clearly specified regulatory contract (e.g., a regulatory license). Under such coupling, regulators will have the independence required to implement impartially the details of a politically popular contract, but will not be empowered to implement any policy of their choosing. Using a large dataset of monopoly franchise auctions, Guasch (2004, table 1.16) reports that the original terms of the franchise contract were renegotiated 61 percent of the time when there was no separate regulatory body responsible for contract administration. The corresponding percentage was only 17 percent when the contracts were administered by a separate regulatory body.

[48] The independence granted the telecommunications regulator in Jamaica is believed to have led to a decline in industry investment as investors feared the regulator would use his autonomy to expropriate investments (Levy and Spiller 1994).

[49] See Laffont and David Martimort (1999) and Martimort (1999b).

policy might be substantially restricted.[50] The tenure of individual regulators might also be limited.[51]

Competition can play a particularly valuable role in disciplining and motivating the incumbent supplier to pursue the best interests of consumers when the regulator's powers to do so are (inherently or intentionally) limited. However, competition is not necessarily a panacea. Weak commitment powers may limit nonfungible investment under both unregulated competition and regulated monopoly. Competitors, like the incumbent supplier, will recognize that their investments may be expropriated by a regulator with substantial expropriation powers and limited commitment powers, and so may be reluctant to undertake the investment required to improve industry performance.[52] Thus, there is seldom an effective substitute for strong commitment powers.

4.6 *Investment and Innovation*

The link between industry structure and the incentive to innovate and to reduce costs is complex even in the absence of regulation.[53] Competing firms may have greater incentive than an unregulated monopoly provider to reduce operating costs in part because industry output (and thus the potential cost savings from a reduction in marginal production cost) is greater under competition than under unregulated monopoly.[54] In contrast, substantial market concentration can encourage innovation for at least two reasons. First, the profit a monopolist generates can serve as a valuable source of research and development (R&D) funding.[55] Second, the prospect of substantial monopoly profit can be a compelling reason to undertake R&D investment.[56]

Regulatory policy can affect infrastructure investment differently than it affects innovative effort and investment designed to reduce operating costs.[57] To illustrate this point, first consider rate of return regulation, which promises a fair return on prudently incurred investment. When expropriation can be avoided, such a promise can deliver strong incentives for infrastructure investment. In contrast, because it requires revenues to track costs closely, rate of return regulation (like other forms of "cost-plus" regulation) typically provides limited incentive for innovation and cost reduction.[58] Now consider price cap regulation, which typically permits revenues to diverge from realized costs for a specified period of time (e.g., four years) but does not promise

[50] See Laffont and Tirole (1990b).

[51] Martimort (1999a) shows that it can be optimal to impose more severe restrictions on a regulatory agency over time as the potential for regulatory capture increases.

[52] If competition undermines a regulator's (otherwise substantial) powers to promise attractive returns to the incumbent monopoly supplier, it is possible that competition may be inadvisable when the incumbent monopolist is investing and otherwise satisfying the long-term needs of consumers reasonably well. In extreme cases, unregulated monopoly might be the most effective way to deliver strong investment incentives if the absence of regulation, itself, constitutes a meaningful commitment not to expropriate a monopolist's earnings. Notice that, although a regulator may have particularly pronounced ability to expropriate investors because of his direct control over a broad range of industry policies, other government entities also may have substantial ability to expropriate investors by imposing excessive profit taxes or indirect taxes, for example.

[53] Wesley M. Cohen and Richard C. Levin (1989) review empirical studies of the relationship between market structure and innovation.

[54] See Kenneth J. Arrow (1962). Of course, this argument presumes that intellectual property protection precludes competitors from immediately copying the innovator's discovery and thereby eliminating the financial gain from innovation.

[55] This argument presumes that capital markets are imperfect. This is a reasonable assumption in the context of innovation because innovators often are unable to convince investors of the merits of their potential innovation without revealing the details of the innovation (and thereby forfeiting some of the potential financial gains from the innovation). See, for example, James J. Anton and Dennis A. Yao (1994, 2002).

[56] See Joseph Schumpeter (1950) and Glenn C. Loury (1979), for example.

[57] Guthrie (forthcoming) provides a more complete review of the literature that examines the impact of regulation on infrastructure investment.

[58] A regulatory policy that delivers no extra profit to a firm as its realized production costs decline effectively expropriates any investment the firm might make in an attempt to secure lower production costs.

specific long-term returns on investment. Although such a policy can provide substantial incentive for short-term innovation and cost reduction,[59] it may provide limited incentive for long-term infrastructure investment.[60] Therefore, the choice between rate of return regulation and price cap regulation will depend in part on the type of investment that is most important to secure. In settings where the top priority is to induce the regulated firm to employ its existing infrastructure more efficiently, price cap regulation may be preferable. In contrast, in settings where it is important to reverse a history of chronic underinvestment in key infrastructure, rate of return regulation may be preferable.[61]

The appropriate choice between rate of return regulation and price cap regulation also is influenced by industry volatility and regulatory commitment powers.[62] As costs and demands change over an extended time period, prices and costs will invariably diverge under price cap regulation, possibly leading either to financial distress or to particularly large profit for the regulated firm.[63] Neither of these outcomes is likely to be credibly sustained in settings where the country's institutions are weak. Therefore, in such settings, rate of return regulation may be preferable to price cap regulation in the presence of considerable industry volatility, particularly if infrastructure investment is desirable.[64] The cost-plus nature of rate of return regulation, which ensures profits are neither excessive nor insufficient, can render its implicit commitment to set prices that track costs closely more credible than the price commitments encompassed in a price cap plan.[65]

5. *Franchise Bidding*

Even when pronounced scale economies render direct competition in the market prohibitively costly, competition for the right to serve as a regulated monopoly supplier (in the form of franchise bidding) can sometimes capture for consumers much of the surplus that strong competition in the market would generate.[66] To illustrate the potential value of franchise bidding, return to the simple model considered in section 3 and suppose two firms bid for the right to serve as the sole provider of the product in question under terms specified by the regulator. Each firm is

[59] The regulatory policy implemented for British Gas in 1986 (which permitted the average price of natural gas to increase at the rate $RPI + Y - X$) is an example of price cap regulation. Sappington (2002) provides an overview of price cap regulation plans. Also see Jan Paul Acton and Vogelsang (1989), Jordan Jay Hillman and Ronald R. Braeutigam (1989), Armstrong, Cowan, and Vickers (1994), and Jeffrey I. Bernstein and Sappington (1999), for example. Notice that a promise to allow revenues to diverge from realized costs can play much the same role as intellectual property rights protection that allows an innovator to derive substantial financial gain from her innovation.

[60] A. Michael Spence (1975) observes that, when quality cannot be regulated directly, rate of return regulation may provide stronger incentives to supply quality than price cap regulation. Sappington (2005a) reviews the literature on service quality regulation in utility industries.

[61] Regulatory policies that deliver a consistent rate of return on investment also can reduce the variance in investors' financial returns, and thereby reduce the regulated firm's cost of capital.

[62] Alternatives to these two forms of regulation also merit consideration. As Schmalensee (1989) notes, earnings sharing plans (which specify explicit rules for sharing the firm's realized earnings with consumers) can help to avoid extreme distributions of rent that can prove difficult to enforce in practice. Sappington (2002) provides a discussion of earnings sharing plans and relevant references to the literature. Vogelsang (2005) explains how earnings sharing policies might be coupled with other regulatory regimes to provide appropriate short-term incentives for cost reduction and long-term incentives for investment.

[63] This is the case even if the price cap plan makes explicit adjustments for changes in economywide output prices and changes in key input prices.

[64] Of course, if a regulator has considerable discretion over which of the firm's investments qualify for the promised rate of return, serious problems of opportunism can arise under rate of return regulation, just as they can arise under price cap regulation.

[65] Guasch (2004, table 1.16) reports more widespread renegotiation of franchise contracts under price cap regulation than under rate of return regulation.

[66] Harold Demsetz (1968) provides the seminal discussion of this observation. Laffont and Tirole (1987), R. Preston McAfee and John McMillan (1987), and Riordan and Sappington (1987), among others, provide formal analyses of the optimal design of monopoly franchises.

privately informed about its own marginal cost of production ($c \in \{c_L, c_H\}$). The two possible cost realizations again are equally likely for each firm. Furthermore, for simplicity, the firms' cost realizations are uncorrelated.[67]

The optimal franchise bidding policy in this setting takes the following form. The two firms report their cost realizations simultaneously. If only one firm reports the low cost realization (c_L), that firm is selected to be the monopoly supplier, sell the product at unit price p_L, and receive transfer payment T_L. If both firms report low costs, one of the firms is selected at random to operate the (p_L, T_L) contract. If both firms report high costs (c_H), one of the firms is selected at random to operate under the (p_H, T_H) contract.

The contracts are optimally designed to ensure the firms report their costs truthfully. Consequently, a firm with the low marginal cost will be selected to serve as the monopoly supplier with probability 3/4. (Thus, franchise bidding secures the sampling benefit of competition.) A firm with the high marginal cost will operate with probability 1/4. Truthful reporting by the low cost firm will be a best response to truthful reporting by the rival potential producer if

$$(7) \qquad \frac{3}{4} U_L \geqq \frac{1}{4}[U_H + \Delta q(p_H)],$$

where $U_i \equiv [p_i - c_i]\, q(p_i) - F + T_i$ for $i = L,H$. The binding participation constraint will be that of the firm with the high marginal cost, so U_H is optimally held to zero. Inequality (7) will optimally be satisfied as an equality to limit the rent afforded a firm with the low marginal cost. Consequently, expected welfare under the optimal pricing policy is:

$$(8) \qquad \frac{3}{4}[w_L(p_L) - \frac{1}{3}[1 + \lambda - \alpha]\Delta q(p_H)] + \frac{1}{4} w_H(p_H).$$

Expression (8) reveals that p_H is optimally chosen to maximize $w_H(\cdot) - [1 + \lambda - \alpha]\Delta q(\cdot)$, while p_L is optimally chosen to maximize $w_L(\cdot)$. Therefore, prices in this franchise bidding setting are the same prices that are implemented in the regulated monopoly setting. (Recall expression (2).) The transfer payment (T_L) to the low-cost supplier is reduced in the franchise bidding setting, though. The firm will reveal its superior capabilities despite being promised a smaller transfer payment because cost exaggeration entails substantial risk of being excluded from the industry. (Thus, franchise bidding secures the rent-reducing benefit of competition.)

Because franchise bidding and monopoly regulation implement the same price for a given cost realization, total expected welfare is the same under the two regimes except that the likelihood of a low-cost supplier increases from 1/2 to 3/4. In the case where there is no cost of social funds (so $\lambda = 0$), expected welfare in the franchise bidding setting is:

$$(9) \qquad \frac{3}{4} v(c_L) + \frac{1}{4} v(c_H + [1 - \alpha]\Delta).$$

It is apparent that expression (9) exceeds expression (5). Expression (9) also exceeds expression (4) when $\rho = 1/2$.[68]

In summary, franchise bidding outperforms both monopoly regulation and duopoly competition in this simple setting.[69] It

[67] The regulator can achieve the ideal (full-information) outcome in this setting if the firms' costs are correlated. See, for example, Joel S. Demski and Sappington (1984) and Jacques Crémer and Richard P. McLean (1985).

[68] This is the case because $v(c_H + [1 - \alpha]\Delta) \geq v(c_H) - [1 - \alpha]\Delta q(c_H)$. Therefore, expression (9) will exceed expression (4) if $2[v(c_L) - v(c_H)] \geq \Delta q(c_H)[1 + \alpha]$. This inequality holds for all values of $\alpha \leq 1$.

[69] In practice, franchise auctions often take different forms. For example, firms often are invited to specify a price at which they are willing to supply the product in question to consumers, and the firm that bids the lowest price might be awarded exclusive production rights. This auction, which involves no transfer payments from firms to the government (or vice versa), typically will result in the most efficient firm serving consumers at a price that reflects its average cost of production. Such an auction resembles normal price competition but has the added advantage that it can permit profitable operation even in the presence of scale economies. Alternatively, the supply price might be determined in advance and firms could bid on the amount they will pay the government (or the payment they will require from the government) for the right (and obligation) to provide the service at the stipulated price.

does so by employing both transfer payments and price regulation to pursue social goals by ensuring the benefit of scale economies through the selection of a single supplier and by securing the sampling and rent-reducing benefits of competition.

More generally, franchise bidding can have its drawbacks. If potential operators do not have comparable skills, accurate information, and substantial financial resources, bidding for the right to serve as the monopoly supplier may not be intense. Firms will have little interest in bidding for a contract they expect to lose to a more capable operator. In addition, a firm will be reluctant to bid aggressively on a contract if it may win the contract not because it is the most capable producer but because it underestimates most severely the true (common) cost of operating the franchise.[70] A firm may be similarly reluctant to bid for a contract when it suspects other bidders (e.g., an incumbent provider of the service in question or a firm that provides similar services in another geographic region) have better information regarding the financial returns the contract is likely to provide. Furthermore, a firm cannot bid more for a contract than its financial resources allow.[71] When pronounced differences in skills or information and/or a dearth of qualified bidders with substantial financial resources limit the intensity of franchise bidding, such bidding will not capture for consumers the surplus they would enjoy if multiple, well-informed firms with similar capabilities competed against each other in the marketplace.[72]

Franchise bidding also may provide opportunities for the selected producer to "hold up" the government. After it is awarded the monopoly franchise, the chosen producer may attempt to renegotiate the terms of the original contract in order to secure more favorable terms.[73] The government may be susceptible to renegotiation demands in order to avoid the appearance of failure in the procurement process or to avoid substantial transaction costs associated with reauctioning the franchise.[74]

Franchise bidding also can fail to provide ideal incentives for investment. If the duration of the franchise contract is short relative to the useful life of a desirable sunk investment, the chosen supplier may be reluctant to undertake the investment if the firm's tenure as the monopoly supplier is likely to

[70] This possibility is known as the "winner's curse." See Vijay Krishna (2002, pp. 84–85), for example.

[71] The terms of the franchise contract can be altered to counteract the problems that arise when bidders have limited financial resources. For example, rather than bid on a lump-sum franchise fee, potential operators might bid on the fraction of realized profit (or revenue) that they are willing to share with their customers. See Tracy R. Lewis and Sappington (2000) for details. Che and Ian Gale (1998, 2000) provide related analyses.

[72] A decision to exclude foreign investors from the bidding process (to promote nationalism or to further national security, for example) can limit substantially the intensity of franchise bidding. Such exclusion also can reduce the value of the franchise since the skills and expertise of the successful bidder affect the profit the enterprise can generate. Roman Frydman, Cheryl Gray, Marek Hessel, and Andrzej Rapaczynski (1999) report that enterprise revenues increase substantially when an SOE is sold to outside managers, but not when it is sold to individuals who managed the SOE before its privatization.

[73] Guasch (2004) analyzes more than 1,000 franchise auctions in Latin America and the Caribbean between 1985 and 2000. He finds that more than 50 percent of electricity franchise contracts and 75 percent of water franchise contracts were renegotiated. The average time between franchise award and renegotiation was approximately two years. Renegotiation was initiated by the chosen operator more often than it was initiated by the government.

[74] When all relevant dimensions of performance are not specified clearly in a regulatory contract, a potential supplier may be able to bid the highest franchise fee, not because it is the least-cost supplier of the services in question, but because it will deliver the least on all of the performance dimensions that are not specified in the regulatory contract. Alejandro M. Manelli and Daniel R. Vincent (1995) identify conditions under which considerations of this sort render it optimal for a regulator to negotiate a contract with a single selected supplier rather than to award the franchise to the firm that bids the most to operate under a contract that does not specify fully all relevant dimensions of performance.

end before it can fully recover the cost of the investment. Consequently, the chosen supplier may adopt an inefficient production technology (one that employs an unduly small level of sunk costs) and all potential suppliers may reduce their bids for the right to serve as the monopolist.

Although long-term contracts can, in principle, help to overcome this problem, long-term contracts seldom are a panacea. In practice, it typically is impossible to delineate all relevant contingencies in a contract.[75] Furthermore, rigid contracts can preclude valuable adaptations to changing industry conditions. In addition, even the most carefully crafted long-term contracts may not be enforced if the prevailing legal institutions are weak. For all these reasons, even long-term contracts typically are unable to deliver ideal investment incentives. In some settings, it may be possible to adjust franchise bidding policies to counteract some of these problems.[76] For example, auction rules that favor the incumbent supplier can enhance the incumbent's incentive to undertake sunk investments in the presence of short-term contracts.[77]

In summary, auctions for the right to be the sole supplier can help to limit monopoly rents and to select the most efficient industry supplier. However, franchise auctions seldom eliminate the need for regulation. Moreover, the same strong regulatory institutions that are necessary for effective monopoly regulation are required for the success of franchise bidding contracts. When these institutions are present, though, franchise bidding can constitute a useful additional instrument that regulators can employ to select and discipline sole providers in network industries.[78]

In concluding this section, we note that yardstick competition can secure many of the same benefits as franchise bidding in settings where different monopolists operate in distinct geographic markets. (For example, different water distribution companies often serve different regions of a country.) In such settings, each monopolist might be compensated on the basis of how its performance compares to the performance of other monopolists. For example, the compensation delivered to each firm might be set equal to an index of the realized costs of the other firms, rather than its own costs. This and other forms of yardstick competition can provide strong incentives for efficient performance by all monopolists when they are known to operate in similar settings. When the firms operate in environments that differ substantially (in geographic area, terrain, weather, or population density, for example), explicit corrections for relevant differences can be important. Such corrections typically will be necessary to avoid compensation that is unduly generous for some firms and unduly meager for others. Appropriate handicapping can be difficult in the presence of limited information about the precise nature of the variation in the firms' operating conditions. However, some relative performance comparisons generally can

[75] See Williamson (1976), for example.

[76] Robin A. Prager (1989), Mark A. Zupan (1989a, 1989b), and Yasuji Otsuka (1997) present evidence which suggests that these difficulties with franchise competition are not always insurmountable in practice.

[77] See Laffont and Tirole (1988). Investment also might be encouraged by adopting technologies that employ more fungible physical assets (e.g., wireless rather than wireline facilities in the telecommunications industry).

[78] Regulators can sometimes benefit by ensuring that alternative suppliers are available to replace the incumbent supplier as needed. Although it can be costly to maintain a second source of production, the ability to readily shift some or all production to the second source can impose useful discipline on a monopoly supplier. See, for example, Rafael Rob (1986), Anton and Yao (1987), Demski, Sappington, and Speller (1987), Riordan (1996), Riordan and Sappington (1989), James D. Dana, Jr. and Kathryn E. Spier (1994), and Anton and Paul J. Gertler (2004).

help to discipline and motivate monopoly suppliers.[79]

6. *Entry Assistance and Anti-Competitive Liberalization Policies*

The foregoing discussion suggests the greatest potential gains from competition will tend to arise when: (1) industry scale economies are limited relative to consumer demand; (2) the industry regulator has limited information, limited resources, and limited instruments with which to craft policy; (3) the regulator's commitment powers are limited; and (4) subsidization of the consumption of some of the dominant supplier's services is either not critical or can be achieved by means other than regulating the supplier's price structure.

The design of liberalization policy is of paramount importance in settings where competition is deemed to be a superior alternative to a prevailing monopoly regime. The purpose of this third segment of the present essay is to discuss the principles that underlie the design of sound liberalization policy.[80] After considering the subtle issue of explicit entry assistance, this section focuses on liberalization policies that can hinder (rather than promote) vibrant, long-term industry competition. Section 7 reviews preferable competition-enhancing liberalization policies.

It is important to emphasize at the outset that the discussion in both this section and the next necessarily entails some subjective judgments. Furthermore, even though the policies considered in this section generally are not recommended, some of the policies can, in theory, enhance welfare in certain settings if the regulator is particularly well informed. Therefore, definitive, unequivocal conclusions about liberalization policies are difficult to draw. The ensuing discussion is intended to provide general guiding principles (to the extent possible) rather than precise, comprehensive prescriptions and definitive conclusions.

Successful liberalization is seldom as simple as removing all legal restrictions on entry into the regulated industry. Entrants face myriad economic barriers to entry, even when legal barriers are removed. These entry barriers include (1) customer inertia due to switching costs or ignorance, for example; (2) incumbent control of key inputs that entrants require for profitable operation; and (3) the prospect of aggressive pricing by incumbent suppliers. Policies that reduce or limit the effects of entry barriers constitute vital components of a successful liberalization plan, as explained in section 7. However, is the reduction of entry barriers sufficient, or should liberalization efforts include additional policies that provide explicit assistance to competitors?

Although such direct assistance has many important drawbacks (as explained further below), it can in principle increase total welfare under some conditions. To illustrate, suppose competition would drive prices close to the entrant's marginal cost of production, as it would, for example, in the setting of section 3.2 when the competitors' costs are highly correlated. In such a setting, an entrant will anticipate limited variable profit from industry operations. Consequently, if the entrant must incur a substantial sunk cost in order to enter the industry, the potential competitor will decline to enter, even if it faces no legal restrictions on entry. Therefore, entry may be least likely to occur without assistance

[79] See Andrei Shleifer (1985), Joel Sobel (1999), Roger Carrington, Tim Coelli, and Eric Groom (2002), and Mehdi Farsi, Massimo Filippino, and William Greene (2005), for example. Armstrong and Sappington (forthcoming, section 4.1) provide further discussion and references. Notice that the entire benefit of yardstick competition arises from the rent-reducing impact of operation by multiple firms. Yardstick competition does not admit a sampling benefit because the firm that supplies a given market is fixed exogenously.

[80] This discussion draws in part from Armstrong, Cowan, and Vickers (1994, chapter 4).

precisely when, due to its intensity, competition is most effective relative to regulation.[81] In such settings, entry assistance (perhaps in the form of a subsidy financed by public funds to cover the sunk entry cost, for example) could increase social welfare in theory. However, the appropriate magnitude of such entry assistance can be difficult to determine in practice. Furthermore, such assistance can promote regulatory capture and unproductive use of public funds.

The general merits of entry assistance depend in part on the external effects of entry. External effects refer to the effects of entry on consumer surplus and the profits of incumbent producers. External effects may cause entry to be unprofitable for an individual firm even though entry would increase consumer surplus and total welfare, as in the setting just described. External effects also may cause entry to be profitable for an individual firm even though entry would reduce total welfare. Such a setting (in which it can be more appropriate to discourage entry than to assist entry) arises naturally when the profit an entrant anticipates from industry participation is derived in part from profit that would otherwise accrue to incumbent producers.[82]

Because the most appropriate level of direct entry assistance can be difficult to determine and because such entry assistance can have undesirable consequences,[83] it generally is not recommended. The following liberalization policies (some of which entail entry assistance and some of which fail to reduce entry barriers adequately) also generally are not recommended.

6.1 *Provide a Temporary Monopoly or Oligopoly*

A common form of liberalization is to announce a future (sometimes distant) date at which competition will be admitted, but explicitly preclude widespread competition before that date.[84] A temporary monopoly (or oligopoly) policy of this sort often is adopted in part to secure the incumbent supplier's support for liberalization. Such support may be important, for example, when the incumbent supplier is a state-owned enterprise with substantial authority to set industry policy or when the incumbent supplier employs a large labor force and so has substantial political power.

In principle, a temporary monopoly policy can have some merit. Intense competition can increase the risk of and limit the return from investment, and thereby reduce the investment of the incumbent supplier. Therefore, in principle, a temporary monopoly could increase the incumbent supplier's investment, to the benefit of consumers.

However, the temporary monopoly also postpones investment by new suppliers, and so may reduce aggregate investment in both the short run and the long run. Furthermore, investment by new entrants may spur retaliatory or defensive effort by the incumbent supplier.[85] This added impetus for investment by an incumbent supplier is eliminated in the short run by a temporary monopoly policy, and potentially reduced by a temporary oligopoly policy. A temporary oligopoly policy also can limit entry to a few selected competitors. Such selection runs the risk of allowing less efficient suppliers to

[81] Entrants may require little assistance when they are competing against an inefficient incumbent supplier that has historically delivered poor service quality to consumers. Consumers often will be anxious to secure service from an alternative supplier under such conditions.

[82] Unassisted entry by the first firm in a market generates positive external effects in that market. The first producer causes consumer surplus to increase and does not affect the profits of other producers.

[83] The information required to assess the magnitude of relevant external effects includes the details of consumer demand, the extent and nature of product heterogeneity, and the production technologies of actual and potential producers. See N. Gregory Mankiw and Michael D. Whinston (1986) for additional analysis of this issue.

[84] Recall that British Gas was permitted to retain its monopoly status (albeit in continually shrinking market segments) during the initial stages of liberalization.

[85] See Woroch (2000) for evidence of this effect.

compete, while excluding the most efficient suppliers from the market. A priori restrictions on the number of competitors also can limit product innovation, facilitate collusion among the selected incumbent producers, and limit industry pressure to reduce operating costs and prices.

A temporary duopoly is sometimes suggested as a means to assist at least one additional industry participant on the grounds that no entry would occur if an entrant were not promised the security of a temporary duopoly. This rationale is not entirely compelling. If a second entrant would render the first entrant unprofitable, the second entrant seems unlikely to enter unless it is more efficient than the first entrant. Furthermore, an incumbent producer may benefit greatly from entry assistance of this type, which suggests such assistance may be poorly targeted.

An additional potential drawback to a temporary monopoly is that it may provide ample time for an incumbent supplier to devise ways to limit the likely success of future competitors. In addition, "temporary" monopolies can become permanent or semipermanent monopolies in settings where the incumbent supplier has substantial political power and where the regulator's commitment powers are limited.

The welfare losses from temporary monopolies and oligopolies can be substantial. In the United States, for example, service quality increased and prices for wireless telecommunications services declined substantially once the relevant markets were opened to more than two providers.[86] During the mandated duopoly period, however, industry prices remained close to their monopoly levels.[87] Similarly, the prices of wireline telecommunications services did not decline substantially in Australia during its mandated duopoly period. In contrast, prices dropped significantly in countries (like Chile and Guatemala, for example) that did not adopt duopoly policies.[88]

6.2 *Exclude Foreign Investors*

Feelings of nationalism and concerns with national security can lead to the imposition of limits on foreign participation in key domestic industries.[89] Although such limits can reduce foreign control of domestic industry, they can impose significant costs on the domestic economy. Limits on foreign ownership can reduce the flow of much-needed capital to the domestic industry. The limits can also serve to exclude the most knowledgeable and experienced operators from the industry, and thereby reduce industry performance. In cases where operating licenses are sold by a domestic government, limits on foreign participation can also reduce the license revenue that accrues to the domestic government.

6.3 *Specify Market Share Targets*

In an attempt to ensure adequate competitive pressure in a liberalized industry, regulators sometimes specify the market share they would like to see competitors achieve.[90] Such a policy suffers from at least two major drawbacks. First, market share is not necessarily a good measure of market power. In particular, an incumbent supplier may have substantial ability to raise the prices it charges for its services even when competitors presently serve a significant portion of customers. This ability to raise prices profitably may stem from the incumbent's superior products or control over key inputs, for example. Consequently, the specification of a target market share for competitors may

[86] See Federal Communications Commission (2003).

[87] See Philip M. Parker and Lars-Hendrik Röller (1997).

[88] See Spiller and Cardilli (1997), for example. The United Kingdom adopted a seven year duopoly policy in its telecommunications industry. Newbery (1999, table 7.4) reports that the productivity of the incumbent producer in the United Kingdom improved significantly only when the duopoly policy was relaxed.

[89] These concerns may be particularly strong in countries where colonial rule has only recently ended.

[90] Recall that target market shares were specified in the U.K. natural gas industry.

not ensure that substantial discipline is imposed on the incumbent supplier.[91]

Second, the specification of market share targets can reduce the intensity of market competition. An incumbent supplier may choose not to compete aggressively against a rival in order to ensure that the rival achieves its market share target. When market share targets are specified, the incumbent may rationally refrain from aggressive competition, recognizing that it may cause the regulator to impose more stringent regulations on the incumbent when competitors fail to achieve their market share targets. The resulting diminution in competitive intensity can result in higher prices and lower service quality for customers and can enable competitors to survive in the marketplace even though they have higher costs than the incumbent supplier.

6.4 *Implement Vague or Excessively Generous Network Access Policies*

Many network industries entail a vertical structure that contains a massive network infrastructure segment (e.g., a basic telecommunications network or a gas or electricity transmission system). The incumbent supplier of retail services often owns or operates the infrastructure segment. Given the prohibitive cost of constructing multiple infrastructures, rival retail operators often are compelled to procure key inputs (e.g., network access) from the incumbent, vertically integrated supplier. The appropriate design of the terms that govern competitors' access to the incumbent's network poses a formidable challenge for regulators. Some undesirable (but not uncommon) elements of network access policy are mentioned here. More desirable elements of network access policy are reviewed in section 7.

One inappropriate element of network access policy is a vague or incomplete statement of the incumbent producer's obligations to supply access to rivals during the liberalization process. If, for example, an incumbent producer is afforded substantial latitude in setting the terms and conditions of network access, the incumbent should be expected to employ this latitude to disadvantage retail rivals. For example, the incumbent should be expected to set high access prices, limit the quality of the inputs delivered to rivals, and delay the provision of access.[92]

A second inappropriate element of network access policy is the failure to establish a timely, functional dispute resolution process. When they operate as rivals at the retail stage of production, the incumbent producer and retail competitors naturally will have divergent interests regarding the terms and conditions of access. These divergent interests often will lead to disputes, even when the incumbent's obligations are reasonably well specified. Prompt resolution of these disputes is necessary to ensure the timely implementation of the liberalization process. By specifying clearly the details of a rapid dispute resolution mechanism, a regulator can help to ensure that an incumbent producer does not disadvantage competitors

[91] It is also possible that an incumbent's market share might overstate the firm's market power. For instance, consider Bertrand price competition in a setting where one firm has slightly lower unit cost than its rivals. In equilibrium, this firm will have a 100 percent market share even though it has limited ability to raise the market price.

[92] Recall that competing suppliers of natural gas enjoyed little success until the terms of access to British Gas's infrastructure were explicitly regulated. Weisman (1995), Nicholas Economides (1998), David Reiffen (1998), David M. Mandy (2000), T. Randolph Beard, David L. Kaserman, and John W. Mayo (2001), and Mandy and Sappington (forthcoming), among others, analyze the incentives of a vertically integrated supplier to disadvantage retail competitors. Of course, a vertically integrated incumbent supplier may not wish to disadvantage a retail competitor if the competitor is substantially more efficient than the incumbent and if the incumbent can secure sufficiently high profit margins on the inputs it sells to the retail competitor. See, for example, Sibley and Weisman (1998) and Rey and Tirole (forthcoming).

by delaying unduly their access to key network infrastructure.[93]

A third inappropriate element of network access policy is the implementation of terms of access that are unduly generous for entrants. To reduce the costs that new suppliers incur when entering a market, regulators may afford new suppliers access to essential infrastructure at prices that are below the incumbent's costs of supplying the access. Short-term subsidies of this nature can reduce the costs that new suppliers incur as they begin to provide service in a formerly monopolized industry. In doing so, such subsidies may help to attract new competitors to the industry and thereby "jump start" competition.[94]

While short-term subsidies could, in principle, prove beneficial in this regard, subsidies that are intended to be in effect for only a short period of time often remain in effect far longer. Long-term subsidies of this sort introduce at least two important problems. First, the subsidies may permit inefficient firms to operate profitably in the industry, thereby increasing industry costs and reducing industry welfare. Second, subsidized access to infrastructure can induce competitors to employ inefficient operating technologies. In particular, a competitor may decide to employ the subsidized access to the incumbent's infrastructure even though the competitor would employ fewer social resources if it built and employed its own infrastructure. Thus, by distorting the technological choices of competitors, subsidized infrastructure access can increase industry operating costs. It can also reduce product and process innovation in the industry by limiting the extent to which competitors construct their own infrastructures.[95]

6.5 *Restrict the Incumbent Asymmetrically*

Stringent, asymmetric regulation of an incumbent supplier that limits the incumbent's ability to compete against entrants can help to attract entry. However, even when entry is desirable, the costs of asymmetric regulation often outweigh its benefits.

To illustrate this more general principle, consider a policy that embeds cross subsidies in the incumbent supplier's pricing structure and prohibits the incumbent from reducing any of its prices in response to competitive pressures (even prices that exceed production costs by a substantial margin). Such a policy has at least four potential drawbacks. First, the cream-skimming induced by cross-subsidies can jeopardize the financial integrity of the incumbent supplier. It can do so by reducing the incumbent's sales of the profitable services targeted by competitors without any offsetting reduction in the sales of the unprofitable, subsidized services. Second, the cross subsidies can increase industry costs by allowing inefficient suppliers to serve customers. Third, when competitors focus their efforts on selling the most profitable services, customers of unprofitable, subsidized services are denied the benefits of competition.[96] Fourth, as explained more fully in section 7.7 below, the design of appropriate access charges can be complicated considerably when the incumbent supplier's retail tariffs do not reflect the supplier's costs.

[93] Recall that few interconnection agreements were negotiated in Chile until after strict mandates for timely negotiation and an effective dispute resolution mechanism were implemented. Spiller and Cardilli (1997) explain how final-offer arbitration of the type implemented in Guatemala's telecommunications industry can facilitate timely negotiation of interconnection agreements.

[94] See Joshua S. Gans and Stephen P. King (2004) for an analysis of this issue. Of course, if access prices are unduly high, competitors may build their own infrastructure even though short-term and long-term industry costs would be minimized if the competitors employed the incumbent's infrastructure.

[95] See, for example, Hausman (1997), Hausman and J. Gregory Sidak (1999, 2005), Gregory Rosston and Noll (2002), and the discussion in section 7.7 below.

[96] Recall that uniform retail prices and distance-based access charges to British Gas's transport pipeline may encourage competitors to serve customers located close to gas landing points.

As one further illustration, consider an asymmetric policy that precludes an incumbent supplier from expanding its product line (e.g., an incumbent supplier of telephone service might be precluded from providing cable television service). It is apparent that such regulations can lead to higher prices, reduced product variety, and lower product quality. When a strong, viable competitor is precluded from a market, the remaining operators may not be compelled to compete as aggressively as they otherwise might. Consequently, prices may rise, superior products may be introduced less frequently, and the quality of existing products may decline.[97] Consumers are harmed by each of these outcomes.

In summary, although restrictions on the scope of an incumbent supplier's operations do not necessarily reduce welfare (see section 7.6 below), they have substantial potential to do so if they are not applied judiciously. Similarly, many of the other types of liberalization considered in this section can, in principle, increase welfare under some conditions. However, these liberalization policies generally are not the best way to promote vigorous, long-term industry competition. Preferable policies are considered next.

7. *Pro-Competitive Liberalization Policies*

The purpose of this section is to describe liberalization policies that generally can help to foster vigorous long-term industry competition while avoiding many of the potential drawbacks discussed in section 6. The liberalization policies discussed in this section focus more on removing entry barriers and unleashing the full force of competition rather than on handicapping or favoring certain competitors. Such policies provide increased potential for ultimately relying on market forces rather than on ongoing detailed regulatory oversight to ensure that consumers are well served.

7.1 *Reduce Customer Switching and Search Costs*

Competition can compel providers to deliver high-quality products to consumers at low prices if consumers are able to easily identify and secure service from the firms that offer the best products at the lowest prices.[98] Therefore, liberalization policies that help to ensure consumers are well informed and are able to switch their service provider easily can stimulate vibrant, enduring competition that may ultimately substitute for regulatory oversight.[99] Specific policies that can be helpful in this regard include the following three.

First, consumers might be afforded ready access to information about the services that competitors offer. Relevant information includes both price and (objective, verifiable) quality information. If consumers generally have access to the Internet, relevant information might be made available at a government (or government accredited) web site. The web site address might be printed on the bills that customers receive from the incumbent supplier. Price comparisons will be most transparent when the services the firms supply are fairly homogenous (such as gas and electricity). Price comparisons can be less meaningful when

[97] Service quality may also decline when universal service obligations require incumbent suppliers to offer service at uneconomic rates. Reductions in service quality can both reduce the cost of supplying the uneconomic service and reduce customer demand for the service. Both effects increase the profit of an incumbent supplier that is saddled with uneconomic universal service obligations.

[98] See Asher Wolinsky (1997) for a model in which a regulator often allows firms to compete directly for customers rather than specifying market boundaries because customers can better discern service quality than the regulator.

[99] Recall that many consumers continue to purchase natural gas from British Gas, even though it tends to charge more for gas than its competitors. This behavior may be explained in part by consumer ignorance regarding the prices charged by competitors.

industry suppliers offer heterogeneous services (e.g., certain types of telecommunications services). Even when it is impractical to provide comparative information, though, basic information about how to contact competing suppliers can increase consumer awareness of competitive alternatives and thereby enhance industry competition.[100]

Second, steps might be taken to reduce the costs that customers incur when they switch suppliers.[101] For example, providers of local telephone service might be required to install technologies that allow a customer to retain the same telephone number regardless of the supplier from which the customer secures service. Absent such "number portability," a consumer might be reluctant to switch suppliers because the switch would require the consumer to inform all friends and associates of her new number or reprint business cards and stationery on which the number appears, for example.[102]

Third, policies might be implemented to reduce asymmetries in the costs that consumers incur when they choose different suppliers. For example, carrier preselection policies in the telecommunications industry allow customers to designate in advance the long-distance carrier that they would like to complete all of their long-distance telephone calls. When a carrier preselection policy is in place, customers do not need to dial additional numbers or otherwise undertake costly or time-consuming activities in order to direct their long-distance telephone business to their preferred carrier. Consequently, new carriers do not need to convince customers to incur substantial costs in order to win their loyalty. The new carriers need only provide higher quality and/or lower prices than incumbent carriers.[103]

More generally, liberalization policies that provide objective information to consumers about the options available to them, reduce customer switching costs, and limit any differential costs that consumers must incur to obtain service from their preferred supplier can foster vibrant, long-term industry competition.

7.2 *Ensure Adequate Monitoring and Data Reporting*

Accurate information about the activities and capabilities of both incumbent suppliers and new operators is of great value when designing regulatory and liberalization policies. This information is essential in assessing the nature and intensity of industry competition and, thus, the extent to which stringent regulatory controls can be relaxed.

Accurate information about the services that competitors supply is important in order to identify the services on which the incumbent supplier might reasonably be afforded more substantial pricing flexibility. Market concentration (as measured by the Herfindahl Index, for example) can sometimes serve as an imperfect indicator of the intensity of market competition. Information about the installed capacity of competitors can be of greater value in assessing both the current and likely future intensity of market

[100] Models of unregulated oligopoly (e.g., Hal R. Varian 1980) often predict that the average equilibrium price in an industry declines as more consumers become informed about the prices competitors charge. Thus, policies that make some consumers more aware of prevailing prices can benefit all consumers.

[101] See Joseph Farrell and Paul Klemperer (forthcoming) for an overview of the economics of switching costs.

[102] Of course, the benefits of policies like number portability must be weighed carefully against the associated costs (e.g., the costs of designing and installing the requisite technology).

[103] Michael Waterson (2003) discusses the potential merits of requiring customers to choose their supplier annually, along the lines of annual policy renewals that are common in the insurance industry. While such a requirement can reduce the differential costs that customers incur in choosing to receive service from a new entrant rather than continuing to receive service from the incumbent supplier, the policy increases transactions costs for all consumers, not only for those who switch suppliers. The same is true of Chile's equal access policy, which, recall, requires callers to specify their preferred carrier every time they place a long distance call.

competition, as can detailed knowledge of the entry barriers that competitors face.

To ensure the timely availability of the information required to formulate appropriate liberalization policies, it is important to establish data reporting requirements. The requirements should specify clearly what data must be reported, how often it must be reported, and which entities must report the data. The data reporting requirements should be changed as infrequently as possible, so as to limit the costs imposed on the reporting entities (and to provide a consistent time series of data).[104]

7.3 *Privatize State-Owned Enterprises*

When the incumbent supplier is owned (wholly or primarily) by the government rather than by private investors, privatization of the state-owned enterprise (SOE) can be an important element of a successful liberalization policy. By hardening soft budget constraints and promoting profit maximization, privatization of an SOE (i.e., selling the firm to private investors) can (1) serve to focus the firm's efforts on reducing its operating costs and delivering high-quality services to consumers and (2) reduce a firm's incentive and ability to engage in below-cost pricing which can preclude the efficient operation of more efficient competitors.[105]

In settings where a government's commitment powers are limited, partial privatization of an SOE may be preferable to full privatization. When the government retains an ownership stake in the firm, the government, like private investors, will suffer financially if it implements policies that reduce the firm's earnings. Consequently, a promise by the government not to expropriate private investors may be more credible when the firm is partially privatized than when it is fully privatized.[106] Widespread domestic distribution of the privately held shares of the (partially) privatized firm can have a similar effect. In the presence of such widespread domestic ownership of the privatized firm, any policy the government might implement that seriously erodes the earnings of the firm would likely evoke widespread, popular opposition. Fearful of the political ramifications of any such widespread opposition, the government will be reluctant to expropriate the privatized firm.[107]

7.4 *Rebalance Retail Tariffs to Better Reflect Costs*

Vibrant long-term industry competition can also be fostered by rebalancing the prices the incumbent supplier charges for its services. Rate rebalancing occurs when prices are aligned more closely with incremental production costs. Ideally, the desired rate rebalancing should be completed at the outset of the liberalization process in order to provide appropriate signals to potential competitors about their likely returns from long-term industry operation. Although prices that diverge from cost can help to achieve desired income distribution, pricing structures that embed cross-subsidies are inappropriate when industry liberalization is under way because they introduce the problems identified in section 6.5.

To avoid these problems, the price charged for each of an incumbent's services should be set at or above the firm's incremental cost of providing the service.[108] Such

[104] Regulators may rationally choose to keep some of the reported data confidential. Confidentiality can protect proprietary business plans and limit undesired information sharing among industry participants.

[105] Sappington and Sidak (2003) note that an SOE's reduced focus on profit can increase its incentives to engage in anticompetitive activities. If, for example, an SOE is more concerned with market share than profit, the SOE may be more willing than a profit-maximizing firm to implement below-cost pricing for an extended period of time.

[106] See, for example, Enrico C. Perotti and Serhat E. Guney (1993), Perotti (1995), and Germa Bel (2003).

[107] See Armstrong, Cowan, and Vickers (1994), Bruno Biais and Perotti (2002), and Bel (2003).

[108] In the presence of scale and scope economies, it may not be possible to set tariffs that reflect marginal production costs without inflicting a financial deficit on the incumbent supplier.

rate rebalancing helps to direct the forces of competition to those services that competitors can supply more efficiently than the incumbent supplier and allow the incumbent supplier to deliver the services that it can supply more efficiently than its rivals.

In settings where rate rebalancing would require dramatic increases in the prices of certain essential services, strong opposition to the rebalancing may arise. To ameliorate this opposition, rates might be rebalanced gradually over time, even though complete rebalancing at the outset of liberalization is preferable, when it is feasible. Alternatively, or in addition, financial support might be provided directly to the customers that would find the price increases to be most burdensome. Such a policy can replace implicit subsidies to all consumers with explicit subsidies to those with the greatest need for financial assistance.[109] If the country's tax system is relatively efficient, general tax revenue can be the best source of funds for the explicit, targeted subsidies in the regulated industry.[110] If the subsidies must be funded entirely within the regulated industry, taxes can be imposed on services that are deemed to be less essential (e.g., the services whose prices were set well above the incumbent's costs prior to liberalization). Any such commodity taxes should apply symmetrically to all suppliers, so as not to distort the competitive process.[111]

7.5 *Allow Adequate, But Not Unlimited, Pricing Flexibility*

If it has no freedom to change the prices it charges for its services, the incumbent supplier will have limited ability to respond to the challenges presented by competitors. Consequently, competitors may survive in the market even if they are less efficient than the incumbent supplier.[112]

Price cap regulation can provide incumbent suppliers some pricing flexibility while limiting undue exercise of market power. Price cap regulation plans typically constrain the rate at which the regulated firm's prices can rise on average, without specifying the exact price that must be charged for any particular service. The application of distinct price cap constraints to distinct groups of services can afford the incumbent supplier some flexibility to respond to competitive challenges without allowing the firm to abuse its market power. To illustrate, the firm might be permitted to increase by five percent annually the average of the prices it charges for services that are supplied by competitors. The firm might also be allowed to increase by one percent annually the average of the prices it charges for services that generally are not supplied by competitors. Pricing restrictions like these that impose different and separate constraints on different groups of services can provide adequate protection for customers who have no competitive alternatives while affording the incumbent supplier a reasonable opportunity to respond to emerging competitive challenges.[113]

In addition, incumbents should generally be precluded from pricing an established service below the incremental cost of providing the service (unless they are required to do so as part of a universal service obligation). As explained above, below-cost pricing

[109] Recall that Chile implemented a program to subsidize the purchase of basic telephone service by citizens in rural regions and low-income urban areas (Laffont 2005, p. 216).

[110] See Atkinson and Stiglitz (1976) and Laffont and Tirole (2000, chapter 6).

[111] Armstrong (2002, section 2.1) discusses the design and implementation of such taxes.

[112] Recall that some observers believe the success enjoyed by new suppliers of local telecommunications services in Chile may be due in part to the limits placed on CTC's ability to reduce prices selectively in response to competitors' prices (Paredes 2005).

[113] Armstrong, Cowan, and Vickers (1994, section 7.5.4) describe how British Telecom used the pricing flexibility it was afforded by price cap regulation to reduce prices substantially on long distance telephone calls (that were also supplied by competitors) while raising prices for local telephone calls (for which there were essentially no other suppliers).

can preclude the operation of more efficient suppliers, and thereby raise industry costs.[114] Below-cost pricing also encourages excessive consumption (i.e., consumption beyond the point at which the marginal benefit of consumption is equal to the marginal cost of production).

It is particularly important to monitor and implement procedures to preclude below-cost pricing by a firm that operates under price cap regulation. This is because such a firm can have a particularly strong incentive to price below cost the services on which it faces particularly intense competition. Under a price cap plan that restricts the rate at which all of the firm's prices can increase on average, a substantial reduction in the price of one service can authorize a pronounced increase in the price of another service. By permitting increases in net revenue on some services that offset corresponding reductions from below-cost pricing on other services, price cap regulation plans can enhance the incentives of an incumbent regulated firm to set predatory prices that can serve to drive its rivals from the market.[115] Therefore, safeguards against below-cost pricing are an important component of price cap regulation plans.

[114] In settings where an incumbent, vertically integrated producer (VIP) otherwise has considerable pricing flexibility, the relevant price floor can include an "imputation" requirement in order to prevent a price squeeze. A price squeeze occurs when a VIP sets a price for a key input that exceeds the price of the VIP's relevant retail service. An imputation requirement obligates the VIP to price its retail service above the sum of the VIP's (downstream) unit cost of producing the retail service and the unit price the VIP charges its retails competitors for the key input. Thus, the lower bound on the price of the VIP's retail product is the usual incremental cost floor, with one exception. The VIP's cost of producing the key input is imputed as the input cost that the VIP imposes on its retail competitors. This price floor is thus the incremental cost the VIP would incur if it faced the same input cost as its rivals. This imputation procedure avoids price squeezes, and thereby helps to ensure that the most efficient providers supply the retail service.

[115] See Armstrong and Vickers (1993).

7.6 *Prevent Disadvantaging of Downstream Competitors*

Vertically integrated producers (VIPs) that both sell essential inputs to retail producers and supply the retail service themselves can disadvantage their rivals through anticompetitive actions other than predatory pricing. For example, consider a setting where a VIP operates under price cap regulation. Suppose the price cap regulation plan imposes a single overarching restriction on the prices of all of the VIP's services, both retail and wholesale services. Under such a restriction, price reductions on retail services effectively authorize price increases on wholesale services. Both price changes are disadvantageous for retail competitors who must purchase the VIP's wholesale services. To limit such disadvantaging of rivals, it generally is advisable to place separate restrictions on the rate at which a VIP's wholesale service prices can rise and on the rate at which its retail service prices can rise.[116]

VIPs can disadvantage retail competitors through means other than strategic pricing. For example, as indicated in section 6.4, a VIP could intentionally reduce the quality of the inputs it delivers to its downstream competitors and thereby limit the competitors' ability to deliver high quality services to their customers.[117] To limit undesirable strategic disadvantaging of rivals, careful monitoring of the quality of the inputs that a VIP delivers to its downstream rivals can be advisable, as can an explicit schedule of penalties for unduly low service quality and, perhaps, rewards for exceptionally high levels of service quality.[118] Structural separation of the

[116] Laffont and Tirole (1996) note that when wholesale and retail services are treated as separate baskets of services in this manner, a VIP's ability and incentive to implement Ramsey prices may be reduced.

[117] Reiffen, Laurence Schumann, and Michael R. Ward (2000) and Paul R. Zimmerman (2003) provide some evidence of this effect.

[118] Lisa V. Wood and Sappington (2004) discuss the design of such reward structures.

VIP's wholesale and retail operations can facilitate accurate monitoring of the relative prices and qualities of the inputs that the VIP supplies to its own retail affiliates and to competing retailers.[119] In assessing the merits of structural separation, this potential benefit should be weighed against any associated costs, including foregone economies of scope.

A more radical alternative to structural separation is ownership separation. Under structural separation, wholesale and retail operations are physically separated but common ownership and control of both operations is permitted. In contrast, ownership separation precludes common ownership and control of wholesale and retail operations. While structural separation can help to limit the ability of a VIP to disadvantage its retail competitors, ownership separation can serve to reduce the incentive a producer of key wholesale services might have to disadvantage the (retail) firms that purchase the wholesale services. When it does not provide retail services, the wholesale provider does not secure any direct gains in the retail market from disadvantaging retail producers. In fact, a reduction in the quality of the inputs it sells generally will reduce the demand for these inputs and thereby reduce the revenues of the wholesale producer. Consequently, by eliminating the key source of potential gains from disadvantaging downstream operators, ownership separation can substantially reduce, if not eliminate, incentives for such disadvantaging.[120]

Of course, the costs of ownership separation must be weighed carefully against its potential benefits. Ownership separation, like structural separation, may sacrifice substantial economies of scope and economies associated with integrated planning of wholesale and retail operations. Ownership separation may also preclude a particularly efficient competitor (the producer of wholesale services) from participating in the retail market and thereby raise industry costs and retail prices. Ownership separation also can entail substantial divestiture costs in cases where a VIP is already providing both wholesale and retail services when ownership separation is considered.[121] The merits of ownership separation also can depend on the prevailing regulatory institutions. A regulator with ample resources, considerable experience, and substantial authority to collect relevant data may be able to prevent a VIP from disadvantaging its rivals. In contrast, a regulator with limited resources, limited experience, and limited information might be unable to do so, rendering vertical separation the best (albeit a costly) way to promote effective competition.

7.7 *Establish Appropriate Access Prices*

The prices that retail competitors must pay to access the infrastructure of an incumbent VIP can have a substantial impact on welfare. A comprehensive assessment of the effects of access prices on welfare is complex. However, some of the key effects can be illustrated in the following simple setting.[122]

Suppose an incumbent VIP incurs marginal cost c_1 in supplying a retail service and

[119] Hausman and Sidak (2005) review the form of structural separation currently required by Ofcom, the U.K. communications regulator. The separation includes the creation of a separate division of British Telecom to supply wholesale services on the same terms and conditions to all retail competitors and an "equality of access" board to oversee the operations of this division.

[120] Vertical divestiture of the incumbent also can simplify the design of access charges by eliminating the effects of access charges on the incentives of a VIP to disadvantage retail competitors. Beard, Kaserman, and Mayo (2001), among others, consider how access charges affect the incentives of a VIP to disadvantage rivals.

[121] See Vickers (1995), Paul J. Hinton, J. Douglas Zona, Schmalensee, and William Taylor (1998), Sand Hyup Lee and Jonathan H. Hamilton (1999), Robert Crandall and Sidak (2002), Stefan Beuhler, Dennis Gärtner, and Daniel Halbheer (2005), and Crew, Kleindorfer, and John Sumpter (2005), for example, for analyses of these and related considerations.

[122] See Armstrong (2002) and Vogelsang (2003) for more detailed surveys of this topic.

marginal cost c_2 in supplying a network input to a single retail competitor. This competitor, the entrant, incurs constant unit cost c in converting one unit of the access service into one unit of its own retail service. (The production technology is fixed proportions, so exactly one unit of the access service is required to produce each unit of the retail service.) The retail services of the incumbent and the entrant are homogeneous, and consumers purchase the product from the firm that sets the lowest retail price.

Suppose the regulator sets retail price p and access price a for the incumbent's services in this setting. Then the entrant will find it profitable to operate if the highest price (p) it can charge for its retail service exceeds its unit cost ($a + c$) of producing the service. Industry production costs are minimized when the least-cost supplier serves all retail customers. The entrant is the least-cost supplier in this setting when ($c_2 + c$) is less than c_1. Therefore, for fixed regulated retail price p, the following access price ensures the entrant will operate when and only when it is least-cost industry supplier:[123]

$$a = c_2 + [p - c_1]. \tag{10}$$

The access price in equation (10) reflects the efficient component pricing rule (ECPR), which states that the access charge should be set equal to the sum of the incumbent VIP's cost of supplying access (c_2) and the opportunity cost (or lost profit, $p - c_1$) the VIP incurs when it loses a unit of business to the entrant.[124] The ECPR implies that when the VIP's regulated retail price exceeds its cost of supplying the retail service (so $p > c_1$), the access charge should exceed the VIP's cost of providing access. This increase in the access charge serves to limit the operation of inefficient entrants. To see why, notice that, if a VIP is obligated to both supply network access at cost and set the price of its retail service above cost ($p > c_1$), an inefficient entrant (i.e., one for which $c_2 + c > c_1$) might find it profitable to enter the market, raising industry production costs. This inefficient entry can be precluded by raising the access charge above cost by the same amount the retail tariff is held above cost ($p - c_1$).[125] Similarly, if the VIP is required to offer a retail service at a price below cost ($p < c_1$), network access can be subsidized accordingly to encourage efficient entry into the subsidized retail market.[126]

Intuitively, one might view the VIP as paying a unit tax of $p - c_1$ when it is required to supply the retail service to consumers at unit price p. To ensure profitable entry only by competitors that are more efficient than the VIP, entrants must pay the same tax the VIP pays. The ECPR implements this tax by raising the access charge above the VIP's cost of supplying access by $p - c_1$.

This simple analysis requires modification in richer settings. Suppose, for example, the retail services of the entrant and the VIP are not homogeneous. In this case, one unit of supply by the entrant does not necessarily reduce the VIP's retail output by exactly one unit, and so the ECPR must be adjusted accordingly.[127] Alternatively, suppose the VIP's retail price is not immutable (perhaps because the VIP operates under a price cap regulation plan that affords the firm some discretion in setting retail prices). In this case, the established access charge will affect

[123] Notice that this discussion takes the incumbent's retail tariff as given. Ideally, wholesale and retail prices should be set simultaneously to achieve social goals, following standard Ramsey principles (Laffont and Tirole 1994).

[124] See Robert Willig (1979) and William J. Baumol (1983) for early discussions of this rule.

[125] This increase in the access charge above the cost of supplying access can help to mitigate the cream-skimming problem discussed in section 7.4 above.

[126] When access charges reflect the ECPR, a VIP is compensated for the profit reduction it incurs when it loses a retail customer to a rival supplier. Consequently, access prices that reflect the ECPR can limit a VIP's incentive to disadvantage retail rivals.

[127] Armstrong, Chris Doyle, and Vickers (1996) provide the details of this and other extensions of the ECPR, along with a more general characterization of optimal access prices.

the VIP's choice of retail tariff. The ECPR requires additional modification to reflect this interaction.[128]

The design of access prices becomes more complicated if an entrant can supply the input itself, by investing in its own infrastructure, for instance. In such a setting, care must be taken to ensure that access charges provide the entrant with appropriate ("make-or-buy") incentives to supply inputs itself rather than purchase them from the incumbent VIP. Two distinct regulatory instruments would be ideal in such a setting. To deliver appropriate make-or-buy incentives to entrants, the VIP's access charge should be set equal to the VIP's cost (c_2) of providing the input. This policy ensures an efficient pattern of production given that entry takes place. Such cost-based access pricing may not ensure industry cost minimization, though, because inefficient retail competitors may find entry profitable or efficient competitors may find entry unprofitable when retail prices diverge from the VIP's production costs ($p \neq c_1$). To ensure efficient entry decisions, cost-based access charges should be accompanied by a tax on the outputs of entrants that reflects the deviation of the VIP's retail price from its cost ($p - c_1$).[129] When the regulator is unable to implement output taxes, the access charge is forced to perform the dual task of providing efficient make-or-buy decisions and efficient entry decisions.[130] Typically, a single instrument cannot achieve two goals, making some compromise inevitable.[131]

In summary, the appropriate design of access charges can require considerable information. Furthermore, the delivery of appropriate incentives for efficient make-or-buy and industry participation decisions can require an extensive set of regulatory instruments (e.g., access prices and output taxes). In settings where a regulator has limited information and limited powers, the design of access charges can present extremely challenging problems. However, these problems can be mitigated if the incumbent VIP's retail tariffs reflect its production costs, as proposed in section 7.4 above. When retail rates reflect costs (so $p = c_1$) in the simple setting considered here, the ECPR in equation (10) collapses to a particularly simple rule: the access charge should be set equal to the VIP's cost of providing access. As noted above, this rule also provides appropriate make-or-buy incentives. Notice that this policy can require little knowledge of consumer demand, and does not require the regulator to be able to control the activities of entrants. In sum, an effective rebalancing of the incumbent VIP's retail tariff greatly simplifies the regulator's task of setting appropriate access charges, and allows access charges to focus on the single task of ensuring appropriate make-or-buy decisions.

Access prices also can affect incentives for network innovation and cost reduction. For example, access prices that provide profit margins on access services that vary inversely with the incumbent VIP's realized access costs can create incentives for the VIP to reduce its costs of supplying access.[132] Clearly, the appropriate design of access charges becomes more complex when access charges serve to motivate cost reduction in addition to securing industry cost minimization and inducing efficient make-or-buy decisions.

[128] See Armstrong (2002, section 2.6) for further discussion of this issue.

[129] See Armstrong (2002, section 2.4) for further discussion of this point.

[130] Sappington (2005b) notes that if retail prices are not regulated, efficient make-or-buy decisions may arise even when access charges do not reflect the incumbent's production costs.

[131] Hausman (1997) and Hausman and Sidak (1999, 2005), among others, note that competitors will rely unduly on an incumbent supplier's infrastructure if access prices are set below the full cost of supplying access. This full cost includes a capital cost that reflects prevailing technological and market risks.

[132] Access prices in the U.S. telecommunications industry reflect estimates of an efficient incumbent supplier would incur in supplying access, which can differ from the actual costs of the incumbent supplier. (See Rosston and Noll 2002, for example.) This pricing policy is intended to induce incumbent suppliers to operate efficiently.

Conceivably, access prices might be set at relatively high levels to systematically encourage entrants to construct their own infrastructure. One reason for doing so (despite the duplicative infrastructure costs such a policy can promote) would be to promote competition among multiple producers in hopes of fostering industry innovation and eliminating the need for long-term regulation of a monopoly supplier of key inputs.[133] However, competition among facilities-based networks will not necessarily eliminate the need for long-term regulation. Even if all operators in the telecommunications industry employ their own infrastructure, each operator must still interconnect with other operators in order to complete calls intended for the customers of other operators. It is conceivable that unregulated negotiations among similarly situated network operators could produce intercarrier compensation arrangements that are broadly consistent with social goals. However, this need not be the case. Firms might negotiate compensation arrangements that inhibit competition or limit new entry, for example.[134]

7.8 *Limit Cost Shifting*

In settings where regulatory controls are relaxed on some, but not all, of the incumbent supplier's operations, it can be important to limit the supplier's ability to engage in cost shifting. Cost shifting occurs when costs that actually are incurred in the production of one set of services (e.g., competitive services) are recorded as costs incurred in the production of a different set of services (e.g., monopoly services). An incumbent supplier may benefit from cost shifting when, for example, its competitive operations are not regulated but the prices it can charge for its monopoly services are linked directly to its measured operating costs. In this situation, the incumbent supplier can secure higher regulated prices without affecting its (pretax) earnings on competitive services by shifting costs from its unregulated to its regulated operations.

The incumbent supplier's ability to undertake such cost shifting can be limited by requiring separate books of account for its regulated and unregulated operations, for example. The firm's incentive to engage in cost shifting can be limited by reducing the extent to which regulated prices are linked to measured operating costs. In particular, price cap regulation can reduce incentives for cost shifting relative to rate of return regulation and other forms of cost-plus regulation.[135]

7.9 *Ensure the Integrated Operation of All Elements of Industry Policy*

It is important to review industry policy in its entirety whenever a major component of industry policy (e.g., liberalization policy) is altered substantially. The review should be designed to ensure the ongoing, effective, integrated operation of all components of industry policy.

The experience in California's electricity industry illustrates the importance of such a comprehensive review. The deregulation of wholesale electricity prices can benefit consumers when a competitive wholesale supply of electricity is available. Encouraging major buyers and suppliers of wholesale electricity to participate in a common exchange system also can provide widespread benefits. Furthermore, a capped retail price for electricity can be a component of a sensible

[133] See Woroch (2002). David Boles de Boer, Christina Enright, and Lewis Evans (2000) note that New Zealand encouraged facilities-based competition among Internet service providers while Australia relied more on regulation to protect consumers. The authors report lower prices in New Zealand than in Australia.

[134] Armstrong (1998) and Laffont, Rey, and Tirole (1998) demonstrate these possibilities formally. Patrick DeGraba (2004) and Sharkey (2004), among others, analyze the effects of different intercarrier compensation arrangements.

[135] Braeutigam and John C. Panzar (1989), Timothy J. Brennan (1990), and Weisman (1993), among others, provide formal analyses of this issue.

regulatory policy. However, even though each of these policies can have individual merit, the combination of these policies can lead to highly undesirable outcomes. As in California, volatile wholesale prices coupled with capped retail prices can jeopardize the financial integrity of retail suppliers of electricity. The problem can become particularly acute when the volatility of wholesale prices is increased by limiting the ability of retail suppliers to negotiate long term contracts for electricity.

Several of the policies cited above can be viewed as corollaries of the general principle that effective, integrated operation of all components of industry policy merit careful examination whenever liberalization is considered. For example, the need to rebalance retail prices and to afford incumbent suppliers appropriate pricing flexibility reflect prudent changes in certain elements of industry policy as other elements change.

7.10 *Increase Antitrust Scrutiny and Enforcement*

A central long-term goal of liberalization is to replace regulatory oversight and control with the disciplining forces of competition. Liberalization might be viewed as a process by which competitive forces are fostered and strengthened to the point where they, alone, can impose effective discipline on some or all of the incumbent supplier's operations. The more pronounced the disciplinary role played by competition, the more important it is to protect competition and ensure it is not subverted by powerful industry participants. This is generally the role of antitrust policy.

Antitrust policy and regulatory policy differ in at least three fundamental respects. First, antitrust policy typically sets guidelines that describe in broad terms acceptable behavior and outcomes. In contrast, regulatory policy often specifies detailed rules (often that apply to particular firms) that define fairly precisely the limits on acceptable behavior and outcomes. Second, having specified ex ante guidelines, antitrust policy typically entails ex post investigations of possible violations of the specified guidelines. Regulation, in contrast, often couples ex ante rules with ongoing industry oversight, rule refinement, and rule enforcement. Third, antitrust policy typically relies on edicts to discontinue anticompetitive behavior and associated fines (often in the form of damage payments to injured parties), while regulation often proscribes specific types of conduct (e.g., price discrimination or expansion into particular markets) and establishes detailed performance requirements and associated reward and penalty structures.[136]

Antitrust and regulatory policy can play vital and complementary roles in the liberalization process. Indeed, it may be important to increase both regulatory oversight and antitrust enforcement as competition develops in the liberalized industry. Thus, the road to deregulation of an industry may not be a straight one. Increased antitrust enforcement and regulatory oversight may both be necessary temporarily to ensure that competition has the opportunity to develop to the point where it can eventually replace regulation as the key source of discipline on the incumbent firm.[137]

In the 1990s, New Zealand decided to rely solely on general competition laws enforced by the courts and by a non-specialized competition authority to govern activities in the telecommunications and electricity sectors. This novel approach was not an immediate failure. However, problems emerged over time, leading the government to realize the need for more orthodox regulatory control. The problems included a heavy case load for the courts and substantial difficulty in proving

[136] See, for example, Marc Bourreau and Pinar Dogan (2001), Sidak (2003), Martin Cave (2004), and Damien Gerardin and Sidak (2005). Dennis Carlton and Randal Picker (2005) note that regulation can allow interested parties greater ongoing intervention in the control process than antitrust.

[137] Ultimately, explicit regulatory control may be replaced in some settings simply by the specter of regulatory intervention. See Amihai Glazer and Henry McMillan (1992), for example.

an incumbent supplier had abused its dominant position. Moreover, legal proceedings proved to move slowly at times when quick decisions were required to keep pace with rapid technological advance. The New Zealand experience suggests that exclusive reliance on competition law may not be sufficient in network industries with emerging competition, even in countries with well-developed case law, strong legal institutions, and substantial judicial resources.[138]

8. *Conclusions*

The foregoing discussion has produced two broad conclusions. First, even the apparently simple choice between regulated monopoly and unregulated competition is not always straightforward in practice. The appropriate choice between these regimes can vary with the relevant technological and demand conditions, with the regulator's skills and resources, with the efficiency of tax systems and capital markets, and with the strength of other prevailing institutions. Second, in settings where competition is preferable to regulated monopoly, the road from monopoly to competition can be a particularly long and winding road. There is no single ideal path from monopoly to competition. The most appropriate liberalization policy depends upon a wide variety of factors, including those relevant to the choice between monopoly and regulation.

The foregoing discussion also has emphasized that, although there is no single liberalization policy that is ideal in all settings, some policies typically are superior to others. Liberalization policies that primarily aid some competitors and handicap others on an ongoing basis can hinder the development of vigorous long-term competition. Therefore, policies such as establishing temporary monopolies or oligopolies, excluding foreign investors, specifying market share targets for industry suppliers, providing entrants with long-term subsidized access to the incumbent's infrastructure, restricting unduly the incumbent supplier's pricing flexibility, and imposing unfunded carrier-of-last-resort obligations exclusively on incumbent suppliers generally are not recommended.

In contrast, liberalization policies that remove barriers to entry and empower consumers to discipline industry suppliers typically are better methods for fostering vigorous long-term industry competition. In particular, policies like reducing customer switching costs, rebalancing the incumbent supplier's prices to better reflect its operating costs, privatizing state-owned enterprises, prohibiting below-cost pricing, and establishing appropriate (access) prices for the use of critical infrastructure generally are recommended.

In addition, careful monitoring of industry operations can be of critical importance during the liberalization process. Accurate, timely data about the nature and intensity of industry competition will allow regulatory policy to adjust quickly to changes in industry conditions. Consequently, although liberalization should ultimately lead to reduced regulatory oversight and control, more pronounced regulatory and antitrust oversight may be required on an interim basis to ensure that regulatory policy is tailored appropriately to the evolving level of competition and that competition is protected.

The road from monopoly to competition is seldom straight and smooth. Detours (e.g., increased regulatory and antitrust scrutiny) may be necessary to ensure safe passage to the intended destination. Bumps in the road (e.g., widespread opposition to rate rebalancing or privatization of state-owned enterprises) should be anticipated in order to facilitate their navigation in as smooth a manner as possible. Thus, the liberalization process can constitute a challenging journey and one that entails considerable uncertainty. Detailed road maps can be of enormous value on such journeys. Unfortunately, though, because every setting in which liberalization might be

[138] Spiller and Cardilli (1997) and Laffont (2005, pp. 200–202).

contemplated is different, road maps that provide sufficient granularity and specificity in every relevant setting are not available. Therefore, one generally must embark on the road to competition armed only with the coarse route markers on which the foregoing discussion has focused.

As emphasized in section 4, the most appropriate liberalization policy can vary considerably according to the institutional setting in which it is being implemented. Therefore, an important role for future research is to develop detailed maps of the best route to competition, i.e., to specify the precise details of liberalization policies that will work well in specific institutional settings. Two approaches to this task seem particularly fruitful. First, the broad principles reviewed in this paper can be translated into precise policy recommendations. (For example, precisely how rapidly should rates be rebalanced, given the prevailing prices, customer wealth, and the operating technologies of the incumbent suppliers and entrants? Alternatively, at precisely what level should access prices be set given the prevailing industry costs structure and retail prices?) Second, case studies and comprehensive empirical work can be undertaken to assess the success and failure of particular liberalization policies in different institutional settings. Together, these two approaches can expand our knowledge of the liberalization policies that will best harness the powers of competition to provide critical industry discipline and perhaps ultimately replace the costly, imperfect discipline that regulatory oversight and control provide.

References

Acemoglu, Daron, Simon Johnson, and James Robinson. 2005. "Institutions as the Fundamental Cause of Long-Run Growth," in *Handbook of Economic Growth, Vol. 1A*. Philippe Aghion and Steven Durlauf, eds. Amsterdam: North-Holland, 385–472.

Acton, Jan Paul, and Ingo Vogelsang. 1989. "Price-Cap Regulation: Introduction." *RAND Journal of Economics*, 20(3): 369–72.

Anton, James J., and Paul J. Gertler. 2004. "Regulation, Local Monopolies and Spatial Competition." *Journal of Regulatory Economics*, 25(2): 115–41.

Anton, James J., and Dennis A. Yao. 1987. "Second Sourcing and the Experience Curve: Price Competition in Defense Procurement." *RAND Journal of Economics*, 18(1): 57–76.

Anton, James J., and Dennis A. Yao. 1994. "Expropriation and Inventions: Appropriable Rents in the Absence of Property Rights." *American Economic Review*, 84(1): 190–209.

Anton, James J., and Dennis A. Yao. 2002. "The Sale of Ideas: Strategic Disclosure, Property Rights, and Contracting." *Review of Economic Studies*, 69(3): 513–31.

Armstrong, Mark. 1998. "Network Interconnection in Telecommunications." *Economic Journal*, 108(448): 545–64.

Armstrong, Mark. 2002. "The Theory of Access Pricing and Interconnection," in *Handbook of Telecommunications Economics, Vol. 1: Structure, Regulation and Competition*. Martin E. Cave, Sumit K. Majumdar, and Ingo Vogelsang, eds. Amsterdam: North-Holland, 295–384.

Armstrong, Mark, Simon Cowan, and John Vickers. 1994. *Regulatory Reform: Economic Analysis and British Experience*. Cambridge: MIT Press.

Armstrong, Mark, Chris Doyle, and John Vickers. 1996. "The Access Pricing Problem: A Synthesis." *Journal of Industrial Economics*, 44(2): 131–50.

Armstrong, Mark, and David E. M. Sappington. Forthcoming. "Recent Developments in the Theory of Regulation," in *Handbook of Industrial Organization, Vol. III*. Mark Armstrong and Robert Porter, eds. Amsterdam: North-Holland.

Armstrong, Mark, and John Vickers. 1993. "Price Discrimination, Competition and Regulation." *Journal of Industrial Economics*, 41(4): 335–59.

Arrow, Kenneth J. 1962. "Economic Welfare and the Allocation of Resources to Invention," in *The Rate and Direction of Inventive Activity*. Richard Nelson, ed. Princeton: Princeton University Press, 609–25.

Atkinson, Anthony B., and Joseph E. Stiglitz. 1976. "The Design of Tax Structure: Direct versus Indirect Taxation." *Journal of Public Economics*, 6(1-2): 55–75.

Aubert, Cecile, and Jean-Jacques Laffont. 2002. "Designing Infrastructure Regulation in Developing Countries," in *Competition Policy in Regulated Industries: Approaches for Emerging Economies*. Paulina Beato and Jean-Jacques Laffont, eds. Washington, D.C.: Inter-American Development Bank; distributed by Johns Hopkins University Press, Baltimore, 1–51.

Auriol, Emmanuelle, and Jean-Jacques Laffont. 1992. "Regulation by Duopoly." *Journal of Economics and Management Strategy*, 1(3): 507–33.

Bakovic, Tonci, Bernard Tenenbaum, and Fiona Woolf. 2003. "Regulation by Contract: A New Way to Privatize Electricity Distribution?" World Bank Energy and Mining Sector Board Discussion Paper No. 7.

Baron, David P. 1989. "Design of Regulatory Mechanisms and Institutions," in *Handbook of Industrial Organization, Vol. 2*. Richard Schmalensee and Robert D. Willig, eds. Amsterdam: North-Holland, 1347–1447.

Baumol, William J. 1983. "Some Subtle Issues in Railroad Regulation." *International Journal of Transport Economics*, 10(1–2): 341–55.

Beard, T. Randolph, David L. Kaserman, and John W. Mayo. 2001. "Regulation, Vertical Integration and Sabotage." *Journal of Industrial Economics*, 49(3): 319–33.

Becker, Gary S. 1983. "A Theory of Competition among Pressure Groups for Political Influence." *Quarterly Journal of Economics*, 98(3): 371–400.

Becker, Gary S. 1985. "Public Policies, Pressure Groups, and Dead Weight Costs." *Journal of Public Economics*, 28(3): 329–47.

Bel, Germa. 2003. "Confidence Building and Politics in Privatization: Some Evidence from Spain." *Economics Letters*, 78(1): 9–16.

Bernstein, Jeffrey I., and David E. M. Sappington. 1999. "Setting the X Factor in Price-Cap Regulation Plans." *Journal of Regulatory Economics*, 16(1): 5–25.

Besley, Timothy, and Stephen Coate. 2003. "Elected versus Appointed Regulators: Theory and Evidence." *Journal of the European Economic Association*, 1(5): 1176–1206.

Biais, Bruno, and Enrico C. Perotti. 2002. "Machiavellian Privatization." *American Economic Review*, 92(1): 240–58.

Blackmon, Glenn. 1994. *Incentive Regulation and the Regulation of Incentives*. Boston: Kluwer Academic Publishers.

Boles de Boer, David; Christina Enright, and Lewis Evans. 2000. "The Internet Service Provider (ISP) Markets of Australia and New Zealand." New Zealand Institute for the Study of Competition and Regulation manuscript. http://www.iscr.org.nz/navigation/research.html.

Borenstein, Severin. 2002. "The Trouble with Electricity Markets: Understanding California's Restructuring Disaster." *Journal of Economic Perspectives*, 16(1): 191–211.

Bourreau, Marc, and Pinar Dogan. 2001. "Regulation and Innovation in the Telecommunications Industry." *Telecommunications Policy*, 25(3): 167–84.

Braeutigam, Ronald R., and John C. Panzar. 1989. "Diversification Incentives under "Price-Based" and "Cost-Based" Regulation." *RAND Journal of Economics*, 20(3): 373–91.

Brennan, Timothy J. 1990. "Cross-Subsidization and Cost Misallocation by Regulated Monopolists." *Journal of Regulatory Economics*, 2(1): 37–51.

Bühler, Stefan, Dennis Gärtner, and Daniel Halbheer. 2005. "Deregulating Network Industries: Dealing with Price–Quality Tradeoffs." Mimeo. University of Zurich.

Burgess, Robin, and Nicholas Stern. 1993. "Taxation and Development." *Journal of Economic Literature*, 31(2): 762–830.

Caillaud, Bernard, Roger Guesnerie, Patrick Rey, and Jean Tirole. 1988. "Government Intervention in Production and Incentives Theory: A Review of Recent Contributions." *RAND Journal of Economics*, 19(1): 1–26.

Carlton, Dennis and Randal Picker. 2005. "Antitrust and Regulation." Mimeo. University of Chicago.

Carrington, Roger, Tim Coelli, and Eric Groom. 2002. "International Benchmarking for Monopoly Price Regulation: The Case of Australian Gas Distribution." *Journal of Regulatory Economics*, 21(2): 191–216.

Cave, Martin. 2004. Economic Aspects of the New Regulatory Regime for Electronic Communications Services, in *The Economics of Antitrust and Regulation in Telecommunications: Perspectives for the New European Regulatory Framework*. Pierre A. Buigues and Patrick Rey, eds. Cheltenham, U.K. and Northampton, Mass.: Elgar, 27–41.

Che, Yeon-Koo. 1995. "Revolving Doors and the Optimal Tolerance for Agency Collusion." *RAND Journal of Economics*, 26(3): 378–97.

Che, Yeon-Koo, and Ian Gale. 1998. "Standard Auctions with Financially Constrained Bidders." *Review of Economic Studies*, 65(1): 1–21.

Che, Yeon-Koo, and Ian Gale. 2000. "The Optimal Mechanism for Selling to a Budget-Constrained Buyer." *Journal of Economic Theory*, 92(2): 198–233.

Cohen, Wesley M., and Richard C. Levin. 1989. "Empirical Studies of Innovation and Market Structure," in *Handbook of Industrial Organization, Vol. 2*. Richard Schmalensee and Robert D. Willig, eds. Amsterdam: North-Holland, 1059–1107.

Cragg, Michael I., and I. Alexander Dyck. 2003. "Privatization and Management Incentives: Evidence from the United Kingdom." *Journal of Law, Economics, and Organization*, 19(1): 176–217.

Crandall, Robert, and J. Gregory Sidak. 2002. "Is Structural Separation of Incumbent Local Exchange Carriers Necessary for Competition?" *Yale Journal on Regulation*, 19(2): 335–411.

Crémer, Jacques, and Richard P. McLean. 1985. "Optimal Selling Strategies under Uncertainty for a Discriminating Monopolist When Demands Are Interdependent." *Econometrica*, 53(2): 345–61.

Crew, Michael A., and Paul R. Kleindorfer. 2002. "Regulatory Economics: Twenty Years of Progress?" *Journal of Regulatory Economics*, 21(1): 5–22.

Crew, Michael A., Paul R. Kleindorfer, and John Sumpter. 2005. "Bringing Competition to Telecommunications by Divesting the RBOCs," in *Obtaining the Best from Regulation and Competition*. Michael A. Crew and Menachem Spiegel, eds. Norwell, Mass.: Kluwer Academic Publishers, 21–40.

Curien, Nicolas, Bruno Jullien, and Patrick Rey. 1998. "Pricing Regulation under Bypass Competition." *RAND Journal of Economics*, 29(2): 259–79.

Dana, James D., Jr., and Kathryn E. Spier. 1994. "Designing a Private Industry: Government Auctions with Endogenous Market Structure." *Journal of Public Economics*, 53(1): 127–47.

DeGraba, Patrick. 2004. "Reconciling the Off-Net Cost Pricing Principle with Efficient Network Utilization." *Information Economics and Policy*, 16(3): 475–94.

Demsetz, Harold. 1968. "Why Regulate Utilities?" *Journal of Law and Economics*, 11(1): 55–65.

Demski, Joel S., and David E. M. Sappington. 1984. "Optimal Incentive Contracts with Multiple Agents." *Journal of Economic Theory*, 33(1): 152–71.

Demski, Joel S., David E. M. Sappington, and Pablo T. Spiller. 1987. "Managing Supplier Switching."

RAND Journal of Economics, 18(1): 77–97.

Dewatripont, Mathias, Ian Jewitt, and Jean Tirole. 1999. "The Economics of Career Concerns, Part II: Application to Missions and Accountability of Government Agencies." *Review of Economic Studies*, 66(1): 199–217.

Dewenter, Kathryn L., and Paul H. Malatesta. 2001. "State-Owned and Privately Owned Firms: An Empirical Analysis of Profitability, Leverage, and Labor Intensity." *American Economic Review*, 91(1): 320–34.

Dewey, James F. 2000. "More Is Less? Regulation in a Rent Seeking World." *Journal of Regulatory Economics*, 18(2): 95–112.

Djankov, Simeon, and Peter Murrell. 2002. "Enterprise Restructuring in Transition: A Quantitative Survey." *Journal of Economic Literature*, 40(3): 739–92.

D'Souza, Juliet, and William L. Megginson. 1999. "The Financial and Operating Performance of Privatized Firms during the 1990s." *Journal of Finance*, 54(4): 1397–1438.

Economides, Nicholas. 1998. "The Incentive for Non-price Discrimination by an Input Monopolist." *International Journal of Industrial Organization*, 16(3): 271–84.

Edwards, Geoff, and Leonard Waverman. 2006. "The Effects of Public Ownership and Regulatory Independence on Regulatory Outcomes: A Study of Interconnect Rates in EU Telecommunications." *Journal of Regulatory Economics*, 29(1): 23–68.

Farrell, Joseph, and Paul Klemperer. Forthcoming. "Coordination and Lock-In: Competition with Switching Costs and Network Effects," in *Handbook of Industrial Organization, Vol. III*. Mark Armstrong and Robert Porter, eds. Amsterdam: North-Holland.

Farsi, Mehdi, Massimo Filippini, and William Greene. 2005. "Efficiency Measurement in Network Industries: Application to the Swiss Railway Companies." *Journal of Regulatory Economics*, 28(1): 69–90.

Federal Communications Commission. 2003. *Eighth Annual Report on the State of Competition in the Commercial Mobile Radio Services Industry*. Washington, D.C.: Federal Communications Commission.

Finsinger, Jorg, and Ingo Vogelsang. 1981. "Alternative Institutional Frameworks for Price Incentive Mechanism." *Kyklos*, 34(3): 388–404.

Finsinger, Jorg, and Ingo Vogelsang. 1982. "Performance Indices for Public Enterprises," in *Public Enterprise in Less Developed Countries*. Leroy Jones, ed. Cambridge: Cambridge University Press, 281–96.

Frydman, Roman, Cheryl Gray, Marek Hessel, and Andrzej Rapaczynski. 1999. "When Does Privatization Work? The Impact of Private Ownership on Corporate Performance in the Transition Economies." *Quarterly Journal of Economics*, 114(4): 1153–91.

Galal, Ahmed. 1996. Chile: Regulatory Specificity, Credbility of Commitment, and Distributional Demands, in *Regulations, Institutions, and Commitment: Comparative Studies of Telecommunications*. Brian Levy and Pablo T. Spiller, eds. Political Economy of Institutions and Decisions series. Cambridge; New York and Melbourne: Cambridge University Press, 121–44.

Gans, Joshua S., and Stephen P. King. 2004. "Access Holidays and the Timing of Infrastructure Investment." *Economic Record*, 80(248): 89–100.

Gerardin, Damien, and J. Gregory Sidak. 2005. "European and American Approaches to Antitrust Remedies and the Institutional Design of Regulation in Telecommunications," in *Handbook of Telecommunications Economics, Vol. II: Technology Evolution and the Internet*. Sumit K. Majumdar, Ingo Vogelsang, and Martin E. Cave, eds. Amsterdam: North-Holland, 517–53.

Glazer, Amihai, and Henry McMillan. 1992. "Pricing by the Firm under Regulatory Threat." *Quarterly Journal of Economics*, 107(3): 1089–99.

Guasch, J. Luis. 2004. *Granting and Renegotiating Infrastructure Concessions: Doing It Right*. Washington, D.C.: World Bank.

Guthrie, Graeme. Forthcoming. "Regulating Infrastructure: The Impact on Risk and Investment." *Journal of Economic Literature*.

Hausman, Jerry A. 1997. "Valuing the Effect of Regulation on New Services in Telecommunications." *Brookings Papers on Economic Activity*, 1–38.

Hausman, Jerry A., and J. Gregory Sidak. 1999. "A Consumer-Welfare Approach to the Mandatory Unbundling of Telecommunications Networks." *Yale Law Journal*, 109(3): 417–505.

Hausman, Jerry A., and J. Gregory Sidak. 2005. "Telecommunications Regulation: Current Approaches with the End in Sight." Mimeo. Massachusetts Institute of Technology.

Henisz, Witold J. 2000. "The Institutional Environment for Economic Growth." *Economics and Politics*, 12(1): 1–31.

Hillman, Jordan Jay, and Ronald R. Braeutigam. 1989. *Price Level Regulation for Diversified Public Utilities: An Assessment*. Norwell, Mass.: Kluwer Academic Publishers.

Hinton, Paul J., J. Douglas Zona, Richard Schmalensee, and William Taylor. 1998. "An Analysis of the Welfare Effects of Long-Distance Market Entry by an Integrated Access and Long-Distance Provider." *Journal of Regulatory Economics*, 13(2): 183–96.

Holmstrom, Bengt. 1999. "Managerial Incentive Problems: A Dynamic Perspective." *Review of Economic Studies*, 66(1): 169–82.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, and Organization*, 7(1): 24–52.

Jackson, Charles, Tracey Kelly, and Jeffrey Rohlfs. 1991. "Estimate of the Loss to the United States Caused by the FCC's Delay in Licensing Cellular Communications." Washington, D.C.: National Economic Research Associates.

Jamison, Mark A., Lynne Holt, and Sanford V. Berg. 2005. "Measuring and Mitigating Regulatory Risk in Private Infrastructure Investment." *Electricity Journal*, 18(6): 36–45.

Joskow, Paul L. 2005. "Incentive Regulation in Theory and Practice: Electricity Distribution and Transmission Networks." Mimeo. Massachusetts Institute of Technology.
Joskow, Paul L., and Edward Kahn. 2002. "A Quantitative Analysis of Pricing Behavior in California's Wholesale Electricity Market during Summer 2000." *Energy Journal*, 23(4): 1–35.
Joskow, Paul L., and Richard Schmalensee. 1986. "Incentive Regulation for Electric Utilities." *Yale Journal on Regulation*, 4(1): 1–49.
Kessides, Ioannis N. 2004. *Reforming Infrastructure: Privatization, Regulation, and Competition*. Washington, D.C.: World Bank.
Kornai, János, Eric Maskin, and Gerard Roland. 2003. "Understanding the Soft Budget Constraint." *Journal of Economic Literature*, 41(4): 1095–1136.
Krishna, Vijay. 2002. *Auction Theory*. San Diego; London and Sydney: Elsevier Science, Academic Press.
Laffont, Jean-Jacques. 1994. "The New Economics of Regulation Ten Years After." *Econometrica*, 62(3): 507–37.
Laffont, Jean-Jacques. 2005. *Regulation and Development*. Cambridge: Cambridge University Press.
Laffont, Jean-Jacques, and David Martimort. 1999. "Separation of Regulators against Collusive Behavior." *RAND Journal of Economics*, 30(2): 232–62.
Laffont, Jean-Jacques, Patrick Rey, and Jean Tirole. 1998. "Network Competition: Overview and Nondiscriminatory Pricing." *RAND Journal of Economics*, 29(1): 1–37.
Laffont, Jean-Jacques, and Jean Tirole. 1987. "Auctioning Incentive Contracts." *Journal of Political Economy*, 95(5): 921–37.
Laffont, Jean-Jacques, and Jean Tirole. 1988. "Repeated Auctions of Incentive Contracts, Investment, and Bidding Parity with an Application to Takeovers." *RAND Journal of Economics*, 19(4): 516–37.
Laffont, Jean-Jacques, and Jean Tirole. 1990a. "Optimal Bypass and Cream Skimming." *American Economic Review*, 80(5): 1042–61.
Laffont, Jean-Jacques, and Jean Tirole. 1990b. "The Politics of Government Decision Making: Regulatory Institutions." *Journal of Law, Economics, and Organization*, 6(1): 1–31.
Laffont, Jean-Jacques, and Jean Tirole. 1993. *A Theory of Incentives in Procurement and Regulation*. Cambridge: MIT Press.
Laffont, Jean-Jacques, and Jean Tirole. 1994. "Access Pricing and Competition." *European Economic Review*, 38(9): 1673–1710.
Laffont, Jean-Jacques, and Jean Tirole. 1996. "Creating Competition through Interconnection: Theory and Practice." *Journal of Regulatory Economics*, 10(3): 227–56.
Laffont, Jean-Jacques, and Jean Tirole. 2000. *Competition in Telecommunications*. Cambridge: MIT Press.
La Porta, Rafael, and Florencio López-de-Silanes. 1999. "The Benefits of Privatization: Evidence from Mexico." *Quarterly Journal of Economics*, 114(4): 1193–1242.
Lee, Sang Hyup, and Jonathan H. Hamilton. 1999. "Using Market Structure to Regulate a Vertically Integrated Monopolist." *Journal of Regulatory Economics*, 15(3): 223–48.
Levine, Paul, Jon Stern, and Francesc Trillas. 2005. "Utility Price Regulation and Time Inconsistency: Comparisons with Monetary Policy." *Oxford Economic Papers*, 57(3): 447–78.
Levy, Brian, and Pablo T. Spiller. 1994. "The Institutional Foundations of Regulatory Commitment: A Comparative Analysis of Telecommunications Regulation." *Journal of Law, Economics, and Organization*, 10(2): 201–46.
Levy, Brian, and Pablo T. Spiller, eds. 1996. *Regulations, Institutions, and Commitment: Comparative Studies of Telecommunications*. Cambridge: Cambridge University Press.
Lewis, Tracy R., and David E. M. Sappington. 2000. "Motivating Wealth-Constrained Actors." *American Economic Review*, 90(4): 944–60.
Loeb, Martin, and Wesley A. Magat. 1979. "A Decentralized Method for Utility Regulation." *Journal of Law and Economics*, 22(2): 399–404.
Loury, Glenn C. 1979. "Market Structure and Innovation." *Quarterly Journal of Economics*, 93(3): 395–410.
Mandy, David M. 2000. "Killing the Goose That May Have Laid the Golden Egg: Only the Data Know Whether Sabotage Pays." *Journal of Regulatory Economics*, 17(2): 157–72.
Mandy, David M., and David E. M. Sappington. Forthcoming. "Incentives for Sabotage in Vertically Related Industries." *Journal of Regulatory Economics.*
Manelli, Alejandro M., and Daniel R. Vincent. 1995. "Optimal Procurement Mechanisms." *Econometrica*, 63(3): 591–620.
Mankiw, N. Gregory, and Michael D. Whinston. 1986. "Free Entry and Social Inefficiency." *RAND Journal of Economics*, 17(1): 48–58.
Mansell, Robert, and Jeffrey Church. 1995. *Traditional and Incentive Regulation: Applications to Natural Gas Pipelines in Canada.* Calgary: The Van Horne Institute.
Martimort, David. 1999a. "The Life Cycle of Regulatory Agencies: Dynamic Capture and Transaction Costs." *Review of Economic Studies*, 66(4): 929–47.
Martimort, David. 1999b. "Renegotiation Design with Multiple Regulators." *Journal of Economic Theory*, 88(2): 261–93.
Maskin, Eric, and Jean Tirole. 2004. "The Politician and the Judge: Accountability in Government." *American Economic Review*, 94(4): 1034–54.
McAfee, R. Preston, and John McMillan. 1987. "Competition for Agency Contracts." *RAND Journal of Economics*, 18(2): 296–307.
Megginson, William L., Robert C. Nash, and Matthias van Randenborgh. 1994. "The Financial and Operating Performance of Newly Privatized Firms: An International Empirical Analysis." *Journal of Finance*, 49(2): 403–52.

Megginson, William L., and Jeffry M. Netter. 2001. "From State to Market: A Survey of Empirical Studies on Privatization." *Journal of Economic Literature*, 39(2): 321–89.

Mizutani, Fumitoshi, and Shuji Uranishi. 2003. "The Post Office vs. Parcel Delivery Companies: Competition Effects on Costs and Productivity." *Journal of Regulatory Economics*, 23(3): 299–319.

Myerson, Roger B. 1979. "Incentive Compatibility and the Bargaining Problem." *Econometrica*, 47(1): 61–73.

Newbery, David M. 1999. *Privatization, Restructuring, and Regulation of Network Utilities*. Cambridge: MIT Press.

Noll, Roger G. 2000. "Telecommunications Reform in Developing Countries," in *Economic Policy Reform: The Second Stage*. Anne O. Krueger, ed. Chicago and London: University of Chicago Press, 183–242.

Otsuka, Yasuji. 1997. "A Welfare Analysis of Local Franchise and Other Types of Regulation: Evidence from the Cable TV Industry." *Journal of Regulatory Economics*, 11(2): 157–80.

Paredes, Ricardo D. 2005. "Lessons from the Deregulation Transition in Chile's Local Telephony Market." *Telecommunications Policy*, 29(5–6): 333–50.

Parker, Philip M., and Lars-Hendrik Röller. 1997. "Collusive Conduct in Duopolies: Multimarket Contact and Cross-Ownership in the Mobile Telephone Industry." *RAND Journal of Economics*, 28(2): 304–22.

Perotti, Enrico C. 1995. "Credible Privatization." *American Economic Review*, 85(4): 847–59.

Perotti, Enrico C., and Serhat E. Guney. 1993. "The Structure of Privatization Plans." *Financial Management*, 22(1): 84–98.

Posner, Richard A. 1974. "Theories of Economic Regulation." *Bell Journal of Economics*, 5(2): 335–58.

Prager, Robin A. 1989. "Franchise Bidding for Natural Monopoly: The Case of Cable Television in Massachusetts." *Journal of Regulatory Economics*, 1(2): 115–31.

Reiffen, David. 1998. "A Regulated Firm's Incentive to Discriminate: A Reevaluation and Extension of Weisman's Result." *Journal of Regulatory Economics*, 14(1): 79–86.

Reiffen, David, Laurence Schumann, and Michael R. Ward. 2000. "Discriminatory Dealing with Downstream Competitors: Evidence from the Cellular Industry." *Journal of Industrial Economics*, 48(3): 253–86.

Rey, Patrick and Jean Tirole. Forthcoming. "A Primer on Foreclosure," in *Handbook of Industrial Organization, Vol. III*. Mark Armstrong and Robert Porter, eds. Amsterdam: North-Holland.

Riordan, Michael H. 1996. "Contracting with Qualified Suppliers." *International Economic Review*, 37(1): 115–28.

Riordan, Michael H., and David E. M. Sappington. 1987. "Awarding Monopoly Franchises." *American Economic Review*, 77(3): 375–87.

Riordan, Michael H., and David E. M. Sappington. 1989. "Second Sourcing." *RAND Journal of Economics*, 20(1): 41–58.

Rob, Rafael. 1986. "The Design of Procurement Contracts." *American Economic Review*, 76(3): 378–89.

Rochet, Jean-Charles, and Lars Stole. 2003. "The Economics of Multidimensional Screening," in *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress, Vol. 1.* Mathias Dewatripont, Lars Hansen, and Stephen Turnovsky, eds. Cambridge: Cambridge University Press, 150–97.

Rosston, Gregory, and Roger G. Noll. 2002. "The Economics of the Supreme Court's Decision on Forward Looking Costs." *Review of Network Economics*, 1(2): 81–89.

Salant, David J. 1995. "Behind the Revolving Door: A New View of Public Utility Regulation." *RAND Journal of Economics*, 26(3): 362–77.

Salant, David J., and Glenn A. Woroch. 1992. "Trigger Price Regulation." *RAND Journal of Economics*, 23(1): 29–51.

Sappington, David E.M. 1994. "Designing Incentive Regulation." *Review of Industrial Organization*, 9(3): 245–72.

Sappington, David E.M. 2002. "Price Regulation and Incentives," in *Handbook of Telecommunications Economics, Vol. 1.* Martin Cave, Sumit Majumdar, and Ingo Vogelsang, eds. Amsterdam: North-Holland, 225–93.

Sappington, David E. M. 2005a. "Regulating Service Quality: A Survey." *Journal of Regulatory Economics*, 27(2): 123–54.

Sappington, David E. M. 2005b. "On the Irrelevance of Input Prices for Make-or-Buy Decisions." *American Economic Review*, 95(5): 1631–38.

Sappington, David E. M., and David S. Sibley. 1988. "Regulating without Cost Information: The Incremental Surplus Subsidy Scheme." *International Economic Review*, 29(2): 297–306.

Sappington, David E. M., and J. Gregory Sidak. 2003. "Incentives for Anticompetitive Behavior by Public Enterprises." *Review of Industrial Organization*, 22(3): 183–206.

Sappington, David E. M., and Joseph E. Stiglitz. 1987. "Privatization, Information and Incentives." *Journal of Policy Analysis and Management*, 6(4): 567–82.

Sappington, David E. M., and Dennis L. Weisman. 1996. *Designing Incentive Regulation for the Telecommunications Industry*. Cambridge: MIT Press.

Schmalensee, Richard. 1989. "Good Regulatory Regimes." *RAND Journal of Economics*, 20(3): 417–36.

Schumpeter, Joseph. 1950. *Capitalism, Socialism, and Democracy*. New York: Harper and Brothers.

Sharkey, William W. 1979. "A Decentralized Method for Utility Regulation: A Comment." *Journal of Law and Economics*, 22(2): 405–07.

Sharkey, William W. 2004. "Pricing Interconnection of Telecommunications Networks." Federal Communications Commission manuscript.

Shleifer, Andrei. 1985. "A Theory of Yardstick

Competition." *RAND Journal of Economics*, 16(3): 319–27.
Sibley, David S., and Dennis L. Weisman. 1998. "Raising Rivals' Costs: The Entry of an Upstream Monopolist into Downstream Markets." *Information Economics and Policy*, 10(4): 451–70.
Sidak, J. Gregory. 2003. "Remedies and the Institutional Design of Regulation in Network Industries." *Michigan State DCL Law Review*, 3: 741–56.
Sobel, Joel. 1999. "A Reexamination of Yardstick Competition." *Journal of Economics and Management Strategy*, 8(1): 33–60.
Spence, A. Michael. 1975. "Monopoly, Quality, and Regulation." *Bell Journal of Economics*, 6(2): 417–29.
Spiller, Pablo T., and Carlo G. Cardilli. 1997. "The Frontier of Telecommunications Deregulation: Small Countries Leading the Pack." *Journal of Economic Perspectives*, 11(4): 127–38.
Spiller, Pablo T., and Ingo Vogelsang. 1997. "The Institutional Foundations of Regulatory Commitment in the UK: The Case of Telecommunications." *Journal of Institutional and Theoretical Economics*, 153(4): 607–29.
Stern, Jon. 2000. "Electricity and Telecommunications Regulatory Institutions in Small and Developing Countries." *Utilities Policy*, 9(3): 131–57.
Stigler, George J. 1971. "The Economic Theory of Regulation." *Bell Journal of Economics*, 2(1): 3–21.
Varian, Hal R. 1980. "A Model of Sales." *American Economic Review*, 70(4): 651–59.
Vickers, John. 1995. "Competition and Regulation in Vertically Related Markets." *Review of Economic Studies*, 62(1): 1–17.
Vogelsang, Ingo. 2002. "Incentive Regulation and Competition in Public Utility Markets: A 20-Year Perspective." *Journal of Regulatory Economics*, 22(1): 5–27.
Vogelsang, Ingo. 2003. "Price Regulation of Access to Telecommunications Networks." *Journal of Economic Literature*, 41(3): 830–62.
Vogelsang, Ingo. 2005. "Electricity Transmission Pricing and Performance-Based Regulation." Mimeo. Boston University.
Wallsten, Scott J. 2001. "An Econometric Analysis of Telecom Competition, Privatization, and Regulation in Africa and Latin America." *Journal of Industrial Economics*, 49(1): 1–19.
Waterson, Michael. 2003. "The Role of Consumers in Competition and Competition Policy." *International Journal of Industrial Organization*, 21(2): 129–50.
Weisman, Dennis L. 1993. "Superior Regulatory Regimes in Theory and Practice." *Journal of Regulatory Economics*, 5(4): 355–66.
Weisman, Dennis L. 1995. "Regulation and the Vertically Integrated Firm: The Case of RBOC Entry into InterLATA Long Distance." *Journal of Regulatory Economics*, 8(3): 249–66.
Williamson, Oliver. 1975. *Markets and Hierarchies: Analysis and Antitrust Implications.* New York: Free Press.
Williamson, Oliver. 1976. "Franchise Bidding for Natural Monopolies—In General and With Respect to CATV." *Bell Journal of Economics*, 7(1): 73–104.
Willig, Robert. 1979. "The Theory of Network Access Pricing," in *Issues in Public Utility Regulation.* Harry Trebing, ed. East Lansing: Michigan State University Press, 109–52.
Wolinsky, Asher. 1997. "Regulation of Duopoly: Managed Competition Versus Regulated Monopolies." *Journal of Economics and Management Strategy*, 6(4): 821–47.
Wood, Lisa V., and David E. M. Sappington. 2004. "On the Design of Performance Measurement Plans in the Telecommunications Industry." *Telecommunications Policy*, 28(11): 801–20.
Woroch, Glenn A. 2000. "Competition's Effect on Investment in Digital Infrastructure." University of California, Berkeley Working Paper.
Woroch, Glenn A. 2002. "Local Network Competition," in *Handbook of Telecommunications Economics, Vol. 1: Structure, Regulation and Competition*. Martin E. Cave, Sumit K. Majumdar and Ingo Vogelsang, eds. Amsterdam: North-Holland, 641–716.
Zimmerman, Paul R. 2003. "Regional Bell Operating Company Entry into Long-Distance and Non-price Discrimination against Rival Interexchange Carriers: Empirical Evidence from Panel Data." *Applied Stochastic Models in Business and Industry*, 19(4): 269–90.
Zupan, Mark A. 1989a. "Cable Franchise Renewals: Do Incumbent Firms Behave Opportunistically?" *RAND Journal of Economics*, 20(4): 473–82.
Zupan, Mark A. 1989b. "The Efficacy of Franchise Bidding Schemes in the Case of Cable Television: Some Systematic Evidence." *Journal of Law and Economics*, 32(2): 401–56.

Army Corps of Engineers Blows Mississippi River Levee

Written by Dave Bohon

Wednesday, 04 May 2011 14:45

The Army Corps of Engineers blew open two miles of levee along the Mississippi River in Missouri May 2, an effort that may have saved the community of Cairo, Illinois, but which flooded 200 square miles of fertile farmland, which Missouri Governor Jay Nixon called “literally the most productive part of our continent.”
The *New York Times* reported May 3 that the massive flood of river released by the Army Corps of Engineers into the 130,000 acre floodway “will soon re-enter the Mississippi near New Madrid [Missouri], through two 5,500-foot stretches blasted out over two days at the lower end of the basin, and the crest will continue to roll on, with the river expected to match or beat its previous record heights at many points along the way.”

Major General Michael Walsh of the Army Corps of Engineers predicted that his team was just at the beginning of its effort to control the river, which had reached record levels over the past few days. The *Times* reported that the Army Corps’ “decision to inundate the 130,000 acres within the spillway’s basin almost certainly saved the town of Cairo, Illinois. The river had reached a record 61.7 feet at Cairo before the explosion and was predicted to rise more than a foot further.” The *Times* quoted Colonel Vernie Reichling Jr. of the Army Corps’ Memphis District as saying that they would most likely have to “fight this river all the way down to the Gulf of Mexico. I don’t see this letting up.”

According to CNN, the first blast on May 2 at the levee near Birds Point, Missouri was followed up by a second the next day at New Madrid, Missouri, with a third planned for May 4 near Hickman, Kentucky. “The second and third blasts, downstream of Birds Point, will allow floodwater to return to the Mississippi River,” CNN reported.

While the main defenses against the Mississippi River were expected to hold, additional flooding was “likely as water backs up into the rivers and tributaries that feed into the Mississippi, and tests ‘non-federal’ levees that line those waterways,” reported the *Times*. Jeff Graschel of the Lower Mississippi River Forecast Center, a subsidiary of the National Oceanic and Atmospheric Administration, told the paper that while federal levees are strong enough to handle what the river will bring, “anything that’s non-federal is a different story.”

Although residents of Cairo, Illinois, and other towns saved by the Army’s actions were grateful, those in the flooded areas, especially those who farm the hundreds of thousands of acres of fertile land, had a different take. Farmer Bryan Feezor told CNN that the sight of his submerged fields made him sick. “Farming is all I ever have done ... and it’s under water, he said. “I really don’t know [what I’m going to do].”

“It’s a life-changing event,” one local farmer, Travis Williams, told CBN News. “My heart goes out to all the farmers who lost their land and homes.”

Even with the Army Corps’ actions, a plan Walsh said had been designed decades ago for such a contingency, and which had “operated as designed," communities and residents along the Mississippi remained in danger. The residents of Cairo remained under a mandatory evacuation, and another six other communities had been issued voluntary evacuation notices, Patti Thompson, a spokeswoman for the Illinois Emergency Management Agency, told CNN. “We’re definitely not out of the woods yet,” she said. “The levees are all very saturated right now and they’re going to continue to have a lot of pressure on them for several days.”

Even with the series of breaches, the National Weather Service continued to predict near-record flooding in southern Illinois, southwest Indiana, western Kentucky and Tennessee, southeastern Missouri, northeastern Arkansas, and even into Mississippi and Louisiana.

Execution of the plan did not come without major conflict, with the state of Missouri taking the Army Corps of Engineers to federal court over the plan. “The state argued the floodwater would deposit silt on the estimated 130,000 acres, and years, along with millions of dollars, would be required to fix the damage,” noted the CNN report. By Sunday, May 1, the case had made its way to the U.S. Supreme Court, which declined to intervene, clearing the way for Walsh to initiate what he called a “heart-wrenching” decision. “I’ve been involved with flooding for 10 years and it takes a long time to recover from something like this,” he said.

Reflecting afterward on the decision to blow the levees, and on the prospect that more would have to be destroyed before it was over, Walsh said that choosing between saving Cairo and other communities or saving the farms in the flood plain was not “easy or hard. It’s simply grave, because the decision leads to loss of property and livelihood,

either in a floodway or in an area that was not designed to flood."

It seemed clear that Walsh sympathized with those who would bear the brunt of his decision, telling one group of Missouri property owners, "I recognize all of your lives will be impacted, but these levees have never been under this pressure before."

According to the Associated Press, even those opposed to the Army Corps' actions "appreciate how Walsh — who is responsible for managing the entire length of the Mississippi River valley, from Canada to the Gulf of Mexico—has handled the situation." Even Governor Nixon, whose administration took the lead in trying to stop the levees from being destroyed, conceded that Walsh "has a very difficult decision to make relatively quickly. He understands the magnitude of the decision on his plate."

Nixon said his focus would now turn to having the Missouri levees rebuilt and to restoring the lost farmland to productivity. But officials warned that it could be early fall before the water drains from the land, and that sediment left by the river could do lasting damage to the once fertile acreage.

Dow Jones Reprints: This copy is for your personal, non-commercial use only. To order presentation-ready copies for distribution to your colleagues, clients or customers, use the Order Reprints tool at the bottom of any article or visit www.djreprints.com

• See a sample reprint in PDF format. • Order a reprint of this article now

THE WALL STREET JOURNAL.
WSJ.com

U.S. NEWS | APRIL 18, 2011

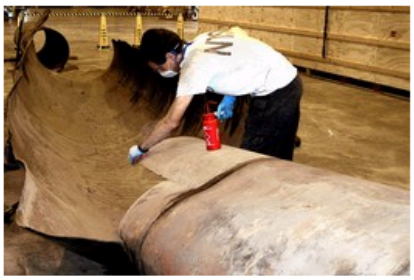
More Utilities Can't Prove Pipe Tests

By DANIEL GILBERT And REBECCA SMITH

Two more big California utilities told state regulators they couldn't prove they have done key safety tests on hundreds of miles of pipelines carrying explosive natural gas.

The disclosure late Friday by gas distribution utilities owned by San Diego-based Sempra Energy came on the heels of a similar admission last month from San Francisco-based PG&E Corp. A PG&E pipeline exploded in September, incinerating a neighborhood in San Bruno, Calif., and killing eight people.

Investigators suspect weak welds in the pipe may have played a role in the explosion.



Associated Press

A technician cleans a ruptured pipeline at a NTSB training center.

Following the San Bruno accident and urgent recommendations by the National Transportation Safety Board, California regulators in January ordered the state's gas utilities to review their safety records and report back on any gaps. The required records included pressure-test data on lines that run through populous areas; these tests are designed to flush out any leaks or flaws in a pipeline by running water or an inert gas through the line at high pressure.

In its joint report, Sempra's Southern California Gas Co. and San Diego Gas & Electric Co. said they were continuing to look, but haven't found complete pressure test records for about 450 miles—or 28%—of their pipelines that run through neighborhoods, or near parks and offices.

EXPERIENCE WSJ PROFESSIONAL

Editors' Deep Dive: Pipeline Safety Attracts Regulatory Heat

NATURAL GAS INTELLIGENCE
FREE California Gas Pipelines, Storage Get More Scrutiny

NATURAL GAS INTELLIGENCE
Texas Toughens Distribution Pipeline Monitoring

THE OIL DAILY
U.S. Transportation Secretary Seeks Tougher Pipeline Safety Measures

Access thousands of business sources not available on the free web. Learn More

"We have not yet located records sufficient to document that the pipelines have been strength-tested per the NTSB recommendations," Sempra said in its filing to the California Public Utilities Commission.

Rick Morrow, the company's vice president of gas engineering, said in an interview on Sunday there were other ways to establish safe operations. He said he was confident Sempra's system was safe because 71% of gas transmission lines in populated areas, including some segments for which the company didn't have pressure-test reports, have been inspected with robotic devices called smart pigs that travel the interior of the pipes.

"I don't want people to conclude we aren't operating our lines safely because our operating history tells a lot about their safety," Mr. Morrow said.

Sempra said it would continue to search for records and in the meantime would lower the pressure at which it moved gas in some lines and increase the frequency of inspections and leak surveys.

Richard Kuprewicz, a pipeline safety consultant based in Redmond, Wash., said he was surprised that the utilities have such gaps in their records, which he said exposed a deficiency in oversight.

"The Public Utilities Commission is not regulating gas transmission pipeline safety in the state of California," Mr. Kuprewicz said in a telephone interview. "The public has a right to be concerned."

A month ago, PG&E said it had been able to find complete records for only two-thirds of its high-pressure lines that run through populated areas.

After the San Bruno explosion, federal accident investigators learned that PG&E's records wrongly identified the pipe that ruptured as seamless, an error that meant the utility never conducted the sort of inspections that might have detected a problem with seam welds in that pipe, which was more than 50 years old. State regulators want to know where similar pipe may have been used but are hitting a wall due to gaps in records.

In a potential settlement with regulatory staff, PG&E agreed to pay a $6 million fine, with $3 million suspended if it met future targets in its continuing records search. The state commission hasn't yet approved the fine and some public officials, including the state attorney general, said it was too small.

Sempra says it is in compliance with all state rules and faces no penalties. A spokeswoman for the California Public Utilities Commission said Sunday the company's filing was "under review."

Copyright 2011 Dow Jones & Company, Inc. All Rights Reserved
This copy is for your personal, non-commercial use only. Distribution and use of this material are governed by our Subscriber Agreement and by copyright law. For non-personal use or to order multiple copies, please contact Dow Jones Reprints at 1-800-843-0008 or visit
www.djreprints.com

THE WALL STREET JOURNAL.

Apple's Mobile Rules To Get FTC Scrutiny

By Thomas Catan

305 words
12 June 2010
The Wall Street Journal
J
B1
English
(Copyright (c) 2010, Dow Jones & Company, Inc.)

WASHINGTON -- The U.S. Federal Trade Commission will investigate whether Apple Inc.'s business practices harm competition in the market for software used on mobile devices, people familiar with the situation said.

For weeks, the FTC has been engaged in negotiations with the Department of Justice over which agency would review allegations by companies that say they're being shut-out of one of the most important emerging computing platforms.

Adobe Systems Inc. has been engaged in a public feud with Apple over its decision to ban Adobe's Flash video technology from Apple devices. This week, Google Inc. complained Apple's new rules on developers could bar Google and other rivals from selling ads inside iPhone and iPad applications, such as games.

Apple has also banned software developers from using other companies' tools to develop software for its devices.

Both Apple and the FTC declined to comment. The decision was reported earlier by Bloomberg News.

This may not be the only antitrust investigation Apple faces. Justice Department lawyers recently contacted companies about Apple's practices in the music business. It isn't clear if it would continue that inquiry.

The Justice Department is already investigating whether Apple and a range of other tech companies improperly agreed not to poach each other's employees.

The FTC will have a wealth of information to mine for its probe. It recently completed a six-month investigation of Google's $750 million acquisition of AdMob Inc., giving its lawyers knowledge of the mobile-ad market that Apple has also entered.

"The Commission has reason to believe that Apple quickly will become a strong mobile advertising network competitor," the FTC said last month.

Telecommunications & Electronic Media

The Opportunity and Peril of Smart Grid

*By Ray Gifford & Eric Gunning**

Last year witnessed "the smart grid" going from an arcane concept about updating the electric distribution system to a full-fledged public policy priority. The Obama Administration views the deployment of smart grid technologies as a linchpin to achieving its new energy economy objectives. These goals include: lowering energy costs for consumers, achieving energy independence, promoting clean and green technologies, and reducing greenhouse gas emissions. To that end, Congress allocated approximately $4 billion in stimulus grants and loans to smart grid deployments and tasked the National Institute for Standards and Technology ("NIST") to develop smart grid interoperability standards.

While smart grid deserves some of the hype that it has received, the reality of smart grid deployment is much more fraught. For smart grid deployment combines familiar challenges of information technology and communications network deployment—standards and interoperability problems, intellectual property issues, and privacy and security—with regulatory issues—prudent investment, rate design, competitive, and monopoly aspects of the service. In turn, this not quite harmonious stew of issues interrelates to present challenges to utilities, information technology vendors and regulators. How these challenges are answered will decide whether smart grid delivers the energy-saving, market-enabling efficiencies that proponents' promise, or if it becomes an industrial policy boondoggle where costs exceed the benefits.

The remainder of this article discusses the major policy issues facing smart grid deployment.

I. What is the Smart Grid?

'Smart Grid' is a general description encompassing distinct components of the electricity distribution network. Analytically, the electric system can be divided into three components: generation, transmission, and distribution. Generation involves just that, generation of electricity from central station power plants. These sources can be coal, nuclear, natural gas, or renewable resources. From there, the electricity is transported over the transmission system, which is an interconnected series of high voltage power lines.[1] In turn, the distribution system takes high voltage power into substations, steps down the voltage and delivers it to consumers.

This is a highly abstracted description of the electric grid system but will suffice for this discussion. The smart grid, for the most part, involves the final part of the grid—the distribution system. "Smart grid" involves creating robust communications paths between end-use consumers (industrial, commercial, and residential users) and upstream to the utilities, or other energy service providers. This communication path, in turn, can be used to manage and monitor voltage and the "cleanness" of power, communicate price to consumers, and customize energy management and usage by consumers.

Each of these smart grid capabilities promises potential benefits to consumers. Better monitoring and control of the distribution system increases reliability. Utilities can know and anticipate where outages occur or where transformers are stressed to the breaking point. In addition, "cleaner" power in the distribution system means less power needs to be generated in the first place.

The ability to communicate price in real-time (RTP) or time-of-use (TOU) promises increased efficiency and conservation. Currently, the vast majority of consumers pay for electricity on an averaged cost basis—an electron used at 4pm is priced the same as one consumed at 4am for the retail customer. Meanwhile, the underlying wholesale cost of generating that electron varies greatly depending on the time of day. During "peak demand," wholesale electricity prices can be five or ten times higher than "off peak" prices. Because there is no price incentive for consumers to conserve during peak demand periods[2] and utilities have a regulatory duty to meet demand, a substantial amount of investment must go to plan and build infrastructure to meet these peaks. Smart grid then, by communicating price to consumers, promises to "shave" energy demand during peak demand times. In turn, this should bring savings to consumers and the electric system by eliminating the need to build generation to meet demand peaks.

Smart grid then is very simple and very complex. It is simple because it involves creating communications paths on the distribution system where consumers have transparency into prices and generation mix; and where distribution operators have more information about demand and system status. It is complex because the architecture and regulatory treatment of this communications path presents many questions.

We turn to those now, beginning with standard setting:

II. NIST Interoperability Roadmap and Standards-Setting

On January 20, 2010, the National Institute of Standards and Technology (NIST) published its "*NIST Framework and Roadmap for Smart Grid Interoperability Standards, Release 1.0*." This report is the culmination of Phase I of NIST's three-phased approach for complying with the congressional mandate to coordinate development of interoperability standards for Smart Grid under the Energy Independence and Security Act of 2007 (EISA). From here, Phases II and III of the approach will be delegated to the newly-created Smart Grid Interoperability Panel (SGIP). This panel will be charged with refining proposed standards and developing new ones as the industry progresses.

** Ray Gifford and Eric Gunning are attorneys with Wilkinson Barker Knauer, LLP, and lead the firm's Smart Grid practice area. Gifford is former Chairman of the Colorado Public Utilities Commission.*

Briefly, the NIST Report:

- Identifies seventy-five existing interoperability standards that are applicable or likely to be applicable to Smart Grid;
- Specifies fifteen high-priority gaps and another fifty-five lower-priority gaps for which standards need to be adopted, receive industry consensus and NIST (or SGIP) approval;
- Establishes priority action plans to address and resolve the identified high-priority gaps;
- Introduces a Smart Grid conceptual model to create a common language for discussing the complex "system of systems"; and
- Passes the baton to the SGIP, a public-private partnership comprised of over 400 members and twenty-two stakeholder groups, to implement Phase II of the NIST interoperability plan which includes, providing a permanent process to "support the ongoing evolution of the Smart Grid Interoperability Framework, identify and address additional gaps, reflect changes in technology and requirements in the standards, and provide ongoing coordination of [standard setting organization] efforts to support timely availability of new or revised Smart Grid Standards." In addition, SGIP will also be responsible for Phase III: Smart Grid device conformity assessments and testing to ensure interoperability.

The issuance of the NIST report closes out on a frenetic year for the federal government in its efforts to spur development and deployment of Smart Grid technologies. These efforts included $3.4 billion in DOE grants awarded in October 2009 as part of the American Reinvestment and Recovery Act of 2009 (ARRA). Whether these efforts will actually catalyze real and substantial growth in Smart Grid remains an open question, and it may be many years before we can assess their success or failure. This determination with respect to NIST's efforts to "solve" interoperability will hinge on a number of practical, legal, and business issues, some of which are outlined below:

A. Federal-State Jurisdictional Issues.

The federal government has some laudable goals with respect to standard setting and spurring Smart Grid deployments. But its jurisdictional authority over the energy markets is more constrained. Through FERC, the federal government can only mandate standards for interstate transmission. It does not have authority over generation, middle-mile and last-mile distribution, or in-home energy management. Therefore, any federal standard setting efforts beyond interstate transmission must rely on condition-laden federal grants and loans (see the DOE grant program), fostering industry adoption (see the SPIG program) and pressuring/encouraging state regulatory commission to buy in. On the last point, there is no guarantee that state regulators will support the federal efforts. State regulators may even bristle on what are perceived to be federal intrusions into state regulatory turf. Standards adoption in large part will therefore depend on states' willingness to mandate NIST standards in state Smart Grid deployments.

Significantly, then, the federal push for standards rests on the hortatory ability of NIST, DOE, and FERC to convince the states and utilities that the federal standards should be followed, or in utility law parlance, would constitute "reasonable and prudent" investments.

B. Preventing Stranded Investments.

NIST's standard setting efforts must avoid, if possible, stranded investments in incompatible technologies. While the over 100 demonstration projects being deployed through the federal stimulus efforts all must incorporate NIST standards, there are hundreds more AMI (automated meter infrastructure) deployments that predate the Smart Grid push. It is inevitable that certain utilities will have deployed non-compliant technologies.

State regulatory commissions will have to face the unenviable task of deciding whether to: (a) allow ratepayers to assume the costs of the utilities' replacement of the stranded technology, or (b) allow the continued use of technologies that are not NIST-standard compliant and thus allow continued use of stranded technologies that will not be supported. Moreover, any utilities seeking cost recovery for AMI or other Smart Grid deployments that are NIST-standard compliant may face significant push-back from the state commissions.

C. What If NIST Gets It Wrong?

Mandating standards is always a tricky endeavor. If NIST focuses its efforts on bolstering a standard that turns out to be deficient or a subpar standard, the federal efforts may have the perverse effect of retarding innovation. It is not clear how nimble NIST and its Smart Grid Interoperability Panel will be in responding to these developments and being able to shift approved standards midstream.

D. Picking Winners and Losers.

Although NIST claims that its standard setting efforts are technology agnostic, there will invariably be winners and losers based on the standards NIST endorses. Again, these winners and losers may not be based entirely on merit but on business decisions regarding technology development, or on political economy factors having very little to do with the underlying technology.

For example, there has been a significant push by technology companies to ensure that Internet protocol (IP) standards are incorporated into the NIST-approved standards. Taking the cue, the NIST includes the IP suite as potential standards and defers to the Internet Engineering Task Force (IETF) to identify the suite of protocols applicable to Smart Grid. Adopting IP will benefit existing technology providers in the Internet space and allow quicker integration to other technologies and sectors. In addition, IP does provide a number of benefits; notably, it is a mature standard with widespread adoption, it enables bandwidth sharing and quality of service measures,

and it is easily scalable. NIST also strongly suggests the adoption of IPv6 (as opposed to IPv4) in developing Smart Grid standards. And yet, non-IP-based standards vendors and advocates raise security concerns about IP-based infrastructure because its very pervasiveness makes it more vulnerable.[3]

E. Privacy Issues.

The NIST report accurately noted that privacy, including the collection of energy consumption data from consumers and the ownership of such data are issues of significant concern to state regulators. As the state regulators strive to enact rules on these issues, it is possible that these decisions will negatively impact the standard setting efforts. This could occur if there is a disconnect between state privacy and ownership rules and the standards regime adopted at the federal level. Although the National Association of Regulatory Utility Commissioners (NARUC) adopted the "Resolution Urging the Adoption of General Privacy Principles for State Commission Use in Considering Privacy Implications of the Use of Utility Customer Information," it is unclear how state regulatory commissions will respond or, as noted by the NIST report, whether state regulatory commissions have statutory authority to make privacy rulings.

With respect to standards setting, the Smart Grid industry is still in its infancy, with a number of hurdles to overcome before it has a comprehensive set of standards that will enable interoperability, plug-and-play interoperability, and stability. The NIST report constitutes a huge initial step—it is hoped—in the right direction.

III. Federal and State Tensions and the Role of the State Governments

The federal smart grid push (through the NIST interoperability mandate) and the $3.4 billion smart grid stimulus package in American Recovery and Reinvestment Act of 2009 ("ARRA") has created tension between federal and state regulators as the majority of the regulatory, cost recovery and policy issues related to grid modernization will occur at the state level. While the Federal Energy Regulation Commission ("FERC") regulates the wholesale energy markets and interstate transmission, state regulatory commissions have jurisdiction over public utilities and intrastate energy distribution and generation. FERC's and NIST's federal authority with respect to standards therefore can only directly impact interstate transmission, not distribution, the last mile, or home area networks, all of which are within the purview of the states. Therefore, the success of smart grid depends in large part on the actions taken by state regulators with respect to, among other things, cost recovery of smart grid investments, market structure, competitive issues, and establishing dynamic retail energy pricing (a necessary component of demand response). Despite the significant role of state regulators, they have been noticeably underrepresented in many smart grid policy discussions and debates.

This is problematic because the states are going to have the primary role in approving smart grid investment.[4] Significantly, state regulators will evaluate:

– Cost recovery and prudent investment tests for utilities' deployment decisions of smart grid technologies;

– Revising the regulatory structure for utilities to more closely align utilities' incentives with smart grid benefits. This could entail decoupling the utilities' profits from the amount of kilowatt-hours (kWhs) consumed and shifting to another regime that compensates utilities for adopting energy efficiency measures, akin to price cap regulation used to regulate telecommunications;

– Dynamic pricing and retail rate structures that enable, among other things, variable retail pricing based on wholesale energy prices;

– Ownership and sharing of customer data generated as a result of the smart grid; and

– Competition issues such as interoperability, unbundling, market regulation and electric network neutrality. [5]

Although it is the Obama Administration's clear belief that grid modernization is a key component to many of the country's energy problems, state regulators will actually need to evaluate the value to energy consumers. Simply put, the benefits of smart grid must prove to be greater than the costs of building out this new communications path. This is public utility law 101—investment in the smart grid must show itself to be reasonable and prudent. And that cost/benefit equation may have different answers in different utility systems.

The federal promotion of smart grid creates tension with states, as it is the state commissions that will be asked to approve utility expenditures on projects partially funded by federal stimulus dollars. Moreover, state regulators may perceive the federal efforts as an attempt to usurp some of the state powers with respect to the prudence of grid investments, interoperability mandates, and grid management.

IV. Network Design and Market Structure

Despite stimulus funding and the rush to develop interoperability standards, significant questions remain regarding what is "smart grid" and what the grid of the future will look like. Debates are ongoing regarding whether the utility grid should have intelligence at the core, be based on an end-to-end architecture similar to telecommunication networks, or take on some hybrid architecture. It is likely that utilities will prefer having the control, intelligence, and innovation located within the core of their networks, as opposed to the edge of the network. This is because utilities' current regulatory model rewards it for increasing the costs of providing electricity by investing in infrastructure, as opposed to seeking efficiency-enhancing benefits from innovative products and services provided over the network. Conversely, IT vendors such as Cisco and IBM favor more intelligence at the edge with Internet protocol being their preferred communications protocol. For now, the regulatory model pushes toward utilities increasing their capital stock, and controlling the communications path and network attachment. Until regulatory institutions like price caps or other efficiency-encouraging regulatory plans

are instituted, the distribution utility will seek to control the smart grid communications path.

In addition to network design, the smart grid industry is puzzling over answering the question: What is the business case for smart grid? For example, demand response depends on solving a number of problems: (a) creating incentives for customers to participate through price or other signals; (b) establishing the appropriate regulatory regime regarding market structure, privacy, etc.; and (c) identifying the "killer application" that will ultimately spur demand response. Beyond demand response, near-term success of grid modernization efforts could probably focus on grid optimization and efficiencies, an area where existing broadband networks could play a role by serving as the middle-mile communications network for the utilities.

For now, no one knows what the smart grid business model looks like. Utility rate design historically biases toward average cost pricing; smart grid makes real time, marginal cost pricing possible. Utility cost recovery biases toward increasing capital stock and selling more electrons; smart grid, if built on existing broadband systems at the least cost, should mean fewer electrons are sold and lower utility capital stock. Utility cost recovery encourages integration forward into the home for energy management; smart grid enables modularity and unbundled access where consumers have more choice of providers, energy management systems, and can monetize their own data.

IV. Conclusion—"Smart Enough Grid"

The core challenge of smart grid is a regulatory one. At the initial level, smart grid is the not-so-simple, but done-before construction of a communications path across the electricity distribution system. At the deeper level, smart grid unsettles the current cost recovery, pricing and business model of distribution utilities. Further, the benefits of smart grid are not uniform across all electric systems, and smart grid must satisfy cost/benefit criteria before regulators are going to authorize cost recovery. There remains great promise to smart grid—its market-making, efficiency-enhancing, and consumer-empowering possibilities are real and profound. In the end, smart grid transformation will have to be an evolution where "smart enough" solutions are tried, and missteps have to be tolerated.

Endnotes

1 The interconnectivity of the transmission grid varies across the nation, and it is not a single, interconnected national grid. The East, West, and Texas constitute separate interconnected transmission grids. Interconnection between these grids, which operate at different frequencies, can only be accomplished through a direct current (DC) tie.

2 In most parts of the country, the peak demand occurs during the summer when demand for cooling is at its greatest. As air conditioning becomes more pervasive, utility summer peaking issues have become more acute.

3 This is exactly the same phenomenon as affects computer platforms. The Windows OS systems face more security threats than, say, a Mac OS or Linux system, in part, because there are so many more machines running Windows. In turn, virus and malware attacks are more likely to focus on the most pervasive platforms because they have more potential victims.

4 Though there are exceptions, state public utility/public service commissions regulate the natural monopoly distribution systems of investor-owned (private) utilities. Publicly owned distribution systems exist in many parts of the country, and rural America is usually served by user-owned distribution co-operatives. Public power and co-operative distribution systems do not usually face a state regulatory approval process for investments like smart grid. That said, investor-owned utilities serve the largest number of customers in the country and stand to gain the most from smart grid investment. Accordingly, the state commissions remain the most important institutions that will evaluate the prudence of smart grid.

5 For example, in Colorado PUC Chairman Binz's comments at a Gridweek conference in September 2009, he specifically indicated that in approving Xcel Energy's SmartGridCity project, he was very interested in privacy; data ownership; market structure (whether utilities would be able to provide services in the home); unbundling of the meter; interoperability and network neutrality issues. *See* Comments of Chairman Ron Binz, Colorado Public Utilities Commission, GridWeek, State and Local Collaboration: Enabling a Recipe for Success Panel, September 22, 2009.



The Silicon Flatirons Roundtable Series on Entrepreneurship, Innovation, and Public Policy Report No. 10[*]

State Regulatory Opportunities and Impediments to Smart Grid

Elizabeth Hartman and Kaleb A. Sieh,
Rapporteurs[**]

Executive Summary

On August 25, 2010, the Silicon Flatirons Center brought together a number of experts in the fields of economics, technology and policy, including public utility commissioners, economics professors, financial analysts, and policy advisors, to discuss regulatory issues facing smart grid deployment in the United States. The discussion was held at the University of Colorado School of Law and touched upon the economics of the smart grid, incentives for the variety of stakeholders in the smart grid effort, the technological challenges of implementing a "smart" electricity generation and distribution system and its constituent technologies, and the policy implications at both state and federal levels.

The theory of New Institutional Economics – an emerging area of economic thought that gives greater weight to the role institutions play in the operation of economic systems – may have significant relevance in this area. As cost of service regulated entities, utilities are not currently incentivized properly to increase efficiency through innovation, while state regulators face the real risk that at this stage smart grid technologies are not yet a proven investment, so for early adopters the costs could outweigh the benefits.

In terms of the technologies involved in the smart grid, one question is whether they should be open or closed and the implications this may have for near- and long-term competition, innovation and policymaking. There are also important privacy questions as to how to approach the large amount of personal data involved.

---

[*] The *State Regulatory Opportunities and Impediments to Smart Grid* event is part of the Silicon Flatirons Roundtable Series on Entrepreneurship, Innovation and Public Policy. Special thanks to Brad Feld, Managing Director of the Foundry Group, who sponsors the Roundtable Series. *Smart Grid* is the tenth installment in the series, following earlier discussions on: (1) Law 2.0; (2) Government 3.0; (3) Open Standards, Open Innovation, and the Rollout of IMS; (4) The Social, Ethical, and Legal Implications of Social Networking; (5) The Promise and Limits of Social Entrepreneurship; (6) The Private Equity Boom; (7) The Entrepreneurial University; (8) Rethinking Software Patents; and (9) The Unintended Consequences of Sarbanes-Oxley. The reports from those discussions can be found at http://www.siliconflatirons.org/publications.php?id=report.

[**] Silicon Flatirons Research Fellows

The smart grid effort will require both state and federal agencies to work together in a form of cooperative federalism to ensure effective implementation. While state governments are taking the leadership role in actually building out the grid, the federal government has an important role to play in both the funding and analysis necessary to support and promote the efficient development of a truly nationwide system. This is especially true within the broader context of global concerns like climate change, economic development and national security.

**

Introduction and Overview

As issues of climate change, economic growth and energy security become more important, the "smart grid" and implementation of smart grid technologies look to be an increasingly important part of addressing these challenges. Recovery Act funding available for smart grid projects in the United States makes this an especially opportune time to invest in energy infrastructure,[1] with policymakers playing a critically important role in determining how these funds can be most efficiently used to correctly incentivize stakeholders in the smart grid system.

In an effort to address some of these issues, on August 25, 2010 the Silicon Flatirons Center brought together a number of economic, technology, and policy experts to discuss issues facing the U.S. smart grid effort.[2] The discussion was held at the University of Colorado School of Law and touched upon the economics of the smart grid, incentives for the variety of stakeholders in the smart grid space, the technological challenges of implementing a smart grid, and the policy implications at both state and federal levels. The roundtable was a continuation of earlier discussions, but with an added focus on the issues regulators will need to address in order for the smart grid to "actually happen."

---

[1] Vice President Joe Biden announced almost $4 billion of stimulus act funding for smart grid projects in April 2009, as detailed in this government press release: http://www.energy.gov/7282.htm (last visited Jan. 23, 2011)

[2] The Roundtable was jointly sponsored by Silicon Flatirons and the Institute for Regulatory Law & Economics (IRLE). The discussion was also held according to the Chatham House Rule, which is used at meetings or discussions to encourage openness and the sharing of information. The Rule itself reads, "When a meeting, or part thereof, is held under the Chatham House Rule, participants are free to use the information received, but neither the identity nor the affiliation of the speaker(s), nor that of any other participant, may be revealed." For more information, see the Chatham House website at http://www.chathamhouse.org.uk/about/chathamhouserule/ (last visited Jan. 23, 2011). In Silicon Flatirons roundtable discussions and the resulting report or summary, the list of attendees and their affiliations is customarily published. Of note, in this discussion the participants were asked to indicate when the comments started to infringe on open proceedings that may be in front of the various government officials so as to protect against any potential *ex parte* requirements.

This report, tracking the structure of the roundtable itself, will proceed in three parts. Part I will discuss the economics of the smart grid effort, including incentives within the pricing system and what effect these can have on innovation and policy. Part II will discuss the technological challenges of implementing a smart grid and related issues. Part III will focus on the federal policy considerations and how these interact with state policy.

Part I – The Economics of Smart Grid

The theory of New Institutional Economics (NIE) is an emerging area of economic thought that gives greater weight to the role institutions play in the operation of economic systems.[3] In the body of writings that has become synonymous with NIE, writers have generally sought to extend neoclassical economic theory by examining how property-rights structures and transaction costs affect economic incentives and behaviors.[4] The NIE model has built up over time as greater institutional detail has been injected into economic models where economic theory may previously have been too abstract to deal with more modern questions.[5]

Viewing smart grid through the NIE lens, much of the initial discussion revolved around the institutions and incentives related to smart grid and its various stakeholders. Although these stakeholders include utilities, vendors, consumers, standards setting organizations, and both federal and state regulators, here the group focused primarily on

---

[3] *See* Eirik G. Furubotn & Rudolf Richter, *The New Institutional Economics: An Assessment*, 1-32, in THE NEW INSTITUTIONAL ECONOMICS: A COLLECTION OF ARTICLES FROM THE JOURNAL OF INSTITUTIONAL AND THEORETICAL ECONOMICS (1991) (introducing and describing New Institutional Economics).

[4] *Id*. at 1. The NIE literature includes, among many other writings: RONALD H. COASE, THE FIRM, THE MARKET AND THE LAW (1988); DOUGLASS C. NORTH, INSTITUTIONS, INSTITUTIONAL CHANGE AND ECONOMIC PERFORMANCE (1990); AND OLIVER E. WILLIAMSON, THE MECHANISMS OF GOVERNANCE (1996).

[5] *Id*. One way to define NIE is in contrast to neo-classical economic analysis,

> While it would be incorrect to say that traditional analysis abstracted completely from institutional structure, there can be little doubt that the usual treatment of institutions was superficial. The existence of political, legal, monetary and other systems was certainly recognized; but either these systems were regarded as neutral in their effect on economic events and ignored, or they were taken as given and then specified in so perfunctory a way as to suggest that institutional influence was not of much importance. By contrast, the new institutional economics seeks, at a minimum, to demonstrate that institutions truly matter. Each distinct organizational structure is said to affect incentives and behavior but, beyond this, the institutions themselves are regarded as legitimate objects of economic analysis. As Coase has argued, it is possible to use theory to analyze institutions so that their operation is explained and made an integral part of the economic model.

Id. at 2 (internal citations omitted).

utilities and state regulators. One theme that emerged was how significant economic challenges will need to be resolved in order to create an incentive structure that works to encourage prudent investment in smart grid development.

A. New Institutional Economics and the Incentives of Different Smart Grid Stakeholders

The roundtable began with a presentation by Ray Gifford, Senior Adjunct Fellow at Silicon Flatirons and a Partner at the law firm of Wilkinson Barker Knauer LLP, on the economics and incentives of regulated utility companies. He started with how the theory of New Institutional Economics applies to smart grid and smart grid technologies, focusing on the impact that institutions can have on economic decisions.

According to Gifford, New Institutional Economics (NIE) brings a different focus and level of sophistication to the micro-analytic questions that face institutions and regulators involved in smart grid policy. NIE shows that institutions and incentives matter, he said, while the smart grid itself confronts institutions, utilities and regulators with the question of how each will address technical innovation in the electric grid. Importantly then, the innovation that smart grid *may* represent – not necessarily an unqualified good – is the need for regulators and utilities to realize there will be different incentives for each stakeholder moving forward. Along these lines, he felt that one goal might be for regulators to adapt and change the incentives presented to utilities.

The New Institutional Economics' analytic premises, as Gifford related, include three fundamental assumptions: (1) bounded rationality – the concept of how humans do not behave in a perfectly logical or rational manner, both on an individual and institutional level; (2) opportunism – players in a market or a regulatory situation are opportunistic and the challenge for regulation is how to plan for that opportunism and think about how to channel it into a more beneficial form; and (3) asset specificity – once capital has been dedicated, especially to electric and communications infrastructure, those assets are not easily changeable and are committed for a significant period of time, on the order of decades and often up to half a century.[6]

Regulators must carefully consider asset specificity when making decisions about the proper time to authorize projects or approve construction in order to avoid costly mistakes, he said, especially if investments are made in a technology that is later abandoned. As a negative example, Gifford pointed to electricity generation "overbuilding" in the 1970's and said that in modern times no stakeholder wanted to be the first to deploy or build the "Commodore 64" of smart grid.

---

[6] *See* Oliver E. Williamson, *The New Institutional Economics: Taking Stock, Looking Ahead*, 38 J. ECON. LIT. 595 (2000), *available at* http://www.jstor.org/stable/2565421.

Comparing smart grid to another industry, Gifford felt the profound innovation cycles the communications industry has gone through might shed some light on the challenges facing smart grid. This comparison could be especially useful, he said, as policymakers move forward and face the challenge of implementing and regulating smart grid technologies in the face of innovation, uncertain information, and unclear benefits.

B. The Institutional Players

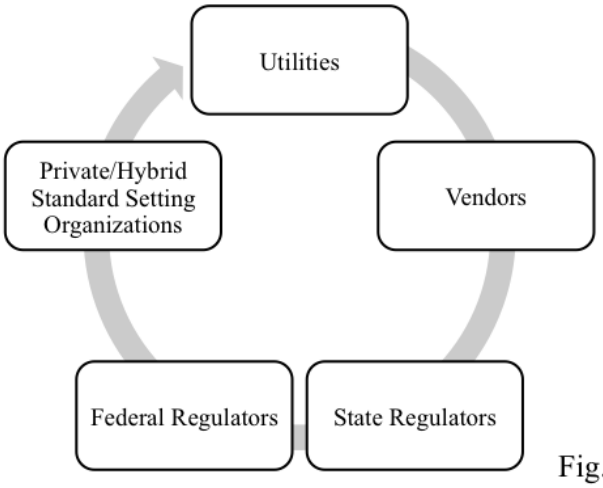


Fig. 1

New Institutional Economics is a helpful lens through which to analyze smart grid because it allows policymakers to examine the incentives and behaviors of all the stakeholders in the ecosystem, for example utilities, vendors, private and hybrid standards setting organizations, and both federal and state regulators alike (see Fig. 1).[7] Gifford analyzed the incentives of each in turn.

Utilities

For utilities, they have some incentives to innovate and change their business models when it comes to smart grid technologies, he said, which is a positive outcome and allows for new business opportunities. Additionally, the regulatory "wind" looks to be favoring smart grid efforts, so utilities have incentives to participate and ensure a guaranteed recovery of their costs.[8]

[7] For a more detailed analysis of the different stakeholders and vendors involved in the United States' smart grid "ecosystem," see DAVID CHENG ET AL., 2010 U.S. SMART GRID VENDOR ECOSYSTEM (2010), *available at* http://www.energy.gov/news/documents/Smart-Grid-Vendor.pdf.

[8] For example, the Federal Communications Commission's recent National Broadband Plan devotes its entire Chapter 12 to energy and the environment, and, among many other recommendations concerning smart grid, recommends

On the negative side, Gifford pointed out how most electricity utilities in the U.S. are currently regulated under a cost of service regime and as a result do not have strong incentives to invest in smart grid technologies or innovate in this area.[9] Despite the positive nature of what innovation and new business models could do for utilities generally, he said, most utility companies have corporate cultures that do not see innovation or changed business models as positive outcomes. Utilities may be motivated by the perception that the industry as a whole is moving towards innovation in the smart grid space, deciding to participate in order to avoid being left behind, but otherwise there are no real incentives to innovate in a cost of service regulated environment. As long as the new business model remains unclear, he said, utilities are not encouraged to change their existing systems.

Utilities are also facing the challenge of addressing the sunk costs – or in other words asset specificity[10] – associated with building out a smart grid. Continual iteration would be a helpful method of determining which particular smart grid technology is the most effective or efficient, but in the utility industry many construction and infrastructure decisions are in place for up to 50 years. This long horizon makes experimentation difficult and costly.

Along these lines, state regulators looking to create investment requirements in new technologies with long payback horizons or depreciation schedules should look to the standards that have already been adopted in leading smart grid states such as Texas and California. One participant mentioned that in California, for example, regulators may require each utility to file a smart grid "plan," which could help keep standards well-defined and maintained across the state.[11] In contrast

---

that "States should reduce impediments and financial disincentives to using commercial service providers for Smart Grid communications." *See* FCC, NATIONAL BROADBAND PLAN (2010), *available at* http://www.broadband.gov/plan/12-energy-and-the-environment/.

[9] While not discussed in depth, Gifford mentioned that there may be distinctions and differences in the incentives faced by utilities regulated under a cost of services regime versus those regulated under a performance- based regime.

[10] Asset specificity refers to physical investments that are specialized and unique to a task, or in other words are not easily redeployed elsewhere. *See* OLIVER WILLIAMSON, THE ECONOMIC INSTITUTIONS OF CAPITALISM (1985). This can be contrasted with an asset's "plasticity" or the range of uses to which an asset may be put. *See* Armen A. Alchian & Susan Woodward, *The Firm is Dead; Long Live the Firm*, 24 J. ECON. LIT. 65, 69 (1988), *available at* http://www.jstor.org/stable/2726609.

[11] The proposed regulation in California to review each smart grid plan individually was eventually modified to stipulate that all proposed plans would be reviewed in a single proceeding held twice annually, in order to improve efficiency and "help ensure some congruity." *See* California Public Utilities Commission, Decision Adopting Requirements For Smart Grid Deployment Plans Pursuant To Senate Bill 17 (Padilla), Chapter 327, Statutes Of 2009 (June 24, 2010) (quoting DRA Comments at 7), *available at* http://docs.cpuc.ca.gov/PUBLISHED/FINAL_DECISION/119902-03.htm#P773_139554 (last visited Jan. 23, 2011).

to these state-mandated guidelines and plans, utilities may also wish to define their own. Although some companies may opt for lower-cost plans while others choose a more gold-plated "Cadillac" version, participants felt that both can be accommodated as long as the standards are flexible and applicable to a wide variety of technologies.

Vendors

Vendors are also an important stakeholder, but Gifford pointed out they currently are not sure whether their customers are the utility or the end-user. If the proper incentive structure was in place, he said, the smart grid market could function more like the solar market with eager involvement of third party vendors. Additionally, smart grid vendors are approaching what looks like a potential "standards war," involving traditional information technology standards-setting issues found in computer operating systems, smart phones, and other technologies.

State Regulators

Returning to the theme of New Institutional Economics, Gifford asked whether the current state regulatory incentives are adequate for the more efficient energy consumption that smart grid represents and, importantly, how can those incentives be modified to accommodate improved outcomes? There are real and compelling reasons to encourage smart grid, he said, including net system savings and environmental benefits. However, there are also risks to pioneering new technologies and no state has yet emerged as the clear leader in innovation and deployment, albeit California appears to have a head start. The city of Boulder, Colorado is also participating in a smart grid project, but some of the recent negative press it has received may call into question the intelligence of pioneering these technologies. As Gifford pointed out, the smart grid does not necessarily meet the "proven investment test" and it is unclear that the costs outweigh the benefits for early adopters. One potential solution, he said, might be price cap plans, which have been successful in other parts of the world.[12]

[12] According to one academic,

> Price cap regulation allows the operator to change its price level according to an index that is typically comprised of an inflation measure, I, and a "productivity offset," which is more commonly called the X-factor. Typically with price cap regulation, the regulator groups services into price or service baskets and establishes an I - X index, called a price cap index, for each basket. Establishing price baskets allows the operator to change prices within the basket as the operator sees fit as long as the average percentage change in prices for the services in the basket does not exceed the price cap index for the basket.

See Mark A. Jamison, *Regulation: Price Cap and Revenue Cap*, in ENCYCLOPEDIA OF ENERGY ENGINEERING AND TECHNOLOGY, VOL. 3, Barney Capehart, ed., pp. 1245-51 (2007), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=959684.

Federal Regulators

One of the major issues facing federal regulators is their lack of jurisdiction over electricity distribution systems, something which is clearly more in the domain of the states. Although the federal government can play an important "convening" and "bully pulpit" role, Gifford felt that the costs and the benefits of smart grid implementation are not uniform across distribution systems (as these can vary according to where the utility sits in its life cycle of assets and what the utilities load profile looks like). Finally, an additional challenge for the federal government is that most utilities are "long" on generation with significant excess capacity, meaning that efforts or technologies that make electricity generation more efficient – and thus reduce customer demand – are not necessarily attractive.

Standard Setting Organizations (SSOs)

Related to both federal and state regulatory issues is the question of standard setting, as it can be one of the important precursors to widespread deployment of smart grid and smart grid technologies. According to Gifford, both utilities and regulators feel the industry is beginning to congregate around a certain set of standards. California, for example, has trended towards generally more open standards, which from both a consumer and regulatory perspective is "safer" overall than closed proprietary standards. Outside of California, he said, utilities seem to generally have a preference for closed proprietary standards, but momentum towards open standards is building and should likely continue.

Gifford concluded his presentation with a comparison to the Ptolemaic astronomy model. He pointed out how Ptolemy theorized the Earth was at the center of the universe and the retrograde motion of the planets in the sky could be explained by epicycles, or in other words a series of nested spheres; this turned out to be an interesting but extremely complicated model.[13] Creating a similarly complex structure for the smart grid, according to Gifford, would result in unnecessary challenges for all stakeholders involved. It is essential that policymakers seriously examine incentive structures that encourage the utility to become the center of the smart grid system. On the one hand then, regulators should give utilities the freedom to make those decisions, while on the other hand remain vigilant in ensuring the incentives faced by utilities in other areas do not "fall apart."

Finally, as the group moved into more broad discussion, Gifford posed a series of questions. First, how can the relevant institutions (utilities, regulators and vendors) be incentivized to invest prudently?

[13] Ptolemy was largely responsible for the "geocentric cosmology" of ancient times, though he was not the first to propose such a model. For more on the life of Ptolemy and the Ptolemaic cosmological model see, Encyclopedia Britannica, *Ptolemy*, ENCYCLOPEDIA BRITANNICA ACADEMIC EDITION, http://www.britannica.com/EBchecked/topic/482098/Ptolemy (last visited Jan. 23, 2011).

Second, how can the closely regulated electricity market accommodate experimentation, failure, and other difficulties in creating standards? Third, how can utilities move away from the engineering model where institutions have centralized control to one of a more loosely distributed network of transactions between generators, distributors, and customers? Finally, how must regulatory institutions adapt to internalize the cost to utilities of smart grid investment and allow both utilities and customers to realize the potential benefits?

C. Discussion

The general discussion started off with one participant asking whether customers, or "end-users," should be included as relevant stakeholders in the smart grid decision-making process. Many other participants thought efforts in this space might not have credibility without the consumer as a main focus. Here, Gifford felt it essential to stop treating consumers as "passive agents" during the regulatory process, instead encouraging them to become more engaged at a transactional level, where he felt their preferences are crucial.[14] A critical part of engagement, he said, will be consumer education, so they can easily understand what has changed and make informed decisions. Considering the different degrees of consumer sophistication, Gifford quipped that the end result of smart grid efforts might actually be the "Smart *Enough* Grid," which would allow different levels of engagement.[15] For example, a very sophisticated consumer might be an entity like Wal-Mart, who will want to take advantage of every available option in order to reduce their electricity costs, while some individual consumers may be less interested in learning about and using some of the more complicated technologies and pricing regimes, but more concerned with stability and ease of use.

The participants generally agreed that any incentive structure failing to consider the needs of the consumer or end user would necessarily be incomplete. One participant wondered if there would ultimately be enough economic benefit for consumers to see value in the smart grid, while another pointed out that even if there was no real

[14] For more on customer participation in the smart grid and how this relates to end-use pricing, see Christopher Russo & Richard Tabors, *Does a Smart Grid Need Smart Customers? The Debate Over End Use Pricing*, Smart Grid News, http://www.smartgridnews.com/artman/publish/Business_Markets_Pricing_News/Does-a-Smart-Grid-Need-Smart-Customers-The-Debate-Over-End-Use-Pricing-1249.html (last visited Jan. 23, 2011).

[15] Here, Gifford was referring to what some have called the "good enough revolution" where over the past decade as the technology and features of products have reached a certain threshold level consumers have begun to prefer "flexibility over high fidelity, convenience over features, quick and dirty over slow and polished" and "[h]aving it here and now is more important than having it perfect." *See* Robert Capps, *The Good Enough Revolution: When Cheap and Simple is Just Fine*, WIRED MAGAZINE, http://www.wired.com/gadgets/miscellaneous/magazine/17-09/ff_goodenough (last visited Jan. 23, 2011).

economic benefit, there may be strong enough prevailing social norms to motivate responsible citizens to participate in smart grid anyway. There may also be new performance capabilities that influence adoption, as economics should not be the only consideration in terms of creating new markets and policies. Further discussion centered around the idea of some kind of benefit or tax credit that could be offered to consumers who switched to smart grid, with the Saver's Switch program in Colorado mentioned as an example.[16] Here though, there were some concerns with introducing potential distortions to the marketplace, which according to some could discourage innovation stemming from the price of electricity being brought to the level of the actual costs of the system.

Another participant wondered if approaching the smart grid analytically and breaking it down segment-by-segment would change the incentives. For example, what if the entire electricity grid was divided into transmission, distribution, and generation rather than divided by individual stakeholder. The group agreed that incentives were complex and that analytically, a middle-mile and last-mile smart grid are different propositions, with many seeing the middle-mile as less difficult from a regulatory perspective.

One participant reiterated that utilities need to be incentivized to internalize the investment proposition, pointing out that currently we have exactly the opposite: in a regulated cost of service environment, the utilities would actually lose money if they became more efficient. For example, he said, currently if a utility sees a 5% efficiency gain through investment in better and more efficient equipment or infrastructure, they actually lose money on that particular capital expenditure, even though there are long run gains in efficiency. Here, one participant thought the driving incentive, at least in California, is whether investment in smart grid technologies is included in the rate base of the utility.[17] Often, according to the participant, if it is in the rate base, the actual performance of the technology is less relevant. In Boulder, he said, the smart grid investment was not in the rate base of the utility.[18]

---

[16] The Saver's Switch program in Colorado operates during the summer and allows the utility, during peak times and by remote control, to "cycle a customer's central air-conditioner compressor units on and off at 15-minute intervals." Those customers who are part of the program receive bill credits from the utility for helping it to "avoid purchasing high-priced electricity from other utilities during peak-use periods or building new power plants." Xcel Energy, Press Release, http://www.xcelenergy.com/Minnesota/Company/Newsroom/News%20Releases/Pages/Denver_Xcel_Energy_to_boost_Saver_s_Switch_participation.aspx (last visited Jan. 23, 2011).

[17] The rate base is very important in determining the profitability of a utility. It is defined as the "total fair value of public utility property that is used in rendering services and that comprises the investment on which a fair rate of return is based in setting utility rates." *See* Miriam-Webster's Dictionary of Law (1996), *available at* http://dictionary.reference.com/browse/rate+base.

[18] Though Xcel Energy has sought further reimbursements for the project and some critics have contended that any additional payments to the utility are

In addition to the rate base, participants mentioned several other regulatory solutions for managing incentives and encouraging innovation, including dynamic pricing,[19] bottom-up pricing,[20] price caps, experimental economics,[21] and a suite of applications designed to allow the utility to provide a number of basic services built around ultra-efficient homes and appliances or devices. Price caps were of significant interest to the participants and were seen as allowing utilities to keep their current profit margins while at the same time incentivizing efficiency. However, there was concern that incentivizing too much efficiency could create a shift to "starving the network" rather than some of the current incentives to "gold plate" it, and thus could require regulations to ensure quality control. One participant pointed out how there may be lessons in other countries' experience with price caps.

Finally, the conversation turned to consumer benefits, where one participant felt there is no demonstrated business case for smart grid technologies. It is hard to "anticipate" consumer benefits, he said, and there is a bit of a chicken and egg problem. Incremental steps are being taken, but there may need to be some sort of "idealism" that helps to provide momentum behind the development of the platform. In order to realize the benefits, he said, it is necessary to open up the network and allow broad innovation on the platform, and this should set the stage for a wide range of unexpected benefits. As an example, he pointed to the Internet and how it has provided an enormous amount of consumer benefit that a regulator would not have been able to predict *ex ante*, or before the innovation on the open platform actually occurred.

---

unwise as the smart grid project in Boulder is "a research-and-development project and its full costs should be borne by Xcel shareholders, not ratepayers." See Mark Jaffe, *Xcel, critics await Colorado PUC's smart-grid rate ruling*, Denver Post (Nov. 28, 2010), http://www.denverpost.com/business/ci_16721011 (last visited Jan. 23, 2011).

[19] Under dynamic pricing, utilities pass through the wholesale costs of electricity to consumers on an ongoing basis by letting retails prices fluctuate based upon wholesale costs. Moving to a system of this nature may require significant and costly changes in the metering infrastructure nationwide, but should lead to reductions in household electricity demand. *See* AHMAD FARUQUI AND SANEM SERGICI, HOUSEHOLD RESPONSE TO DYNAMIC PRICING OF ELECTRICITY—A SURVEY OF THE EXPERIMENTAL EVIDENCE (2009) (describing the results of fifteen different dynamic pricing experiments), *available at* http://www.hks.harvard.edu/hepg/.

[20] Bottom up pricing is where the rate maker considers "all [the] components of a utility rate for possible competition, including transmission, distribution, customer service, metering, marketing, taxes, etc. as well as the energy component (which is generally considered the "top" of the rate)," whereas in contrast a top-down approach "involves an examination in the reverse order and typically results in only the energy component becoming subject to competition, instead of all aspects of the utility rate." *See* Glossary of Energy Market Terms, http://www.energybuyer.org/glossaryAB.htm (last visited Jan. 23, 2011).

[21] Experimental economics is a somewhat recent development in the field of economics and focuses on "the systematic evaluation of economics theories under controlled laboratory conditions." *SEE* DOUGLAS D. DAVIS & CHARLES A. HOLT, EXPERIMENTAL ECONOMICS (1992).

One participant felt that many benefits of smart grid are "upstream" and difficult to quantify. He pointed to improved "smart" appliances and how many of them require "smart rates" for their full efficiency benefits to be realized. In this area though, he said, manufacturers are having trouble deciding whether the utilities or the consumer is their market. Another participant felt that if overall efficiency was the goal, investment should be in overall system efficiency rather than in individual smart grid technologies because of the potentially greater benefits to this approach. He thought that retrofitting utility customer facilities or buildings would have a larger effect than simply installing smart appliances. Additionally, participants voiced some concern that appliance manufacturers were hindering the smart grid effort at the state level by not being as active in state proceedings as they are at the national level, and in this respect not helping to create the evidentiary record necessary for changes at the state level.

Consumer backlash was another concern. As one participant related, in California and other places there has been resistance and consumers have "opted out" of pricing structures, with some voters turning against smart meters and inverted block rates.[22] One reason might be that consumers can sometimes distrust change generally, but another reason might also be that consumers often end up paying more for smart grid technologies, at least in the short run. Additionally, the new systems might be too complicated for consumers who want a simple, consistent solution. But not all consumers are the same, so while some may enjoy the added choice that smart grid can offer, others may prefer to stick with the system they are already using. This is part of both the value and the cost of the smart grid: the ability to segment consumers – both businesses and individuals – and move from a one-size-fits-all mentality to an understanding of how the value may vary significantly based on the type of consumer.

Speaking again to the business case, one participant pointed out one significant benefit of smart grid technologies, which is reducing the need to build new electricity generation facilities. Some participants felt this was not a tangible or large enough benefit to change perceptions. Also, some participants were concerned there may be less of a savings in this area – or at least the savings would be spread out over a longer time horizon and be less impactful – since most utilities are currently "long" on electricity generation and have excess capacity as is.[23]

---

[22] The consumer backlash in California has been broad and sustained. *See* Tom Zeller Jr., *Smart Meters Draw Complaints of Inaccuracy*, NEW YORK TIMES (Nov. 12, 2010), http://www.nytimes.com/2010/11/13/business/13meter.html?_r=1&ref=business (describing how California residents have turned against smart meters because of inaccurate measurements of power usage and the resulting increase in utility bills) (last visited Jan. 23, 2011).

[23] Although most utilities may have excess capacity for normal demand cycles, they are not necessarily long on "peak" generation – the additional and at times significant electricity generation requirements at times of peak demand. Smart

From the discussion of the benefits and business case for smart grid technologies, the conversation then moved to the technology considerations.

Part II – Technology Considerations of Smart Grid

The technologies used in building out the smart grid will have significant implications for its architecture, both in terms of capabilities *and* limitations. Since electricity is an integral part of almost all citizens' daily lives – from charging a cell phone to surfing the Internet to viewing television and listening to music – and as more and more of our civic lives and the governmental process move online and onto devices that require electricity to operate, where and how electricity is delivered is becoming vitally important. How we choose to architect the technology will have significant implications in other areas as well, such as policymaking.

A. Signaling Networks in Smart Grid Distribution Systems

The second segment of the discussion began with a presentation by Dale Hatfield, Executive Director of the Silicon Flatirons Center, Adjunct Professor at the University of Colorado, and former Chief Technologist at the Federal Communications Commission, where his self-acknowledged goal was to place the smart grid effort into a larger historical and technological context.

Hatfield began by giving some background on smart grid and some general information on the electricity grid. He pointed out that there were four major components in the electricity grid: (1) electricity generation; (2) transmission; (3) distribution; and (4) the customer premises' network and devices. He said there was no single or formal definition of smart grid, as it is sometimes used as a generic term that describes applying computer intelligence and networking to otherwise "dumb" electricity systems, and other times it is used to describe a modernized grid that enables "bidirectional flows of energy and uses two-way communication and control capabilities that will lead to an array of functionalities and applications."[24] Some of the goals of the smart grid, according to Hatfield, are to improve the "reliability, security, and efficiency" of the electric system, by improving the delivery systems

---

grid may help to better balance the high demand that can cause blackouts. *See* Mike Orcutt, *How a Smarter Grid Can Prevent Blackouts*, POPULAR MECHANICS (Aug. 6, 2010), http://www.popularmechanics.com/science/energy/efficiency/how-a-smarter-grid-can-prevent-blackouts (last visited Jan. 23, 2011).

[24] *See* NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, FRAMEWORK AND ROADMAP FOR SMART GRID INTEROPERABILITY STANDARDS, RELEASE 1.0 AT 13 FN. 6 (2010), *available at* http://collaborate.nist.gov/twiki-sggrid/pub/SmartGrid/IKBFramework/NISTFrameworkAndRoadmapForSmartGridInteroperability_Release1final.pdf.

of large electricity generation systems and by improvements and innovations in distributed-generation and storage.

Hatfield explained that although there are islands of intelligence within the electricity network today, the distribution component of the grid has been largely lacking, making it necessary for power companies to rely on customer-reported outages when a disruption in service occurs rather than on automatic alerts by the network itself. In order to better illustrate how this distribution component could become more intelligent, Hatfield felt some examples from other industries might be helpful.

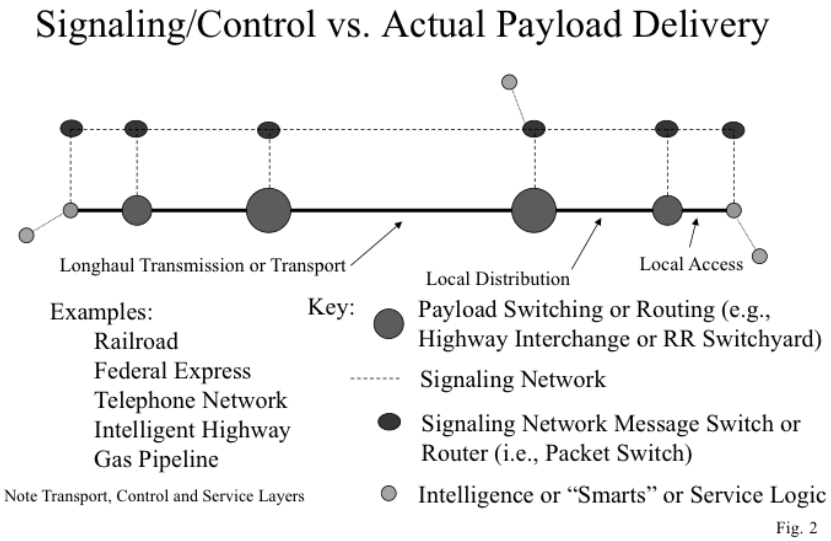


Fig. 2

Along these lines, Hatfield said it is important to distinguish between the two core elements of any network: the signaling and control system versus the actual payload delivery system (see Fig. 2). For instance, in the railroad industry and its railroad track network there are the rails themselves, which are for the purpose of actually moving the trains – or in other words the payload delivery system – and there are the parallel telegraph lines that run along beside the rails to coordinate communications concerning which train should go where and at what time – or in other words the signaling system.

Traditional Telephone Network

Logical Components: (1) Signaling network (and associated logic) and (2) Conversation network

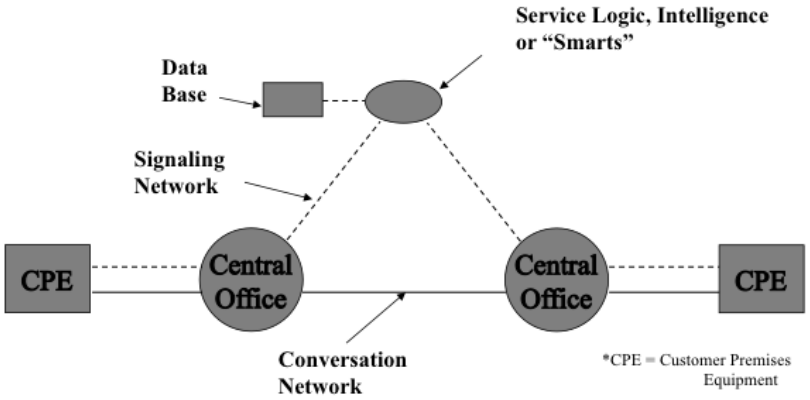


Fig. 3

The telephone network operates similarly, he said, and has lines in place to both carry actual phone calls and the attendant signaling network that determines where the call should go (see Fig. 3). This signaling network, according to Hatfield, is really the nervous system of the whole organization and determines the ability of the network to deliver calls intelligently and efficiently. As Hatfield related, the signaling network and its associated assets were key strategic assets of the traditional telephone industry prior to the rise of the Internet. The phone network is a good example of a relatively intelligent system that is able to accomplish complex tasks. Introducing computer power to the telephone database made the network even more powerful and allowed the routing of calls to the correct geographic location, time zone, and even language depending on how much information is known about the caller.

According to Hatfield, the intelligence of the signaling network is a critical component of any system's efficiency and competitiveness, which has important implications when deciding whether to use a closed proprietary signaling system versus an open platform that many different operators can use to freely develop innovations. Policy that allows open access to the signaling network for the smart grid would encourage strong innovation and competition in the industry. In Hatfield's opinion, the electric industry is comparatively late to the game in terms of adding signaling capability to their system, considering how railroad engineers were able to add this critical element to their network more than a century ago.

Another industry that is ahead of the power grid in terms of signaling capability, Hatfield related, is the mail and package delivery

business, where FedEx arguably revolutionized overnight package delivery services by introducing a dedicated signaling network. This parallel signaling and control network dramatically improved the efficiency and reliability of a physical payload delivery network that had existed for decades. The increased efficiency gave FedEx a competitive edge that forced UPS to quickly adapt by making deals with cellular network providers to create its own improvised signaling network.[25]

Turning back to the main topic, Hatfield pointed out how the electric power industry has been slow to fully adopt the notion of a parallel signaling and control network, especially in terms of distribution. The extension of a parallel digital communications (or signaling) network from power generators to individual end users – or in other words the smart grid – has the potential to revolutionize the electric industry.

Working with existing signaling networks to bring intelligence to an unrelated content delivery system, as UPS did with cellular phone network providers, is a strategy that could prove useful to the electric industry as well. Existing networks like cable, Internet and cellular are already in place connecting US homes, and utilities could leverage these networks to improve intelligence and efficiency in the electric grid.[26]

When implementing a signaling network to gain increased efficiency in a delivery network, Hatfield said that a key consideration is

---

[25] As described in their corporate history, UPS added tracking through wireless networks in 2000. *See* United Parcel Service, Company History 2000-2007, http://www.ups.com/content/us/en/about/history/2007.html (last visited Jan. 23, 2011).

[26] As an aside, Hatfield felt that there were some other networks in the U.S. that might be able to use this same technique. For example, the U.S. highway system could utilize existing communications networks to bring greater intelligence to their transportation grid, creating wireless alert systems for accidents or congested traffic ahead so drivers are able to reroute their path or adjust their speed accordingly. Of course, the notion of an intelligent highway system has been around for quite some time and the subject of a large of amount of research by the US Department of Transportation and others, see e.g., U.S. Department of Transportation, Research and Innovative Technology Administration, Intelligent Transportation Systems, http://www.its.dot.gov/ (last visited Jan. 23, 2011). There is still plenty of room for improvement though, see Joyce Wenger, et al., *The Smart Highway: A Smart Idea?*, Strategy and Business, Feb. 26, 2008, http://www.strategy-business.com/article/li00064?gko=9148d (last visited Jan. 23, 2011. Just like with the electric grid, Hatfield felt this limited highway network intelligence is rather conspicuously behind the sophistication of other networks such as the gas pipeline system that alerts oil companies automatically if there is a rupture in the pipe anywhere along the thousands of miles that it runs through sometimes harsh and remote environments. Another pipeline system that is considering implementing network intelligence is the water utility industry. For more on the smart *water* grid, see Devin Coldewey, *i2O: An Intelligent Grid For Water Systems That Could Save Millions Of Gallons*, TECH CRUNCH (Oct. 15, 2010) http://techcrunch.com/2010/10/15/i2o-an-intelligent-grid-for-water-systems-that-could-save-millions-of-gallons/ (last visited Jan. 23, 2011).

where the intelligence of the system is located: in the center, at the edge, or within the devices that are actually part of the network itself. For the electric grid, this means that, with suitable instrumentation in the house, monitoring and control of power usage could occur: (1) at the power company itself; (2) at each individual home; (3) or within each individual appliance . The question of where the intelligence or control functionality is located is important because it relates to competitive situations where one entity may have an advantage in terms of controlling the signaling network data.

Concluding his presentation, Hatfield described where some of the trends for locating intelligence within the smart grid are going and how, moving forward, policy makers can help work towardsan optimized and efficient system architecture. He pointed out how, similar to the way in which the intelligence in communications networks has migrated over time from inside the network to, in the case of the Internet, the edge of the system and into each individual computing devices, the trend in the electric grid may also be moving intelligence towards the edge and into houses and appliances, rather than towards a central operating station. Moving forward, he said, the task for policymakers will be to apply the historical lessons of parallel intelligence network systems to best define the interfaces and protocols that will work for the smart grid system.

According to Hatfield, how the U.S. goes about defining this network will have enormous consequences and whether it is either open or closed, proprietary or "open source," and centralized or with intelligence located at the edge will have a tremendous impact on future innovation.[27] He said it is critical that the system architecture is optimized for efficiency in the future.

B. Discussion

Following the presentation, the floor was opened to discussion. Picking up where Hatfield had left off, the first question asked where the best place to house the intelligence in the smart grid network might be. Hatfield replied it was difficult to say with certainty where the "best" location is, but that he felt flexibility was a key consideration to allow for

---

[27] By statute the development of a smart grid has been made a national policy goal in the U.S., and Energy Independence and Security Act of 2007 specifies that:

> [T]he interoperability framework should be "flexible, uniform, and technology neutral." The law also instructs that the framework should accommodate "traditional, centralized generation and distribution resources" while also facilitating incorporation of new, innovative Smart Grid technologies, such as distributed renewable energy resources and energy storage.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, FRAMEWORK AND ROADMAP FOR SMART GRID INTEROPERABILITY STANDARDS, RELEASE 1.0 (2010), *available at*: http://www.nist.gov/public_affairs/releases/upload/smartgrid_interoperability_final.pdf.

adjustments or the possibility of a combination of locations in the future.[28] Another benefit of increased intelligence in the grid would be to link electricity demand to the distribution and generation systems in real time with the attendant rate schedules. Some participants felt the main challenge boils down to a question of timing: the new rate schedules are only beginning to develop, so the first order of business will be to implement the technology and then the rates can be passed on to consumers (with the hoped-for result of shifting consumption from high-cost to lower-cost times of the day). There is money on the table here, according to some participants, especially if these rate schedules can be properly managed through consumer education and the changes in consumption patterns actually occur. One difficulty is the number of bigger players making significant profits with the current energy-pricing regime, leading some participants to comment that there may be pushback to the rate schedule idea.

The discussion then turned to the architecture of the signaling network, focusing on what structures are already in place that could be utilized as part of the smart grid system. One participant mentioned the cable network, focusing on how it is a nationwide communications network with broad penetration. Participants also mentioned the use of wireless technologies and radio frequencies, though using radio spectrum for smart grid technologies would require working with the Federal Communications Commission (FCC). Referring back to the concept of system architecture as policy, one participant wondered how regulators would unbundle the services each network offers to the extent necessary to support innovation and healthy competition, and of course to make sure that these stayed unbundled.

There was also some discussion as to how to best organize the business arrangements and agreements on some of the existing networks to overcome potential regulatory and economic challenges. Here, there were questions as to the economic incentives inherent in the way these partnerships are structured. Unlike innovative markets where first movers stand to gain from additional rents, for example in the smart phone industry where early adopter customers are willing to pay premiums for the first models, utility regulation does not permit discrimination in rate classes and it reduces the economic incentive to innovate.

There may be different incentives for the consumer and utility companies as well, with some questions around who will ultimately pay for the smart grid effort and whether the horizons of the different stakeholders are in sync. Here, one participant felt the near term benefits and business case are mostly on the side of the utility companies, but that consumers and the federal government are the ones being forced to pay the upfront costs. He wondered if this was one cause for the consumer backlash that had been seen, and if the utility companies should be

[28] As a clarification, Hatfield said that when using the word "intelligence" he was primarily referring to "routing" and the ability to quickly and automatically redirect electricity around a failure in the grid.

forced to justify the increases in the rates charged to customers. Another participant agreed that there might be more benefit on the side of utility companies and felt there should be a way for the utility to pass this benefit on to the consumer. One possible solution, mentioned previously in the roundtable, is price caps.[29] In comparison to cost of service regulation, according to one participant, price caps have shown successful results in the UK and Canada, suggesting that US regulators would do well to consider them.[30] One final comment pointed out that there might be differences in the incentives of the regulated versus deregulated utility companies, and that innovation may come first in the deregulated areas.

In addition to the economic incentives around innovation, participants felt the ownership model for the smart grid signaling system may be important as well. Once the intelligence network is in place, it can also serve other uses, for example connecting the elderly to a medical alert network in case they need assistance. If the utility is the owner of a proprietary, closed signaling network, they may or may not allow the system to be used for these other purposes, which is why an open platform would be preferable for purposes of innovation and maximizing the benefits to society generally.

Another participant pointed out how utilities have been conditioned over decades to favor stability over innovation in their corporate culture, adding a further challenge to regulators. Along these lines, some participants thought that collaboration among government agencies involved in the smart grid might be helpful. Here, many thought the Federal Energy Regulatory Commission (FERC) might have a role in generally helping utility companies to develop a new culture that is more supportive of innovation.[31]

Finally, one participant asked if separating out the intelligence in the signaling network versus the payload delivery system was a proper goal. He said the intelligence seems to be focused on the information collected and used for the routing decision, whereas the information and intelligence as to the payload is really the customer's decision at the edge of the network. Possibly, he said, this means that what is needed first is more of a smart *information* system and that any upgrade to the physical facilities should come later. The participant thought the proper focus instead might be building out an information system through the use of smart meters and other similar technologies.

---

[29] Though one participant thought that price caps would require the regulator to "vigilantly" enforce quality of service requirements.

[30] For more information on price caps in the UK, see Richard Green, Has Price Cap Regulation of UK Utilities Been a Success?, Public Policy For The Private Sector (The World Bank Group) (Nov. 1997), *available at*: http://cdi.mecon.gov.ar/biblio/docelec/bm/ppps/N132.pdf.

[31] The Federal Energy Regulatory Commission's general mission is to "[a]ssist consumers in obtaining reliable, efficient and sustainable energy services at a reasonable cost through appropriate regulatory and market means." *See* Federal Energy Regulatory Commission, http://www.ferc.gov/about/about.asp (last visited Jan. 23, 2011).

There was a final brief discussion concerning smart meters that focused on the difference between implementing Advanced Metering Infrastructure (AMI), which enables two-way communications with the meter, versus the existing system of Automatic Meter Reading (AMR).[32] Participants felt that utility companies would like the ability to communicate with meters using AMI to control energy use, which could result in real operational savings, but the downstream consumer benefits are not yet clear.

Part III – Federal Policy Implications

The third and final part of the discussion examined the federal and state policy considerations of the smart grid, where there may be an opportunity for both levels of government to work together in a form of cooperative federalism that should result in the optimized implementation of new technologies. While state governments are taking the leadership role in actually building out the grid, the federal government plays an important role in providing both the analysis and funding necessary to support efficient development of the system nationwide, and also placing the smart grid effort within the context of the larger global concerns of climate change, economic development and national security. Additionally, consumer participation in the smart grid implementation process is a critical component.

A. Roles for Federal versus State Governments in Smart Grid Implementation

There is no question that the energy industry is behind others as far as using information to become more efficient. According to Phil Weiser, Senior Advisor to the Director for Technology and Innovation at the National Economic Council, the Department of Energy is positioned to help shape the evolution of these technologies. However, in Weiser's

---

[32] According to the Electric Power Research Institute:

> Advanced metering systems are comprised of state-of-the-art electronic/digital hardware and software, which combine interval data measurement with continuously available remote communications. These systems enable measurement of detailed, time-based information and frequent collection and transmittal of such information to various parties. AMI or Advanced Metering Infrastructure typically refers to the full measurement and collection system that includes meters at the customer site, communication networks between the customer and a service provider, such as an electric, gas, or water utility, and data reception and management systems that make the information available to the service provider.

See ELECTRIC POWER RESEARCH INSTITUTE ADVANCED METERING INFRASTRUCTURE (AMI) (2007), *available at* http://www.ferc.gov/eventcalendar/Files/20070423091846-EPRI%20-%20Advanced%20Metering.pdf.

opinion, there are a couple of principal substantive challenges to implementing the technology.

First, the smart grid needs to be viewed in an appropriate context; it will not, for example, revolutionize society on the order of the Internet. Referring to any new technology as the "next Internet," he said, is an unfair comparison, as the Internet revolutionized everything from banking to entertainment and it is therefore unlikely that any new technology will soon appear and have such a dramatic impact on human social realities. Second, Weiser also pointed out that the term "smart grid" often ill defined or at least under-defined. Referring to the smart grid as a precisely defined concept is imprecise, as there are actually many different components that are not necessarily equal when it comes to the cost benefit analysis. At a minimum, smart grid includes transmission, distribution, and customer premise devices, and more broadly includes a set of smart grid technologies. Smart grid technologies are not, stated simply, merely smart meters, as sometimes is depicted. Finally, despite the issues with expectations and definitions for smart grid, the promise of smart grid technology is real and critical to the electricity network of the future, which will face challenges such as the increased use of renewable energy sources, management of peak demand, and the accommodation of electric vehicles.

Implementing an efficient smart grid will be critical in managing these challenges and others. Initially, Weiser pointed out that renewable energy supplies are variable in their supply characteristics as compared to coal or natural gas. Renewables are highly sensitive to weather conditions and other factors that can vary drastically – sometimes on a daily basis – so for stability purposes the United States will likely need to continue to use more reliable sources in the generation mix. The use of renewables could increase significantly if there were better storage technologies capable of capturing the renewable energy supply, but unfortunately the technology has not quite reached that threshold yet.

One of the things smart grid technology can do is to help stabilize the energy supply by helping to clarify the size and extent of demand so that the supply can respond appropriately. These new technologies may also serve as a final means of quantifying exactly how much capacity is necessary in regions where additional generation capacity still needs to be constructed. One looming challenge is electric vehicles; there will be a large number of people charging their vehicles at the same time of day, for example when they return home from work, creating new and excessive demand on the grid during peak times.[33]

[33] A University of California Berkeley study models the effects of plug-in vehicle charging on the grid under a number of different scenarios and assumptions, with most scenarios showing significant stress on current electricity generation and distribution systems. *See* NICHOLAS DEFOREST, ET AL., IMPACT OF WIDESPREAD ELECTRIC VEHICLE ADOPTION ON THE ELECTRIC UTILITY BUSINESS – THREATS AND OPPORTUNITIES 9-15 (2009), *available at* http://cet.berkeley.edu/dl/Utilities_Final_8-31-09.pdf.

The challenges of smart grid are certainly significant, and the federal government will need to work closely with state governments to meet those challenges. According to Weiser, there are three primary institutional issues that will impact the approach to smart grid implementation. First, state governments are in a leadership position in terms of actually implementing smart grid policies, while the federal government is in a position to provide the states with better research and analytics to understand what is working and what is not. He said that one significant obstacle states will face is a legacy regulatory system that values stability and constancy rather than innovation and change, which as discussed previously is part of a culture that is reflected in the utilities themselves. As California deals with the challenges of implementing a statewide smart grid and the city of Boulder, Colorado deals with similar obstacles at the municipal level, he said it is important to recognize the importance of learning from lessons that comes from early experiences.

Second, according to Weiser the federal government can catalyze innovation by providing funding.[34] Pointing to some of the lessons learned in the energy projects funded as part of the Recovery Act stimulus package, he felt the federal government should be able to: (1) take leadership in standard setting; (2) develop recommendations for best practices; (3) provide encouragement for experimentation; and (4) create a structure to analyze technological and marketplace developments. Working together on smart grid, he said, the state and federal actors can develop a form of cooperative federalism that both respects and embraces state authority and the opportunities for experimentation, but also benefits from federal guidance and support where there is a comparative advantage at that level. Weiser felt this type of cooperative federalism will be essential for finding the optimal approach to smart grid technology.

Federal and state regulators may also be able to work together on smart grid by using a funding sequence similar to the stages that venture capital firms use when investing in start-up companies, with the first round of funding coming from the federal government followed by later rounds by the states.

Consumer participation in the smart grid implementation process is also critical. Weiser felt that dynamic pricing does not have to be real-time, but utilities do need to provide some degree of signaling and an opportunity for consumers to react to those signals. One technique is giving customers the opportunity to "set it and forget it" when it comes to their power usage settings. Approaches such as this are critical to engaging the consumer and evaluating what technologies are most cost

[34] For example, as part of the American Recovery and Reinvestment Act of 2009, Congress set aside $300 Million in tax credits for expenditures associated with investment in plug-in electric vehicles in order to support and promote that technology. *See* NORTH AMERICAN ELECTRIC RELIABILITY CORPORATION, RELIABILITY IMPACTS OF CLIMATE CHANGE INITIATIVES: TECHNOLOGY ASSESSMENT AND SCENARIO DEVELOPMENT 77 (2010), *available at* http://www.nerc.com/files/RICCI_2010.pdf.

effective. Implementation of smart grid should examine the use of less expensive and simpler technologies such as smart meters and automated thermostats to avoid causing consumer distrust through excessive expense. Weiser felt the technology should put a premium on being easily adaptable to meet the needs of a wide range of consumers.

Weiser concluded with four "animating" questions, including: (1) how do consumers view smart grid technologies and how can the smart grid revolution be more consumer friendly; (2) how to place a structure around ongoing experimentation, evaluation of different successes or failures, and assess what really constitutes a cost effective grid; (3) how to define the role for federal government policy moving forward, including the structure for research and development funding and ongoing evaluation; and (4) evaluating what areas can benefit from defining best practices, for example privacy standards for the data involved in smart grid operations, as well as cyber security issues.

B. Discussion

Starting off the discussion, one participant felt that global and national issues – such as greenhouse gas emissions, economic development, and energy security – are what drive the discussion at the federal level, but that what is actually needed is a more incremental approach, and this "gradualism" should lower the cost impact on consumers. Along these lines, one participant thought that the focus for policy makers in Washington, D.C. should be where there is a pending "paradigm shift" that will provide an opportunity for significant movement in an area where otherwise, using an incremental approach, the US would not be able to reach its goals.

Another participant felt there needed to be a more holistic long-term view of the smart grid effort, and that this should be a multi-step and extended effort involving integrated resource planning processes. The goal, he said, should be getting to a smart grid "formula" and it might be that the key is in requiring utilities to give communities and states an annual business plan of sorts that addresses the electricity generation mix.

Most agreed that the federal and state governments are "in this together" but that the difficulty is in framing the challenge so as to increase engagement by all the stakeholders. Here, participants thought that education might be the key – for example through educating the consumers and then giving them the tools to automate their consumption. This way consumers would understand more about what their consumption means in terms of electricity production and distribution, and would have the ability to sync pricing (reflecting actual system conditions) with their own demand, moving the consumer from a passive to an active role. One participant thought that many other countries have more advanced regulators and regulation systems for their electricity grids, and getting the international perspective may help inform new models. As an aside, another participant thought that New Institutional

Economics (NIE) might help the US find new regulatory institutions that both support innovation and solve some of these regulatory conundrums.

There were some questions as to what the "killer app" in the smart grid space might be. One participant pointed to the telecommunications sector and how it transformed over time. While historically there had been some speculation as to what the killer app in that space might be – the one to break up the Bell telephone system – that app turned out to be the Internet and cell phones. The thought here was that for things in the smart grid space to shake out in a natural and sustainable way, like the changes in telecommunications, finding the killer app may take time. Looking at the current state of smart grid, some participants thought that things were moving pretty fast, especially from the consumer's perspective, and that utilities and policy makers should be careful and avoid assuming that all consumers will buy into the same things.

Highlighting the consumer perspective, one participant felt that California might end up as an example of what *not* to do when it comes to implementing smart grid technologies. In California, they related, there have been significant "hiccups," for example when some wrong technologies were installed and then needed to be replaced by correct technology – at a significant cost to the rate-paying customers. Along these lines, there have been numerous changes to the rates, increases in consumer education, and more, but now many cities are "refusing" the new smart meters. With all of this, and especially with only marginal savings to consumers on their bills – with at times short -term increases to achieve long-term savings – it is becoming more difficult to convince consumers that smart grid is something helpful. In order to address some of these issues it may be helpful to start talking to consumer groups at the federal level.

Privacy questions when it comes to third-party access to data

In terms of getting information to the consumers of electricity and the innovators in the smart grid space, one participant wondered if there was an opportunity to accomplish some of the consumer education at the edge of the network, and asked how the innovators on the supply side were getting their information. He pointed out that often utilities do not share the type of information necessary for third parties to innovate, and that for smart grid to work the people at the edge need access to the appropriate information. There was some discussion of groups working to provide consumers with this kind of information, and one participant anecdotally pointed to an effort by Google to gain access to this data through private agreements with utilities.[35]

[35] Currently only available to utility customers in San Diego, California, the project is called the Google Power Meter. It is an online tool that uses the "information provided by utility smart meters" to allow the consumer to view their "home's energy consumption from anywhere online." *See* Google Power Meter, http://www.google.com/powermeter/about/about.html (last visited Jan. 23, 2011).

Another participant pointed out that the incentives for utilities on this issue are all wrong, because providing access to this data would in essence make utilities the primary motivator behind reducing their own customers' demand. Because of the respective incentives, participants thought one focus might be on federally mandated data sharing. This brought up the question of data access versus data ownership. The group generally felt that consumers owned their electricity usage data and there was unanimous agreement that consumers should have access to it.[36] One participant pointed out that there might be a chicken and egg problem, because how does a regulator mandate access to certain types of data if they are not even being captured? There were many questions around what data is being captured currently, how to capture more data, whether this simply meant better meters, and who or what entities should be considered third parties.

Another major issue surrounded the privacy implications of sharing consumer energy usage data. Here, the discussion began by asking what it meant to allow access to user data. Paul Ohm, Associate Professor of Law at the University of Colorado Law School, felt that data can either be useful or private, but not both.[37] For example, if you have a powerful enough computer and a sufficient number of unique variables, then data anonymization is practically impossible, or at least easily "re-individualized." Currently, he said, most privacy laws have some sort of safe harbor for anonymized data that has been "stripped" of personally identifiable information, but with increases in computing power this anonymization is mostly illusory. There needs to be a new approach to privacy, especially when looking at information that might show such things as what appliance is used in the home at what time of day and for how long.[38]

Ohm said that one approach might be a notice and consent regime wherein the utility gives notice to the consumer as to what information it gathers and disseminates, but he pointed out that many privacy experts have given up on this model because the average consumer does not rationally understand the risks involved. Thus, he said, the government may need to become paternalistic at some point and give people the level of privacy they "deserve" and not the level to which

---

[36] One participant pointed out that there was pending California legislation on this point, which has now been passed into law. It addresses such things as third party disclosure policies, data security and protection, and liability. *See* California Senate Bill 1476, http://www.leginfo.ca.gov/pub/09-10/bill/sen/sb_1451-1500/sb_1476_bill_20100929_chaptered.html (last visited Jan. 23, 2011).

[37] For a more thorough discussion of Professor Ohm's views on privacy, see Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 701, *available at* http://uclalawreview.org/?p=1353.

[38] It is possible to get at this type of information through more detailed knowledge of electricity use combined/compared against databases of the current draws of different appliances and electrical devices. *See* Jon Beyea, *The Smart Electricity Grid and Scientific Research*, SCIENCE (May 2010), p. 979, *available at* http://www.sciencemag.org/content/328/5981/979.full.pdf.

they agree. Maybe the answer is legislation or procedures that "nudge" consumers in certain directions on privacy choices for a set number of years until there is more experience and a better culture of privacy in the populace.

There is also a motive problem, according to Ohm, because there will be a treasure trove of personal information on electricity usage in the smart grid data that companies will want to monetize, from personal information in the home to business information in the workplace.[39] This raises a host of questions. For example, who within the organization has access to the information? What happens if it is "leaked"? What procedures should be taken, or can be required to be taken, to prevent leaks? Must the owner of a building give consent to sharing the energy usage of a building or can the tenants give consent? Ohm felt that drawing bright lines on these issues would be very difficult.

Other participants highlighted how privacy issues are not specific to smart grid, and pointed out how the same issue is involved in the information tracked through customer discount cards at supermarkets and other personalized coupons. The difference though, according to other participants, is that many people do not look favorably on their utility company and there is generally no competition or choice at the consumer level; people cannot choose their utility company and they are "forced" to pay many regulatory fees. Another participant pointed out that this bridge had already been crossed in the telecommunication space by granting access to customer consumption information, and that there were processes and procedures there that could be applied to the smart grid data. Another useful model might be the protection of medical information through the Health Insurance Portability and Accountability Act (HIPAA)[40] because personal medical information is arguably more sensitive than energy usage, and is highly regulated, but is still open to use by researchers.[41] An added wrinkle, according to Ohm, is that there is significant judicial precedent that makes gathering information about what goes on inside someone's home uniquely invasive.[42]

---

[39] One participant called this "Google envy."

[40] 42 U.S.C. § 1320d (2010).

[41] For a list of the privacy rules for HIPPA compliant researchers, see the National Institute of Health website at http://privacyruleandresearch.nih.gov/faq.asp (last visited Jan. 23, 2011).

[42] The Supreme Court has recently upheld the relatively high expectations of privacy in the home by prohibiting the use of a thermal imager to gather details about the home that would have been previously inaccessible without a physical trespass—at least until such time as the technology to do so becomes widely available to the general public. *See* Kyllo v. United States, 533 U.S. 27 (2001). But there seems to be some tension with the generally lower expectations of privacy in the information contained in common business records that most utilities gather as a matter of course. *See* Jack L Lerner & Deirdre K. Mulligan, *Taking the "Long View" on the Fourth Amendment: Stored Records and the Sanctity of the Home*, STAN. TECH. L. REV. 3 (2008), *available at* http://stlr.stanford.edu/pdf/lerner-mulligan-long-view.pdf.

Some participants felt that the privacy discussion might be overcomplicating what smart grid is. They pointed out how on the Internet many people give out vast amounts of significantly personal information every day, whereas smart grid technologies are at their core only about giving and receiving electricity price and demand data. Some participants felt there would be vendors who would offer various levels of data protection along with their services and that this would end up being a market decision by consumers.

Conclusion

Implementation of smart grid technologies in the United States faces many challenges, from economic, technological and policy standpoints. As cost of service regulated entities, utilities are not currently well incentivized to increase efficiency through innovation, while state regulators face the real risk that at this stage, the smart grid is not a proven investment. Architectural concerns of whether the signaling network will be open or closed and where the intelligence will be located are critical for ensuring that the smart grid design is optimized to allow for the greatest efficiency moving forward. Finally, state and federal governments will likely need to work together in a form of cooperative federalism to ensure the optimized implementation of new smart grid technologies.

Attachment A – Smart Grid Roundtable

(Alphabetical by Last Name)

Srihna Apridee	University of Colorado, Interdisciplinary Telecommunications Program, *Student*
Rimvydas Baltaduonis	Gettysburg College, *Assistant Professor, Economics*
Frank Barnes	University of Colorado, *Professor of Electrical Engineering*
Brad Bernthal	Silicon Flatirons, CU Law, *Director of EI, Professor*
Tim Brown	University of Colorado, Interdisciplinary Telecommunications Program, Electrical and Computer Engineering, Computer Science, *Professor*
John Colgan	Illinois Commerce Commission, *Commissioner*
Kelly Crandall	University of Colorado Law School, *Alumnus*
Jeffery Earl	Indiana Regulatory Commission, *Law Judge*
Sherman Elliott	Illinois Commerce Commission, *Commissioner*
Giancarlo Estrada	Arizona, *Policy Advisor to Chairman Kris Mayes*
Manuel Flores	Illinois Commerce Commission, *Chairman*
Ray Gifford	Silicon Flatirons Center, *Senior Adjunct Fellow*
Dian Grueneich	California Public Utilities Commission, *Commissioner*
Eric Gunning	Wilkinson Barker Knauer, LLP, *Associate*
Mark Handschy	U.S. Department of Energy, *Senior Advisor*
Dale Hatfield	Silicon Flatirons Center, *Executive Director*
Steven Hauser	Grid Integration National Renewable Energy Laboratory, *Vice President*
Bill Levis	Colorado Department of Regulatory Agencies, *Consumer Counsel*
Maureen McLaughlin	U.S. Department of Energy, *Senior Legal Advisor*
Paul Ohm	University of Colorado Law School, *Associate Professor*
Preston Padden	Silicon Flatirons Center, *Senior Fellow*
Kaleb Sieh	Silicon Flatirons Center, *Research Fellow*
Gregory Sopkin	Colorado Public Utilities Commission, *Former Chairman*
Jess Totten	Competitive Markets Division Public Utility Commission of Texas, *Director*
Bob Veneck	Technical Operations Indiana Utility Regulatory Commission, *Executive Director*
Phil Weiser	National Economic Council, The White House, *Senior Advisor to the Director for Technology and Innovation*

Journal of Economic Perspectives—Volume 12, Number 3—Summer 1998—Pages 89-110

U.S. Industry Adjustment to Economic Deregulation

Clifford Winston

Economic deregulation does not happen overnight. It takes time for lawmakers and regulators to dismantle regulatory regimes, and then it takes more time for the deregulated industries to adjust to their new competitive environment. Federal regulatory agencies and the U.S. Congress began liberalizing pricing, entry, and exit in the transportation, financial, energy, and communications industries during the mid-1970s. But while some smaller industries such as intercity bus transportation and air cargo have been fully deregulated during that time, the only major American industries fully deregulated to date are airlines and motor carriers. As to the pace of industry adjustment, consider airlines. Economic deregulation of the airline industry began in 1974 when the Civil Aeronautics Board first encouraged experiments with discount fares. It was completed in 1983 when all regulations on fares, entry, and exit were eliminated. Yet more than 20 years after the deregulatory process began, the airline industry continues to shed inefficiencies that accumulated over decades of regulation and to find new ways to market their service. Indeed, the airlines' latest marketing innovation—selling discount fares on the Internet—would be illegal if fares were still regulated.

It is not surprising that deregulated (or partially deregulated) industries are slow to achieve maximum efficiency. When regulatory restrictions on pricing, operations, and entry (especially from new firms), have been enforced for decades, managers and employees of regulated firms settle into patterns of inefficient production and missed opportunities for technological advance and entry into new markets. Deregulation frees an industry from the state's control over prices, entry, and exit, although, of course, firms are still subject to antitrust laws and safety regulations. After deregulation, some costs usually fall in short order, but it takes

■ *Clifford Winston is Senior Fellow, Brookings Institution, Washington, D.C.*

firms a long time to tear down decades-old barriers to efficiency and to adopt more efficient production and marketing practices.

Deregulation is a long-term process, from which society will continue to reap benefits as firms continue to adjust to free market competition and as more and more industries are more fully deregulated. Policymakers, however, take a more myopic view—and often place a greater weight on minimizing disruptions to the public than on maximizing economic efficiency.[1] This emphasis may play a useful role in gaining passage of deregulatory legislation. Nonetheless, the concern over disruptions can also lead to counterproductive outcomes. For example, the Essential Air Services program was established in 1978 to assuage congressional fears that airline service to small communities would disappear under deregulation. However, Morrison and Winston (1986) suggest that the subsidies provided under this program were not only unnecessary, but may have been responsible for reducing the potential benefits that could be generated in these markets.

Policymakers are impatient, and when deregulation fails to produce sufficient immediate benefits to their constituents, they face temptation to reregulate or to limit further deregulation. For example, Congress deregulated cable television in 1984 and then reregulated it in 1992, even though the reregulation in all likelihood resulted in a loss in consumers' welfare (Crandall and Furchtgott-Roth, 1996). Similarly, some policymakers were expressing concern early in 1997 that the 1996 Telecommunications Act had not already led to declines in rates—while failing to acknowledge that this legislation still left the telecommunications industry substantially regulated.

Policymakers and the public tend to notice only what deregulation has done for them *lately*, forgetting that society will continue to reap benefits from deregulation as industries continue to adjust to the new competitive order. As a result, they often undervalue the benefits of deregulation. To make sure that policymakers and the public do not sell deregulation short, it is useful to examine how the long-run efficiency benefits of deregulation are achieved. In doing so, it will become clear that the U.S. industries' adjustment to deregulation is marked by many common elements. These same elements will probably govern the evolution of other industries—telecommunications, cable television, and electricity—that are now embarking on substantial deregulation. It appears likely that each will become far more efficient than it was under regulation and in the process generate substantial benefits to consumers.

Although I believe that the academic community has reached a consensus that deregulation's net benefits to consumers are substantial, I recognize that this view may not be compelling to certain policymakers who do not care about the average consumer in the country, but only about their constituents. For example, at a recent hearing on the airline industry in the U.S. House of Representatives, Congressman

[1] Hahn and Hird (1991), Winston (1993), and Joskow and Noll (1994) provide comprehensive overviews of regulatory reform. Peltzman (1989), Noll (1989, 1996), and Winston and Crandall (1994) assess political forces underlying regulatory reform.

Conyers from Detroit, Michigan, ignored estimates of the large nationwide benefits of airline deregulation and questioned the wisdom of the entire policy because he felt fares at Detroit's Wayne airport were "too high." Setting aside the accuracy of Conyers's assessment, his perspective illustrates that public debates about deregulation persist after it has long been in place. It is therefore important to identify the long run benefits of deregulation to respond to policy proposals that may emerge from these debates.

Theoretical and Empirical Considerations

Neoclassical economic theory predicts that both a profit-maximizing monopolist and a perfectly competitive firm will operate on the technologically efficient production frontier; that is, any profit-maximizing firm should always wish to minimize its costs, regardless of how much competition it faces. But a regulated firm is in a different situation. Although regulated firms can choose their technologies and operating practices, these choices are made subject to the state's control over prices, entry, and exit and without the challenges posed by unrestricted competition from incumbent firms and new entrants, hence managers and employees face a rather different set of incentives in searching for greater efficiency. I am not aware of a formal model of industry adjustment to deregulation that provides theoretical guidance on this issue. Predictions of the effects of deregulation were generally guided by static models that assumed technology and operations would not be significantly affected by the change in the regulatory regime (Winston, 1993). Nonetheless, one can draw on some basic theoretical ideas to identify the central factors that will cause an industry to become more efficient as it adjusts to deregulation.

First, an old-fashioned view is that regulation improves welfare because it helps control monopoly. This view has been discredited because it is clear that regulation primarily limits competition among firms, and this lack of competition causes an industry to accumulate substantial managerial slack or "X-inefficiency"; that is, firms do not minimize the cost of producing a given level of output. When an industry is deregulated, unrestricted competition among incumbent firms and from new entrants forces the industry to shed such inefficiencies and to seek out innovations in marketing, operations, and technology. Nickell (1996) presents empirical evidence that shows that increased numbers of competitors are associated with a significantly higher rate of total factor productivity growth.

Second, particular regulations can force firms to operate in an inefficient manner. In a seminal paper, Averch and Johnson (1962) showed that a hypothetical form of regulation, rate-of-return regulation, forced regulated firms to choose their inputs in an inefficient manner; because the regulators determined a firm's profits as a percentage of the firm's capital investment (or "rate base"), firms had a strong incentive to overinvest in capital. One could modify this model to characterize

situations where an actual regulation leads to an operating inefficiency. For example, entry barriers prevented firms, such as airlines and motor carriers, from developing their networks optimally; exit barriers prevented firms, such as railroads, from shedding excess capacity; and price regulations prevented firms, such as natural gas pipelines, from efficiently marketing their capacity during peak and off-peak periods.

Finally, regulations also prevent firms from responding effectively to external disturbances, such as a recession or a large unanticipated change in prices or interest rates. An industry subject to regulation may, because of its cartel status, be somewhat insulated from these shocks. For example, if fuel prices rise, regulated electric utility firms could appeal to regulators to raise prices. But if regulators refuse such appeals, regulated firms lack the flexibility to respond to shocks (Joskow, 1974). Deregulated firms gain the ability to respond more effectively to external disturbances.

To sum up: an industry's adjustment to deregulation will, in theory, be shaped by intensified competition and increased operating freedoms that will cause the industry to become more technologically advanced, to adopt more efficient operating and marketing practices, and to respond more effectively to external shocks. Moreover, one must remember that deregulated industries are facing these challenges and opportunities for the first time, often with a burden or legacy of investments, operating practices, and sunk costs that inhibit change and competition. In what follows, I will present empirical evidence from the airline, trucking, railroad, banking, and natural gas industries that illustrates how these factors govern an industry's adjustment to deregulation. Regulatory reform of the former industries began in the mid-1970s, and their current regulatory status is summarized in Table 1. I focus on these industries because they are major industries that have been subject to regulation for decades and have recently experienced substantial regulatory reform. Thus I do not consider the intercity bus industry because it is comparatively very small (its annual revenues are only $900 million), nor do I consider petroleum production, because it was regulated for a short period. The experience drawn from the industries in Table 1 will constitute evidence that I will use to assess the likely adjustment of other industries to deregulation.

Some of the evidence that I report should be qualified at the outset. The appropriate way to measure the effects of deregulating an industry is a counterfactual analysis that estimates the price, cost, and service changes that are solely attributable to deregulation, and thus would not have occurred had an industry still been regulated. But such an analysis can be complicated. It may need to account for the business cycle, for those elements of technological change due to deregulation and those not, and for changes in the characteristics of products, including service quality. Some argue that an appropriate counterfactual should also allow for regulators to learn from past errors and for regulation to have improved as it hypothetically continued, although there is little evidence that regulators learn and change in this way. To sidestep such complications, a simpler approach is to evaluate changes in real costs and prices over time; for example, one can compare real

Table 1
Current Status of Regulatory Reforms

Industry	*Regulatory Status*
Airlines	Pricing, entry, and exit have been fully deregulated.
Motor Carriers	Interstate and intrastate rates, entry, and exit have been fully deregulated.
Railroads	Most rates, including contract rates, have been deregulated, but "tariff" rates for certain commodities are still subject to maximum rate "guidelines."
Banking	Ceilings on interest rates (except for demand deposits) have been eliminated by deregulation, and Congress has authorized interstate banking to take place during the next few years. Thrift institutions have been allowed into consumer and business lending, while affiliates of banks have been permitted some degree of securities underwriting.
Natural Gas	Prices at the wellhead have been fully deregulated. Independent shippers now have access to pipelines at rates regulated by the Federal Energy Regulatory Commission. By obtaining "interruptible" service (i.e., a pipeline owner can stop service to a customer when demand is high under conditions specified by a contract), shippers can obtain discounts from these rates. Shippers can also sell their surplus pipeline capacity to other entities. This competitive re-selling has also led to rates below the tariff rates. Finally, market-based or negotiated rates for pipeline storage service, hub service, and even transportation services are being allowed.

airline costs in 1977 the first year before deregulation with real airline costs for the most recent period during which cost data have been collected. This simpler approach is taken in many studies of deregulation, thus I will draw on these estimates here. Although these analyses do not constitute a rigorous counterfactual, their conclusions are strongly suggestive and qualitatively consistent with the few rigorous counterfactuals that have been performed, typically using less recent data.

Changes in Entry and Exit and the Extent of Competition

As the airline, trucking, railroad, banking, and natural gas industries have been deregulated, competition has intensified, both among incumbent firms and because of new entrants.

Many of the new entrants that have made their presence felt have come to be identified as "low-cost" or "independent."[2] In the airline industry, Southwest Airlines, the premier low-cost carrier, is generally acknowledged to have set the standard for

[2] For more detailed discussion of the material in this section, see Morrison and Winston (1995) for the airline industry; Corsi (1996a) for the "less-than-truckload" trucking industry; Corsi (1996b) for the "truckload" trucking industry; Grimm and Windle (forthcoming) for the railroad industry; Berger, Kashyap, and Scalise (1995) for the banking industry; and DeVany and Walls (1994), Herbert and Kreil (1996), and Energy Information Administration (1995) for the natural gas industry.

efficiency and consistent service that other carriers are striving to reach. ValuJet, another low-cost carrier, also played an important competitive role until its license was suspended in June 1996 because of safety concerns (it has subsequently returned and is now known as AirTran); its presence in Atlanta influenced Delta to lower its costs and prices. The trucking industry is composed of two sectors: "less-than-truckload" (LTL), which uses a network of terminals to consolidate shipments of more than one shipper's goods on a truck, and "truckload" trucking, which provides point-to-point service for one shipper's goods that fill an entire truck. In the LTL segment, low-cost non-union regional carriers, such as Con-Way Transportation Services, have become an important competitive force. In response to this challenge, the national LTL carriers, such as Roadway, have pursued an aggressive policy of purchasing regional (non-union) carriers and operating them as independent business units. The dominant force in "truckload" trucking has become the "Advanced Truckload" or "High Service" mega-carrier, such as Schneider National. These carriers have become so efficient that they are capturing substantial traffic from firms that historically, in response to high regulated truck rates and poor service, provided their own trucking services. For-hire trucking operations are now roughly 25 percent less expensive than private carriage, which relies primarily on (more costly) unionized labor, according to estimates from DRI/McGraw-Hill (cited in Corsi, 1996b). In the rail industry, no large carrier has entered for several decades, because of the huge capital requirements to start a railroad and the ubiquitous coverage of existing networks. But since deregulation there has been a substantial increase in the number of smaller low-cost (non-union) railroads such as Montana Rail Link, which have formed a small system from track that they purchased from large railroads. And competition among incumbent railroads is much more intense. Darius Gaskins has frequently pointed out that when he was a Berkeley economics professor, he taught theories that predicted that duopolists had considerable ability to maintain high prices. But as the CEO of Burlington Northern railroad, he learned that duopolists' prices often reflected fierce competition. In banking, new competition has arisen from interstate bank expansions and from non-bank intermediaries like mutual funds, insurance and financial companies, and corporations issuing debt. Finally, in the natural gas industry, independent producers have increased their share of gas production since deregulation, and independent shippers now have competitive access to pipelines.

Substantial merger activity has generally occurred within a decade of an industry's deregulation. As in any industry, whether recently deregulated or not, the contribution of these mergers to social welfare has been subject to controversy. For example, end-to-end mergers have helped railroads become more efficient, but parallel mergers may have reduced competition between railroads while producing few efficiency gains. Similarly, some airline mergers have been found to enhance social welfare, while others have not.[3] Major mergers often continue as firms adjust

[3] Winston, Corsi, Grimm, and Evans (1990) discuss railroad mergers in the deregulated environment, while Morrison and Winston (1989) and Werden, Joskow, and Johnson (1991) estimate the effects of airline mergers since deregulation.

to deregulation. In the airline industry, American and U.S. Airways, Northwest and Continental, and United and Delta have all recently proposed alliances. In the railroad industry, the imminent absorption of Conrail by Norfolk Southern and CSX may lead to other merger proposals. In the trucking industry, national carriers continue to acquire regional carriers, and in the wake of the UPS strike, Federal Express has acquired RPS to compete more effectively for small freight shipments. The growth of nationwide banking has already led to a great deal of merger activity in the banking industry and as indicated by the recent Norwest and Wells Fargo mega-merger, seems likely to lead to more.

In some cases, incumbent firms, particularly in the airline, banking, and railroad industries, have seen mergers as the way to enter new markets. Conversely, because regulation enabled many inefficient firms to stay in business, it is not surprising that mergers have also facilitated large-scale exit, especially in the LTL trucking industry and in the banking industry, as weaker firms that were unable to compete effectively in a deregulated environment sought a merger partner to remain in business.

Following deregulation, the net result of entry, exit, and mergers has generally been that competition in actual *markets* becomes more intense, although the total number of firms in an industry nationwide may either rise or fall. That is, travelers can choose among more airlines when taking a trip; consumers can choose among more lending institutions when seeking a loan; shippers can choose among UPS, Federal Express, or revitalized rail, truck, or intermodal (truck-rail) carriers to ship their goods; and companies can choose among more utilities or an array of marketers to purchase their natural gas.

After deregulation, competition has even developed in markets once thought to have "natural" entry barriers. In other words, experience rather than theory is more instructive about which parts of an industry are competitive and which parts are subject to natural monopoly. Some rail shippers that may appear to be captive to one major railroad have the option of using railroads that have formed a small system or using another major railroad with trackage rights over the "monopoly" carrier's track. In addition, firms contemplating a new plant location or expansion can sign long-term railroad rate contracts before committing to new investment, thus exploiting competitive alternatives before making a decision. Thousands of natural gas end users now contract with pipelines to ship and store their gas and they can release (sell) their unused capacity to other entities. Natural gas end users and rail shippers can benefit from source competition—for example, Alabama utilities that face rate increases on coal shipped from Colorado can switch to coal from Kentucky if they can access a different railroad—and from product competition—for example, utilities that face rate increases on natural gas can shift to coal or oil if their technology permits such substitution.

Innovations in Marketing, Operations, and Technology

The intensified competition resulting from deregulation causes firms to make innovations in marketing, operations, technology, and governance that enable

them to become more efficient, improve their service quality, introduce new products and services, and become more responsive to consumers' preferences.[4] One might suspect that regulation could foster innovations because firms were more profitable in this environment and thus had the resources to devote to innovative activity. I point out, however, that regulation generally did not significantly increase industry profitability.

One way that deregulated firms have marketed their services more effectively is by offering an array of price-service packages to potential customers. Airlines now offer travelers a wide range of fares from discount fares with various travel restrictions—some of the lowest discount fares are exclusively available on the Internet—to much higher fares with no travel restrictions. This has contributed to planes flying with a higher percentage of seats filled—a rise in the so-called "load factor." Railroads and motor carriers now often negotiate price-service packages with shippers directly. During regulation, railroad traffic was effectively prevented from moving under contract rates. Today, more than half of all rail traffic moves at a negotiated contract rate, which has facilitated more efficient use of capacity and services that are tailored to shippers' production and inventory policies. Banks have greatly expanded the scope of financial services they offer customers and provided new tools, such as derivative securities and financial guarantees, to help their customers diversify risk. In the natural gas industry, the development of market centers and hubs has enabled marketers to help business consumers contract for rates and service that match their specific requirements.

Deregulated firms have made a number of innovations in their operations and technology. Among these innovations to date are the airlines' accelerated development of hub-and-spoke networks that have enabled carriers to increase their load factors. Also, the hub-and-spoke system has brought much greater service frequency, which was a dramatic surprise to many. When fares were regulated, airlines competed excessively in flight frequency. It was thus expected that flight frequency would fall when airlines were deregulated. But with the connections made possible by hub-and-spoke networks, one additional aircraft departure from a spoke airport to a hub airport provides many alternatives on connecting flights, so that service frequency has actually risen. Airlines' development of computer reservation systems has also facilitated improvements in scheduling and handling flight reservations. In the trucking and railroad industries, improvements in network design, the expanded use of intermodal operations, double stack rail cars, and increased use of computer systems have enabled faster, more reliable service. For example, Norfolk Southern once tracked its cars and locomotives by posting a video camera at the entrance to each rail yard; now, an electronic scanner automatically tracks each car's arrival. Truck drivers for Schneider National used to have scheduled check-in

[4] For more detailed discussion of the material in this section, see Morrison and Winston (1995) for the airline industry; Corsi (1996a, 1996b) for the trucking industry; Winston, Corsi, Grimm, and Evans (1990) for the railroad industry; Berger, Kashyap, and Scalise (1995) for the banking industry; and Henning, Tucker, and Liu (1995) and Herbert (1996) for the natural gas industry.

phone calls to headquarters; now, they roam the highways with small satellite dishes attached to their cabs, and use the data transmitted by technicians in the company's headquarters to track maintenance needs and the location of shipments. Banks have applied advances in information processing and communications to lower consumers' and their own transactions costs. Natural gas companies have used computer technology to lower their operating and maintenance expenses, and have improved their service to customers by unbundling production, transportation, and distribution.

It could be argued that many of the innovations by deregulated firms have been the direct result of advances in information technology, and thus would have occurred regardless of whether their industries were deregulated. The argument, of course, is not that deregulation was a primary force in the information technology revolution. But the benefits from these advances were realized only because firms had the incentive and operating freedom to design new production systems and write computer programs to optimize operations. Under regulation, these firms had little incentive or competitive pressure to do so, and regulators did not design regulations to stimulate innovative activity.

Deregulated firms have also been spurred by external shocks to improve the efficiency of their operations, and in particular to match their capacity with consumer demand. The airline industry, for example, has traditionally made capacity commitments roughly two years in advance because of the lead times to acquire aircraft. However, the high income elasticity of demand for air travel—together with a dose of unpredictability—created overcapacity in the early 1990s, which led to intense fare wars and huge industry losses (Morrison and Winston, 1996). Carriers have subsequently adjusted their operations by expanding their capacity more slowly in the face of growing demand. In addition, American and Delta Airlines have made long-term purchasing commitments to Boeing Aircraft, in return for expedited delivery of planes, which puts American and Delta in a better position to tailor their capacity to demand.[5] Railroads have used contracts to align their cars and equipment with shippers' demand and reduced their vulnerability to problems caused by overcapacity. Contracting for natural gas pipeline storage and capacity, and the development of spot and futures markets have helped the gas industry respond effectively to the business cycle.

The experience of the banking industry with external shocks differs from that of other deregulated industries, because deregulation was followed by a crisis for the savings and loan industry and for many banks, culminating in substantial costs for taxpayers to bail out the savings and loan (S&L) deposit insurance fund. But deregulation—which in this industry meant freeing up the interest rates that savings and loans could pay, and allowing interstate banking and new financial instruments—can be held responsible for only a fraction of the savings and loan

[5] Boeing has agreed not to enforce these commitments as a condition for obtaining the European Commission's approval of its merger with McDonnell-Douglas. American and Delta, however, are unlikely to change their commitments.

debacle. The earlier regulations, under which S&L's were limited in the interest they could pay and the loans they could make, could have literally meant the annihilation of the S&L industry when market interest rates climbed sharply in the late 1970s. The S&L disaster was caused by a sequence of events: the legacy of previous regulation, unfavorable interest rate movements, partial deregulation, lax regulatory supervision, moral hazard problems created by federal deposit insurance, and unwarranted optimism over real estate prices in certain areas. The lesson here is not to fear deregulation itself; indeed, the deregulated banking and S&L industries are today vibrant and healthy. However, deregulation can be undermined if it is not supplemented with appropriate policies. In the case of a deregulated banking industry, a vigilant safety-and-soundness regulatory regime was essential, given federal deposit insurance. In other industries, the appropriate supplemental policies mean that deregulated industries continue to be subject to antitrust laws, safety regulations, and so on.

Finally, deregulation has also led to improvements in corporate governance that contribute to innovative activity. Winston, Corsi, Grimm, and Evans (1990) report that under deregulation, railroad managers are younger, better educated, and have fewer years of company service than under regulation, and Kole and Lehn (1996) and Meyer and Oster (1984) argue that airline managers have become much more entrepreneurial in the deregulated environment.

Improvements in Industry Efficiency and Consumer Welfare

By stimulating competition and giving firms greater incentives and freedom to seek out efficiency, deregulation has generated substantial improvements in consumer welfare that will continue over time. Because firms will continue to innovate in ways that they would not have under regulation, it is important to bear in mind that any current summary of the gains from deregulation is likely to be a lower bound.

Industry Efficiency

The evidence to date suggests that since deregulation, each industry has substantially improved its productivity and reduced its real operating costs from 25 percent to 75 percent. Table 2 gives some examples of what has happened in the main industries considered here. Of course, deregulation should not be given total credit for all of the decline in each industry's real operating costs. For example, some of the decline in rail operating costs can be attributed to the long run trend in railroads' traffic mix to include a greater proportion of low-cost bulk traffic. Some of the decline in truckload carriers' operating costs can be attributed to these carriers largely eliminating the more costly less-than-truckload traffic from their traffic mix. Nonetheless, deregulation does deserve credit for most of the fall in costs. In general, these cost savings have not significantly increased industry profitability. Rather, the intensity of competition under deregulation has forced much of the

Table 2
Improvements in Industry Efficiency

Industry	*Improvements*
Airlines	Average industry load factors have increased from roughly 52 percent the decade preceding deregulation to roughly 62 percent since deregulation. Real costs per revenue ton-mile have declined at least 25 percent since deregulation. Industry profits have been very volatile during deregulation, although higher, on average, than they would have been under regulation.
Less-Than-Truckload Trucking	Carriers have substantially reduced their empty miles since deregulation. Real operating costs per vehicle mile have fallen 35 percent, but operating profits are slightly lower than they would have been under regulation.
Truckload Trucking	Carriers have substantially reduced their empty miles since deregulation. Real operating costs per vehicle mile have fallen at least 75 percent, but operating profits are slightly lower than they would have been under regulation.
Railroads	Railroads have abandoned one third of their track miles since deregulation. Real operating costs per ton-mile have fallen 60 percent, and rail profits are much higher than they would have been under regulation.
Banking	The real cost of an electronic deposit has fallen 80 percent since deregulation. Operating costs have declined 8 percent in the long run because of branch deregulation. Recent industry returns on equity exceed those just before deregulation.
Natural Gas	Pipeline capacity has been much more efficiently utilized during peak and off-peak periods since deregulation. Real operating and maintenance expenses in transmission and distribution have fallen roughly 35 percent.

Sources: For airlines, Morrison and Winston (1998); for less-than-truckload trucking, Corsi (1996a); for truckload trucking, Corsi (1996b); for railroads, Winston, Corsi, Grimm, and Evans (1990) and Association of American Railroads (1996); for banking, Berger, Kashyap, and Scalise (1995) and Jayaratne and Strahan (1998); and for natural gas, Henning, Tucker, and Liu (1995) and Herbert (1996).

savings to be passed on to consumers in lower prices. The one industry where profitability *has* increased substantially is the railroad industry, which had earned low profits under regulation and experienced many bankruptcies. As hoped for by the architects of the 1980 Staggers Act deregulating the railroads, rail has profited greatly from deregulation. During 1971–80, the industry's return on equity was less than 3 percent, which along with widespread bankruptcies in the northeast and midwest heightened concerns about whether this industry could be financially viable in the private sector; during the mid-1990s, the industry return on equity has been more than 8 percent.

The intertwined pattern created by deregulation—more competition, restructuring, opportunities for new markets, and further cost reductions—seems likely to continue. The deregulation of international airline markets, which is slowly proceeding through bilateral negotiations between the United States and other countries, will bring new competition to U.S. carriers, along with new possibilities for route networks and marketing. Airlines are currently exploring a number of options to raise labor productivity such as having (lower cost) pilots who work for their

affiliates fly smaller (regional) jets that have been recently introduced on short-haul routes. The trucking industry has experienced a serious shortage of long-distance drivers, but the increasing use of intermodal operations has helped to alleviate this problem. Of course, intermodal operations also bring more competition in freight markets. Although truckers have significantly reduced their empty mileage, more progress can be made through further load consolidations, and by for-hire motor carriers making further inroads into private trucking operations (Corsi, 1996b). Although railroads have abandoned a lot of track already, they will continue that process. Railroads can also improve their fleet by taking advantage of technological innovations that have made locomotives more powerful and enabled rail cars to carry heavier loads without wearing out track. The spread of smaller low-cost rail carriers should spur the industry to continue to reduce its labor costs. Finally, it is possible that rail mergers will continue to the point where only two large transcontinental railroads, both highly efficient, would remain in the industry. But even this restructuring (if it comes to pass) would still leave two large railroads in the east and two in the west, and thus have little effect on the current level of rail competition between any two given destinations. The growth of nationwide banks and securitization will lead to more consolidations in the banking industry, but also to more competition in local banking markets. The recent development of electronic banking is opening up a new set of markets where banks are likely to compete with computer and phone companies, according to the American Bankers Association. Finally, continued unbundling of natural gas production, transportation, and distribution will increase competition in this industry, particularly at the distribution level.

Innovations in marketing, operations, and technology are often difficult to anticipate, so the path of any particular industry as it adjusts to deregulation will be unpredictable. Nonetheless, it is striking that nearly 20 years after deregulation began, even industries with low sunk costs and a simple technology, like truckload trucking, are continuing to adjust to deregulation in much the same way as industries with large sunk costs and a more complicated technology, like railroads.

Consumer Welfare

Consumers have turned out to be the primary beneficiaries of deregulation. The evidence to date suggests that since deregulation, each industry has significantly improved its service quality and reduced its real average prices from 30 to 75 percent, as summarized in Table 3. As in the case of the decline in operating costs, one can attribute most but not all of the decline in prices to deregulation. Some alternate reasons for lower costs in rail and trucking, which also apply to prices, were given earlier. Certain critics of airline deregulation have argued that fares were already declining before deregulation and that the decline in real fares since deregulation would have occurred anyway. However, Morrison and Winston (1995) show that deregulation has accelerated the de-

Table 3
Improvements in Consumer Welfare

Industry	*Improvements*
Airlines	Average fares are roughly 33 percent lower in real terms since deregulation, and service frequency has improved significantly.
Less-Than-Truckload Trucking	Average rates per vehicle mile have declined at least 35 percent in real terms since deregulation, and service times have improved significantly.
Truckload Trucking	Average rates per vehicle mile have declined at least 75 percent in real terms since deregulation, and, because of the emergence of "Advanced Truckload" carriers, service times have also improved significantly.
Railroads	Average rates per ton mile have declined more than 50 percent in real terms since deregulation, average transit time has fallen at least 20 percent, and the standard deviation of transit time has fallen even more than 20 percent.
Banking	Consumers have benefitted from higher interest rates, from better opportunities to manage risk, and from more banking offices and automated teller machines.
Natural Gas	Average prices for residential customers have declined at least 30 percent in real terms since deregulation, and average prices for commercial and industrial customers have declined even more than 30 percent. In addition, service has been more reliable as shortages have been almost completely eliminated.

Sources: For airlines, Morrison and Winston (1998); for less-than-truckload trucking, Corsi (1996a); for truckload trucking, Corsi (1996b); for railroads, Winston, Corsi, Grimm, and Evans (1990) and Association of American Railroads (1996); for banking, Berger, Kashyap, and Scalise (1995) and Jayaratne and Strahan (1998); and for natural gas, Costello and Duann (1996) and Crandall and Ellig (1997).

cline in real fares, and that the decline would have been much less had fares still been regulated.

It is difficult to itemize the sources of the price declines in each industry according to changes in competition supplied by incumbent firms and new entrants and by improvements in efficiency. Morrison and Winston (1998) are able to provide such a decomposition for the airline industry. They report that competition supplied by incumbent carriers accounts for 18 percent of the savings from lower real fares since deregulation, competition supplied by Southwest Airlines accounts for 31 percent of the savings, competition supplied by other new entrants accounts for 10 percent of the savings, and improvements in carriers' operating efficiencies account for 41 percent of the savings.

Deregulation's considerable success has surprised the economics profession in at least some respects. The profession correctly predicted that consumers would gain from lower airline, motor carrier, and natural gas prices, and from higher interest rates on bank deposits, but the profession expected rail rates to rise (or at least not to fall very much), and failed to predict the large gains from improved service quality in all industries (Winston, 1993). Accounting for changes in prices and service quality, a conservative estimate of the annual net benefits that consum-

ers have received just from deregulation of intercity transportation—airlines, railroads, and motor carriers—amounts to roughly $50 billion in 1996 dollars (Morrison and Winston, 1998). These benefits have been achieved without compromising safety.[6]

To be sure, consumers have not shared equally in the gains from deregulation. Long-distance air travelers on high density routes have benefitted more than short-distance travelers on low density routes; large rail and truck shippers have benefitted more than small rail and truck shippers; large savers have benefitted more than small savers; and commercial and industrial natural gas customers have benefitted more than residential natural gas customers. But this distribution of benefits has generally had a rational economic basis. Because one objective of economic regulation is commonly to "equalize" prices and services across all consumer groups and geographical areas, it is not surprising that the transition from regulation to market forces has not "equalized" deregulated prices and services for all customers. In any industry, buyers who make large purchases or reside in markets with many competitors or lower costs will generally receive lower prices and better service than buyers who make small purchases or reside in markets with few competitors or higher costs. In the airline industry, for example, Morrison and Winston (1997) show that roughly 90 percent of the difference in the gains to travelers on high traffic density routes and low traffic density routes can be explained by the higher costs of serving travelers on low density routes, which are attributable to the lesser economies of scale of smaller planes and lower load factors on such routes.

As in other industries, institutions are evolving to maintain competitive forces in many deregulated markets. In shipping, third-party logistics firms have sprung up to analyze shipper distribution patterns and logistics costs and to determine, with the aid of sophisticated software, the lowest-cost routes and the carriers with the lowest rates. These firms are also able to achieve cost savings for shippers by leveraging the volumes of all of their clients to obtain discounts from carriers. Travel agents and large firms can obtain lower fares for travelers who can be organized into a bargaining unit. Recently, GM and Chrysler set up an agreement with Pro Air, a start-up carrier, to pay a monthly fee for five years in exchange for unlimited passage on the carrier for their employees. Marketers can obtain lower natural gas prices for an aggregate of residential users, and so on. Many of these institutions are just developing. In the future, they are likely to become more prominent in deregulated markets and enable more consumers to share in the ongoing efficiency gains from deregulation.

[6] The effects of deregulation on wages and employment vary by industry. For example, truckers have experienced significant losses in wages but railroad workers have not, while railroad employment has been cut but airline employment has risen. In general, the losses to labor from deregulation represent at most a modest fraction of consumers' gains, see Winston (1993) and the accompanying article by James Peoples in this issue.

Implications for Other Industries

The evidence suggests that deregulation has been and continues to be a success, and further, that this success is the result of an adjustment process that is common to all major industries that have been completely or partially deregulated. Is this conclusion likely to apply to the major industries—telecommunications, electricity, and cable television—that are currently facing substantial regulatory reform?[7]

Deregulation in these areas is barely underway and moving very slowly. In telecommunications, prices for interstate long-distance service have been deregulated for incumbent carriers—subject to various restrictions on geographic conformity of rates and interconnection charges that are kept above costs by state and federal regulators to cross-subsidize local service. The prices of intrastate long-distance and local service are still regulated by state regulatory commissions. The 1996 Telecommunications Act envisioned widespread entry by the many players in telecommunications and cable TV into each other's markets, but also left in place a number of existing regulations, including those at the state level, and set up some new preconditions and restrictions on such competition. In electricity, states are exploring deregulation of generation, and allowing access to transmission and eventually to distribution. Cable television rates were deregulated in 1984 and then reregulated in 1992, but, in accordance with the 1996 telecommunications legislation, rates for smaller systems have now been deregulated again. Remaining cable television regulations are to be phased out by 1999. Some in the policy-making community and the press are attempting to assess the overall benefits of deregulation by looking at the short-term effects of these partial steps toward deregulation. I will take a long-run perspective, which begins by assessing whether the greater operating freedoms afforded by deregulation will be accompanied by intensified competition in these industries, and then considers both the likely problems of transition and the possible innovations that may benefit the public.

Competition and Potential New Entry

Until recently, many economists would have said that telecommunications, electricity, and cable TV industries were natural monopolies on the grounds that they have substantial economies of scale, making entry difficult. But recent technological advances have changed the rules of competition in these industries. In telecommunications and cable TV, digital transmission technologies over fiber optics or the electromagnetic spectrum are creating new opportunities for competition in parts of these industries. In electricity, the advent of small scale combined cycle gas turbine (CCGT)

[7] For more detailed discussion of the material in this section, see Crandall and Waverman (1995) and Harris and Kraft (1997) for the telecommunications industry; Crandall and Furchgott-Roth (1996) for the cable television industry; and Smith (1996), White (1996), Joskow (1997), and Kwoka (1997) for the electricity industry.

technology is doing the same. In all three industries, both incumbents and potential entrants are providing a potentially large source of competitors.

The current incumbent players in telecommunications include both local service companies, like the regional Bell operating companies, several independent companies, and a number of small rural companies; and long-distance telephone service, like AT&T, MCI, Sprint, and some smaller carriers. They are potential entrants into each other's markets. Other potential entrants into both the local and long-distance markets include cellular and other wireless companies, cable TV companies, and electric utilities, who often own fiber optic cable networks that could carry phone calls. Even Internet services now offer long-distance and international voice calls. Similarly, cable TV companies could face competition from telephone companies, electric utilities, and cellular systems, as well as from direct broadcast satellite systems and multichannel microwave distribution services.

Competition arises in electricity when services are unbundled, so that a single company no longer provides power generation, transmission, and retail distribution services. Independent power producers are potential entrants in generation; retail power merchants and generation companies are potential entrants in transmission; and municipalities could become potential entrants in distribution. An additional source of competition will arise if customers can negotiate with generation companies for power and arrange their own transmission, as has occurred in natural gas.

As in the case of other deregulated industries, the strongest competitive forces could be provided by new low-cost competitors, who are already starting to emerge. For example, "dial around" long-distance firms, which can be accessed by dialing a five-digit code, are offering lower long-distance rates than larger carriers because they have much lower marketing and promotional expenses. In electricity, pilot experiments in certain states will enable consumers to obtain lower electricity prices from independent start-up companies, which use the latest technologies and have much lower costs than older incumbent utilities. The incumbent utilities have enormous sunk costs of plant already installed and contractual commitments to buy independently generated power at very high rates. Free entry may even create markets for a generic service, such as communications or energy, which could eventually enable consumers to choose among several firms to supply all their energy and communications services. The Potomac Electric Power Company has recently announced a new venture with RCN Corporation of Princeton, New Jersey, to offer alternative local telephone service, high-speed Internet connections, and cable television to consumers in the Washington, D.C., area.

The rapid pace of merger activity in the communications and energy industries suggests that firms in these industries are positioning themselves for deregulated competition. Long-distance and local telephone companies, regional Bell operating companies, and cable and telephone companies are proposing mergers. The value of mergers between electric utilities and between electric and natural gas utilities rose from $2 billion in 1982 to nearly $20 billion in 1996. Antitrust authorities will again be required to mediate between the private interests of firms and the efficiency objectives of society. However, most of the proposed restructuring is likely

to help industries shed inefficiencies developed over decades of regulation. The probable pattern is that the number of competitors may fall over time, but the intensity of competition in communications and electricity markets should become much greater than it was under regulation.

Potential Problems of the Transition

It is clear that intense competition can develop in communications and electricity markets. But the legacy of sunk costs created by regulation presents problems for the transition from regulation to deregulation of these industries. Meyer and Tye (1988) argue that the transition should encourage suppliers to sign the same kind of contracts with each other and with customers that they would have signed had there been no sunk cost legacy of regulation, thereby causing regulatory institutions to wither away. In practice, the competing interests of incumbent firms, new entrants, and consumers—as mediated through the existing regulatory and legislative structure—is likely to inhibit prompt or fully efficient solutions. The major problems to be addressed include competitive access, stranded assets, and universal service.

The competitive access issue begins with the observation that to facilitate competition in telecommunications and electricity, competitors must have the option of interconnecting with existing phone networks and electricity transmission lines. As one can imagine, incumbent carriers and new entrants often disagree as to what charge is reasonable; the new entrants argue that they should be charged only at marginal cost, while the incumbent carriers argue for higher charges that would allow them to recover their earlier capital costs. Parties have been able to agree on access charges in the railroad, natural gas, and airline industries, and market mechanisms have also developed to promote access. In Canada, railroads have engaged in "reciprocal switching," thus enabling shippers to have access to more than one railroad. In natural gas, re-selling of capacity has facilitated access to pipelines. To facilitate entry at airports, airlines are now able to buy or lease slots and gates from each other. The Federal Communications Commission has been trying to issue legally binding guidelines, without much success, on how the regional Bell operating companies and other local exchange companies must open their markets to competition and under what conditions the regional Bell companies can enter long-distance markets. AT&T's recently proposed acquisition of TCI, a cable television company, should sharply expand its ability to provide local service and could result in entry of local phone companies into long-distance markets. Congressional legislation has been introduced that would require the federal government to phase in competition in electricity at the consumer level in any state that has not done so by the year 2000.[8]

The "stranded assets" problem arises because incumbent firms are concerned that with deregulation, they will be stuck with uneconomic capital investments

[8] In the case of communications, regulatory authorities can also enhance social welfare by continuing to reallocate spectrum capacity to facilitate competition in cellular service (Kwerel and Williams, 1992).

urged upon them by regulators in the past, while having to compete against firms that are using more efficient technologies. The appropriate compensation, if any, for these stranded assets has emerged as a controversial issue in electricity and telecommunications; Kahn (1997) provides a balanced perspective on this issue in the electricity industry. But the issue is not new. In the natural gas industry, deregulation forced pipeline owners to accept from producers contractually priced gas supplies that they were unable to sell to utilities, because the utilities could find lower-cost sources of natural gas. Pipelines were allowed to recover, through a temporary fixed charge on consumers, as much as half the cost of "buying out" existing commitments to producers. In fact, based on data furnished by pipelines, transition costs predicted by the Federal Energy Regulatory Commission to be $44 billion turned out to be only $13 billion (Crandall and Ellig, 1997). After the break-up of AT&T, the depreciation rates for the obsolete capital caused by the break-up were generally allowed to increase (Costello and Graniere, 1996). When competition accelerated the obsolescence of existing investments, regulators resorted to price caps, but these prices caps were not explicitly set to allow the further recovery of stranded assets. Instead, telecommunications firms were responsible for recovering the undepreciated portion of obsolete capital through the increased profitability allowed by price cap regulation. The stranded cost stakes in electricity are huge, because incumbent utilities have been prodded by regulators to make suboptimal investments and sign long-term contracts that involve costly technologies, like nuclear power and some alternative energy sources, so that their costs will be much higher than new entrants' costs. But several states have already proposed ways to help utilities pay off their construction debts, usually through some fixed charge on consumers' monthly bills, thus enabling deregulation to move forward. Indeed, regulators in California have approved a plan to allow all utility customers to buy electricity from any supplier of their choosing in 1998.

The provision of universal service is not only important to policymakers, but to many supporters of deregulation as well. But if the goal is to provide reasonable prices and service to an area on an ongoing basis, firms need to have incentives to do that, and universal service requirements tend to reduce those incentives. For example, in airline and railroad deregulation, efforts were made to "protect" air travelers in rural areas by subsidizing carriers to serve these small communities and to "protect" shippers of certain bulk commodities like coal by instituting maximum rate guidelines. In fact, the airline subsidy program reduced carriers' incentives to provide high-quality, low-cost service to small communities (Morrison and Winston, 1986), while maximum rail rate guidelines have had little effect on prices because only a handful of rates have ever been successfully challenged. Another way of providing universal service is to require that large companies cross-subsidize certain users. Such cross-subsidies are difficult to sustain in a competitive market, where services can often be unbundled. Moreover, the gains to the community may not be as large as expected. In telecommunications, for example, Crandall (1997) finds that the inefficient cross-subsidization of local service through supra-compet-

itive long-distance rates has provided very small benefits to low-income consumers, because many of these consumers are heavy users of long-distance service. In these ways, policies to preserve universal service have a tendency to undermine the competitive forces that can enable consumers to obtain reasonable prices and service.

Opportunities for Innovations

Deregulation should not receive credit for the technological changes that have opened the possibility of intense competition in the communications and energy industries. But deregulation should allow and encourage more rapid application of these innovations, and by providing greater operating freedom and a more competitive environment it should stimulate new innovations. For example, communications regulation has cost consumers by delaying the introduction of new services, such as voice messaging and cellular telephone (Hausman, 1997). But after just partial deregulation, the demand of communications firms for fiber-optic cable to provide an array of services at lower cost has already exceeded expectations. There is also evidence that countries with several competitive telecommunications firms have much greater penetration of cellular phone service than countries with state-run monopolies, and that state regulatory reforms, such as using rate caps rather than rate-of-return regulation, have had a beneficial impact on companies' deployment of new technologies that allow much higher speed service, such as high-speed Internet connections (Blustein, 1997; Taylor, Zakardas, and Zona, 1992; Greenstein, McMaster, and Spiller, 1994). Low orbiting satellites represent another technology that could lead to lower costs, and new and improved services. Unbundling electricity production and emerging third parties could lead to various innovations; for example, consider a power marketing firm that also offers a bundled service of optimal climate control and conservation, together with the sale of power.

So what can we expect from deregulating communications and electricity? As with other deregulated industries, the adjustment process will be long and difficult, but the potential benefits to consumers appear to be huge. Remember, real operating costs in other industries fell roughly 25–75 percent following deregulation. Costs should also fall substantially in deregulated communications and electricity industries, as firms are free to shed inefficient technologies and are stimulated by new competition to develop and implement cost-reducing innovations, which should bring prices much closer to marginal costs. In telecommunications, the annual benefits from optimal pricing alone have been estimated to be as high as $30 billion (Crandall and Waverman, 1995). Deregulation of cable television is likely to lead to a more rapid expansion of viewing options (Crandall and Furchtgott-Roth, 1996). The electricity industry will come to be characterized by a competitive generation market, access to the transmission network for all transactions, and a high degree of service unbundling. Although incumbent utilities are likely to suffer losses, new entrants are likely to be profitable, and residential and commercial consumers of electricity are likely to gain considerably more than $10 billion an-

nually from price and entry deregulation (White, 1996). Some gains are already taking place as public transit companies are negotiating contracts with alternative suppliers of electricity or renegotiating contracts with current suppliers that substantially lower their electricity costs.

Conclusion

Some policymakers and economists appear reluctant to draw generalizations from the U.S. experience with deregulation over the last two decades. Industries, it is said, are different: they have different technologies, entry requirements, and so on. That deregulation works in one industry does not imply it will work in others.

This paper suggests, however, that industries are likely to behave quite similarly when it comes to adjusting to deregulation, and that their adjustment, while time-consuming, will raise consumer welfare—significantly even at first, and increasingly over time. Markets will become more competitive. Firms will develop innovations to become more efficient and more responsive to consumers. The benefits to society will grow as the adjustment continues. Indeed, the success of the U.S. deregulation experiment has gained attention throughout the world as evidenced by countries such as Japan and members of the European Union embarking on policies to liberalize competition in transportation, communications, and banking.

In the absence of any real, accumulating evidence about the impact of regulatory reform, the ideological overtones in the past debate over deregulation have been understandable. But we now have some 20 years of evidence. It is time to refocus the debate directly on that evidence and broaden our vision of where markets can replace governments in determining the most efficient way to create and sell products and services. This process is taking place in telecommunications, electricity, and cable TV. Looking ahead, government control over prices, entry, and exit in ocean transportation clearly warrants the deregulatory reforms proposed by the Clinton administration. Similar proposals are also warranted at the state and local level for taxis, limousines, and occupational licensure. It may be even more important to question the extent of the government's quasi-monopolist role in providing and regulating urban transportation (Lave, 1985; Winston and Shirley, 1998) and education (Chubb and Moe, 1990). These are big steps. But most of the theoretical arguments for greater competition in these activities parallel the arguments for deregulation. The empirical evidence may also turn out to be the same.

■ *This paper was originally presented at the Conference of Economists at Australia National University. I am grateful to W. McKibbin for commissioning the paper and to S. King and J. Quiggan for discussing it. Ongoing guidance and encouragement have been provided by A. Kahn, J. Meyer, and T. Taylor. I have also benefitted from discussions with M. Baily, A. Berger, T. Corsi, K. Costello, R. Crandall, S. Dennis, C. Grimm, R. Hahn, J. Herbert, E. Kwerel, J. Kwoka, R. Litan, S. Morrison, P. Nivola, J. Ratner, P. Viton, and M. White. Finally, B. De Long, A. Krueger, and B. Szittya improved the presentation of the paper.*

References

Association of American Railroads, "Analysis of Class I Railroads, 1980–1995," Association of American Railroads, Washington DC, 1996.

Averch, Harvey, and Leland Johnson, "Behavior of the Firm under Regulatory Constraint," *American Economic Review,* December 1962, *52,* 1052–69.

Berger, Allen N., Anil K. Kashyap, and Joseph M. Scalise, "The Transformation of the U.S. Banking Industry: What a Long, Strange Trip It's Been," *Brookings Papers on Economic Activity,* 1995, *2,* 55–218.

Blustein, Paul, "Nations that Play Tele-Monopoly Do Not Pass Go," *Washington Post,* February 27, 1997, p. E1.

Chubb, John E., and Terry M. Moe, *Politics, Markets & America's Schools.* Brookings, Washington DC, 1990.

Corsi, Thomas M., "Current and Alternative Federal Size and Weight Policies: Less-Than-Truckload Motor Carriers," College of Business and Management working paper, University of Maryland, 1996a.

Corsi, Thomas M., "Current and Alternative Federal Size and Weight Policies: Truckload Motor Carriers," College of Business and Management working paper, University of Maryland, 1996b.

Costello, Kenneth W., and Daniel J. Duann, "Turning Up the Heat in the Natural Gas Industry," *Regulation,* 1996, *1,* 52–59.

Costello, Kenneth W., and Robert J. Graniere, "Lessons Learned from Restructuring of Former Comprehensively Regulated Industries," The National Regulatory Research Institute, Ohio State University, July 1996.

Crandall, Robert W., "Universal Service Subsidies and Consumer Welfare: Long-Distance Access Charges," Brookings Institution working paper, April 1997.

Crandall, Robert W., and Jerry Ellig, *Economic Deregulation and Customer Choice: Lessons for the Electricity Industry.* Center for Market Processes, George Mason University, 1997.

Crandall, Robert W., and Harold Furchtgott-Roth, *Cable TV: Regulation or Competition?* Brookings, Washington DC, 1996.

Crandall, Robert W., and Leonard Waverman, *Talk is Cheap: The Promise of Regulatory Reform in North-American Telecommunications.* Brookings, Washington DC, 1995.

DeVany, Arthur, and W. David Walls, "Natural Gas Industry Transformation, Competitive Institutions and the Role of Regulation," *Energy Policy, 22*:9, 1994, 755–63.

Energy Information Administration, "Oil and Gas Development in the United States in the Early 1990s: An Expanded Role for Independent Producers," U.S. Department of Energy, Washington DC, October 1995.

Greenstein, Shane, Susan McMaster, and Pablo T. Spiller, "The Effect of Incentive Regulation on Local Telephone Companies' Development of Digital Infrastructure," presented at the American Enterprise Institute, Washington DC, 1994.

Grimm, Curtis M., and Robert Windle, "The Rationale for Deregulation." In James Peoples, ed. *Regulatory Reform and Labor Markets.* Norwall, MA: Kluwer Press, forthcoming.

Hahn, Robert W., and John A. Hird, "The Costs and Benefits of Regulation: Review and Synthesis," *Yale Journal on Regulation,* Winter 1991, *8,* 233–78.

→ **Harris, Robert G., and C. Jeffrey Kraft,** "Meddling Through: Regulating Local Telephone Competition in the United States," *Journal of Economic Perspectives,* Fall 1997, *11*:4, 93–112.

Hausman, Jerry A., "Valuing the Effect of Regulation on New Services in Telecommunications," *Brookings Papers on Economic Activity: Microeconomics,* 1997.

Henning, Bruce, Lee Tucker, and Cindy Liu, "Productivity Improvements in the Natural Gas Distribution and Transmission Industry," *Gas Energy Review,* February 1995, *23,* 17–20.

Herbert, John H., "The Emergence of Natural Gas Market Centers," Energy Information Administration, U.S. Department of Energy, Washington DC, August 1996.

Herbert, John H., and Erik Kreil, "US Natural Gas Markets: How Efficient Are They?" *Energy Policy,* 1996, *24*:1, 1–5.

→ **Jayaratne, Jith, and Philip E. Strahan,** "Entry Restrictions, Industry Evolution, and Dynamic Efficiency: Evidence from Commercial Banking," *Journal of Law and Economics,* April 1998, *41,* 239–73.

→ **Joskow, Paul L.,** "Inflation and Environmental Concern: Structural Change in the Process of Public Utility Price Regulation," *Journal of Law and Economics,* October 1974, *17*:2, 291–327.

→ **Joskow, Paul L.,** "Restructuring, Competition and Regulatory Reform in the U.S. Electricity Sector," *Journal of Economic Perspectives,* Summer 1997, *11*:3, 119–38.

Joskow, Paul L., and Roger G. Noll, "Economic Regulation." In Martin Feldstein, ed. *American Economic Policy in the 1980s.* NBER, Chi-

cago, IL: University of Chicago Press, 1994, 367–440.

Kahn, Alfred E., "Competition and Stranded Costs Re-Revisited," *Natural Resources Journal,* Winter 1997, *37*:1, 29–42.

Kole, Stacey R., and Kenneth Lehn, "Deregulation and the Adaptation of Governance Structure: The Case of the U.S. Airline Industry," Graduate School of Management working paper, University of Rochester, July 1996.

Kwerel, Evan R., and John R. Williams, "Changing Channels: Voluntary Reallocation of UHF Television Spectrum," Federal Communications Commission, Washington DC, November 1992.

Kwoka, John, *Power Structure: Ownership, Integration, and Competition in the U.S. Electric Power Industry.* Kluwer Press: Norwall, Massachusetts, 1997.

Lave, Charles A., ed., *The Private Challenge to Public Transportation.* Pacific Studies in Public Policy, Cambridge, MA: Ballinger Press, 1985.

Meyer, John R., and Clinton V. Oster, Jr., *Deregulation and the New Airline Entrepreneurs.* Cambridge, MA: MIT Press, 1984.

Meyer, John R., and William B. Tye, "Toward Achieving Workable Competition in Industries Undergoing a Transition to Deregulation: A Contractual Equilibrium Approach," *Yale Journal on Regulation,* Summer 1988, *5,* 273–97.

Morrison, Steven A., and Clifford Winston, "Causes and Consequences of Airline Fare Wars," *Brookings Papers on Economic Activity: Microeconomics,* 1996, 85–131.

Morrison, Steven A., and Clifford Winston, *The Economic Effects of Airline Deregulation.* Brookings, Washington DC, 1986.

Morrison, Steven A., and Clifford Winston, "Enhancing the Performance of the Deregulated Air Transportation System," *Brookings Papers on Economic Activity: Microeconomics,* 1989, 61–112.

Morrison, Steven A., and Clifford Winston, *The Evolution of the Airline Industry.* Brookings, Washington DC, 1995.

Morrison, Steven A., and Clifford Winston, "The Fare Skies: Air Transportation and Middle America," *Brookings Review,* Fall 1997, 42–45.

Morrison, Steven A., and Clifford Winston, "Regulatory Reform of U.S. Intercity Transportation." In Jose A. Gomez-Ibanez, William B. Tye, and Clifford Winston, eds. *Essays in Transportation Economics and Policy: A Handbook in Honor of John R. Meyer.* Brookings, Washington DC, 1998.

Nickell, Stephen J., "Competition and Corporate Performance," *Journal of Political Economy,* August 1996, *104,* 724–46.

Noll, Roger G., "Economic Perspectives on the Politics of Regulation." In Richard Schmalensee and Robert Willig, eds. *The Handbook of Industrial Organization, Volume 2.* Amsterdam: North Holland, 1989, 1253–87.

Noll, Roger G., "The Economics and Politics of the Slowdown in Regulatory Reform," Brookings Institution working paper, Washington DC, 1996.

Peltzman, Sam, "The Economic Theory of Regulation after a Decade of Deregulation," *Brookings Papers on Economic Activity: Microeconomics,* 1989, 1–41.

Smith, Vernon L., "Regulatory Reform in the Electric Power Industry," *Regulation,* 1996, *1,* 33–46.

Taylor, William E., Charles J. Zakardas, and Douglas J. Zona, "Incentive Regulation and the Diffusion of New Technology," presented at the International Telecommunications Annual Conference, 1992.

Werden, Gregory J., Andrew S. Joskow, and Richard L. Johnson, "The Effects of Mergers on Price and Output: Two Case Studies from the Airline Industry," *Managerial and Decision Economics,* October 1991, *12,* 341–52.

White, Matthew W., "Playing with Power: Price and Entry Deregulation in Retail Electricity Markets," *Brookings Papers on Economic Activity: Microeconomics,* 1996, 201–67.

Winston, Clifford, "Economic Deregulation: Days of Reckoning for Microeconomists," *Journal of Economic Literature,* September 1993, *31,* 1263–89.

Winston, Clifford, Thomas M. Corsi, Curtis M. Grimm, and Carol A. Evans, *The Economic Effects of Surface Freight Deregulation.* Brookings, Washington DC, 1990.

Winston, Clifford, and Robert W. Crandall, "Explaining Regulatory Policy," *Brookings Papers on Economic Activity: Microeconomics,* 1994, 1–49.

Winston, Clifford, and Chad Shirley, *Alternate Route: Toward Efficient Urban Transportation.* Brookings Institution, Washington DC, 1998.

915 F2d 17 Town of Concord Massachusetts v. Boston Edison Company

915 F.2d 17

59 USLW 2201, 1990-2 Trade Cases 69,193

TOWN OF CONCORD, MASSACHUSETTS, et al., Plaintiffs, Appellees,

v.

BOSTON EDISON COMPANY, Defendant, Appellant.

No. 89-1872.

United States Court of Appeals,

First Circuit.

Heard April 5, 1990.

Decided Sept. 21, 1990.

William L. Patton with whom Robert J. Stillman, Ropes & Gray, Boston, Mass., George F. Bruder, Carmen L. Gentile, Bruder, Gentile & Marcoux, Washington, D.C., John J. Desmond, III and Jeffrey N. Stevens, Boston, Mass., were on brief, for defendant, appellant.

Charles F. Wheatley, Jr., with whom Don Charles Uthus, Peter A. Goldsmith, Timothy P. Ingram, Pace J. McConkie, Allen D. Freemyer, Wheatley & Ranquist, Arlington, Va., Charles R. Parrott, Nutter, McClennen & Fish, Boston, Mass., Albert S. Robinson, and Grindle, Robinson & Kertzman, Wellesley, Mass., were on brief, for plaintiffs, appellees.

Before BREYER, Chief Judge, and CAMPBELL and SELYA, Circuit Judges.

BREYER, Chief Judge.

1

The chief question raised by this appeal is whether a pricing practice known as a price squeeze violates the antitrust laws when it takes place in a fully regulated industry.

2

To understand the nature of a price squeeze the reader must keep three basic facts in mind. First, a firm can engage in a price squeeze only if it operates at two levels of an industry, and only if its competitors at one level are also its customers. Alcoa, the defendant in a famous antitrust case,

provides an example of such a firm. See United States v. Aluminum Co., 148 F.2d 416 (2d Cir.1945) (Alcoa). For several decades, Alcoa controlled the domestic production of almost all of America's aluminum ingot. It sold its ingot to independent fabricators, some of whom turned the ingot into aluminum sheet; it also fabricated sheet itself. Both Alcoa and the independents sold the finished sheet to the same group of consumers. Alcoa, therefore, operated at two levels of the aluminum business--ingot production and sheet production--and it competed with its customers (the independent sheet fabricators) at the second level.

3

Second, a price squeeze occurs when the integrated firm's price at the first level is too high, or its price at the second level is too low, for the independent to cover its costs and stay in business. Suppose, hypothetically, that Alcoa's price for ingot was $100 per ton; that the independents' costs of fabricating ingot into sheet was $50 per ton; and that Alcoa's price for sheet was $145 per ton. Under these circumstances, the independents, with ingot costs of $100 and fabricating costs of $50, would have no "room" to make a profit, for they could not charge more than $145 for sheet without losing all of their business to Alcoa. Alcoa's prices of $100 for ingot and $145 for sheet would squeeze the independents out of business.

4

Third, Judge Learned Hand, in United States v. Aluminum Co., supra, wrote that a price squeeze violates Sherman Act Sec. 2, 15 U.S.C. Sec. 2(a), when (1) the firm conducting the squeeze has monopoly power at the first industry level, (2) its price at this level is "higher than a 'fair price,' " and (3) its price at the second level is so low that its competitors cannot match the price and still make a "living profit." See id. at 437-38. Other courts, using substantially similar language, have reached the same conclusion. See, e.g., Bonjorno v. Kaiser Aluminum & Chem. Corp., 752 F.2d 802, 808-11 (3d Cir.1984), cert. denied, 477 U.S. 908, 106 S.Ct. 3284, 91 L.Ed.2d 572 (1986); George C. Frey Ready-Mixed Concrete, Inc. v. Pine Hill Concrete Mix Corp., 554 F.2d 551, 553 (2d Cir.1977); Carl Hizel & Sons, Inc. v. Browning-Ferris Indus., Inc., 600 F.Supp. 161, 161-62 (D.Colo.1985).

5

This case raises a narrow question about the price squeeze theory of antitrust liability: does Sherman Act Sec. 2 forbid a governmentally regulated firm with fully regulated prices--prices that are regulated at both industry levels--from asking regulators to approve prices that could create a price squeeze? Despite language in some cases suggesting that the answer is yes, see pp. 28-29, infra, our

analysis of the likely effects of a price squeeze in a fully regulated industry leads us to conclude that the answer is no. Effective price regulation at both the first and second industry levels makes it unlikely that requesting such rates will ordinarily create a serious risk of significant anticompetitive harm. At the same time, regulatory circumstances create a significant risk that a court's efforts to stop such price requests will bring about the very harms--diminished efficiency, higher prices--that the antitrust laws seek to prevent. We conclude, therefore, that price regulation will, in most cases, prevent a price squeeze from constituting an "exclusionary practice" of the sort that Sherman Act Sec. 2 forbids.

6

We also conclude that, regardless, the plaintiffs in this particular case did not demonstrate the existence of an unlawful exclusionary practice, for the evidence does not support a critical jury finding. It does not show that the defendant, Boston Edison, possessed monopoly power in the primary product market.

7

For these reasons, we reverse a judgment in favor of the plaintiffs.

8

* Background

9

The private firms that supply American homes and businesses with electricity are called "investor-owned utilities." Most of these firms are fully integrated, operating at all three levels of the electric power industry: (1) they produce (or generate) electricity, (2) they transmit electricity from generators to local distributors, and (3) they distribute electricity at the local level. See generally Joskow, Mixing Regulatory and Antitrust Policies in the Electric Power Industry: The Price Squeeze and Retail Market Competition, in Antitrust and Regulation 175-78 (F.M. Fisher ed. 1985) [hereinafter Joskow]; Meeks, Concentration in the Electric Power Industry: The Impact of Antitrust Policy, 72 Colum.L.Rev. 64, 67-69 (1972) [hereinafter Meeks]. Despite being fully integrated, these firms typically do not distribute only electricity that they themselves have generated. Rather, in many parts of the country, groups of firms have formed power "pools" through which they coordinate the generation, transmission, and distribution of electricity throughout a large geographical area. These pooling arrangements aim to enhance reliable and efficient electric service by matching customer

demands (which change minute by minute) with available low-cost supply sources. As a result of power pooling, an integrated utility will often wind up distributing electricity generated by a different, interconnected company. See generally 16 U.S.C. Sec. 824a-1 (governing pooling agreements); New England Power Pool Agreement, 56 F.P.C. 1562 (1976) (describing the New England Power Pool), petitions for review denied, Municipalities of Groton v. FERC, 587 F.2d 1296 (D.C.Cir.1978); C. Phillips, Jr., The Regulation of Public Utilities 585-92 (1988) (discussing coordination among electric utilities) [hereinafter C. Phillips]. Even in the absence of explicit pooling arrangements, the many physical interconnections among American utilities, combined with the tendency of electricity to flow instantaneously along interconnected lines to wherever it is demanded, make it likely that electricity distributed by one utility will be supplied by another. One legal consequence of these interconnections and pooling arrangements is that most electricity flows "interstate," permitting federal, as well as state, regulation of electricity. See Cincinnati Gas & Elec. Co. v. FPC, 376 F.2d 506, 507-09 (6th Cir.), cert. denied, 389 U.S. 842, 88 S.Ct. 77, 19 L.Ed.2d 106 (1967); Public Serv. Co. v. FPC, 375 F.2d 100 (7th Cir.), cert. denied, 387 U.S. 931, 87 S.Ct. 2054, 18 L.Ed.2d 992 (1967); see generally FPC v. Florida Power & Light Co., 404 U.S. 453, 92 S.Ct. 637, 30 L.Ed.2d 600 (1972).

10

Investor-owned utilities of the type we have described supply approximately 75% of the nation's electricity. The remainder is supplied by government- and consumer-owned systems. Unlike their investor-owned counterparts, these systems tend not to be integrated. The federal government, for example, generates about 10% of the nation's electricity (at systems such as the Tennessee Valley Authority), but sells most of it to independent distributors. Approximately 2,000 municipalities and cooperatives own small-scale distribution companies, but only a handful generate their own electricity; most purchase the electricity they distribute from investor-owned utilities or from the federal government. In contrast to investor-owned systems, municipal systems typically enjoy generous tax and financing subsidies and substantial freedom from regulation. See generally Joskow at 175-78; Joskow and Schmalensee, Markets for Power: An Analysis of Electric Utility Deregulation 17-20 (1983); Lopatka, The Electric Utility Price Squeeze as an Antitrust Cause of Action, 31 UCLA L.Rev. 563, 568 & n. 29 (1984) [hereinafter Lopatka]; Meeks, 72 Colum.L.J. at 67-69.

11

The record before us reveals facts typical of the national pattern. The defendant, Boston Edison, is a fully integrated investor-owned utility. It belongs to the New England Power Pool ("NEPOOL"). It transmits electricity (either self-generated or, e.g., secured from other NEPOOL members) to 52 cities

and towns in eastern Massachusetts. Each of these towns has only one distribution system serving all of the towns' inhabitants. In 39 of the towns, Edison owns the distribution system. In the remaining towns, either the town itself or another utility owns the distribution system. Concord and Wellesley, the plaintiffs, are two of the towns that own their own distribution systems. They take all of their electricity over Edison's transmission lines, and buy most of it directly from Boston Edison.

12

Edison's rates are completely regulated. The Federal Energy Regulatory Commission ("FERC," or the "Commission") regulates the rates that Edison charges when it sells electricity at "wholesale" (i.e., when it sells electricity to other firms, including municipal distributors, for resale to final customers). See 16 U.S.C. Sec. 824 (regulating the "transmission ... and ... sale of electric energy at wholesale in interstate commerce"). Before Edison can increase its rates, it must publish the new rates in the Federal Register and give notice of the proposed increase to each of its wholesale customers. See 18 C.F.R. Sec. 35.8 (1990). If an interested party raises a legitimate question about the rate increase, FERC normally will suspend the new rates for up to five months, investigate them, and, if it has not completed its investigation by then, permit them to take effect subject to refund in the event it ultimately rejects them. See 16 U.S.C. Sec. 824d(d, e). Similarly, the Massachusetts Department of Public Utilities ("DPU") regulates the "retail" rates that Boston Edison charges to consumers in the 39 cities and towns in which it distributes electricity. See M.G.L. c. 164, Secs. 93-94G. In contrast, as we mentioned earlier, no agency regulates the retail rates that municipal distributors, such as Concord and Wellesley, charge to their customers, for these distributorships are non-profit entities indirectly controlled (through the ballot) by the very customers they serve.

13

Between 1984 and 1987, Edison filed a series of wholesale rate increases with FERC. Wellesley and Concord vigorously opposed those rate increases before the Commission. In each case, after briefly suspending the rates, the Commission allowed them to take effect subject to refund. The towns sought judicial review of FERC's decisions. Eventually, with one exception not relevant here, see Towns of Wellesley, Concord and Norwood v. FERC, 786 F.2d 463 (1st Cir.1986), this court upheld as lawful the relevant FERC determinations, see Towns of Concord and Wellesley v. FERC, 844 F.2d 891 (1st Cir.1988); Boston Edison Co. v. FERC, 885 F.2d 962 (1st Cir.1989).

14

Concord and Wellesley then attacked the same wholesale rate increases on another legal front. They brought this antitrust action in federal district court. They pointed out that these wholesale rate increases filed with FERC were not matched by corresponding retail rate increases. Thus, over a three-year period, Edison's rates to Concord and Wellesley increased while its rates to its own retail customers in the 39 towns surrounding Concord and Wellesley stayed roughly the same. The effect of this disparity, said Concord and Wellesley, was to put the two towns in a price squeeze. If they did not raise their own retail rates to reflect the higher cost of buying electricity from Edison, they would make less money; but if they raised their retail rates while Edison's stayed the same, customers for whom they and Edison compete (say, a small manufacturer with high energy costs) might decide to forego Concord and Wellesley and settle in a town served by Edison. In light of the resulting potential harm to their electricity distributing businesses, Concord and Wellesley argued that the wholesale price increases, in the absence of corresponding retail price increases, amounted to unlawful monopolization. See Sherman Act Sec. 2, 15 U.S.C. Sec. 2(a). A jury agreed, and the district court declined to overturn its verdict. See Town of Concord v. Boston Edison Co., 721 F.Supp. 1456 (D.Mass.1989). We disagree with the district court, and we reverse the judgment.

II

The Price Squeeze

15

The basic legal principles that govern this monopolization case are well settled. Even though Boston Edison is a regulated firm, it has no blanket immunity from the antitrust laws. See Otter Tail Power Co. v. United States, 410 U.S. 366, 372-75, 93 S.Ct. 1022, 1027-28, 35 L.Ed.2d 359 (1973); see also Cantor v. Detroit Edison, 428 U.S. 579, 596 n. 35, 96 S.Ct. 3110, 3120 n. 35, 49 L.Ed.2d 1141 (1976); Gulf States Utils. Co. v. FPC, 411 U.S. 747, 758-59, 93 S.Ct. 1870, 1877-78, 36 L.Ed.2d 635 (1973). The antitrust law relevant in this case makes it unlawful for a firm to "monopolize ... any part of the trade or commerce among the several States." Sherman Act Sec. 2, 15 U.S.C. Sec. 2(a). To prove a violation of this statute, a plaintiff must demonstrate (1) that the defendant possesses "monopoly power in the relevant market," and (2) that the defendant has acquired or maintained that power by improper means. See United States v. Grinnell Corp., 384 U.S. 563, 570-71, 86 S.Ct. 1698, 1703-04, 16 L.Ed.2d 778 (1966); Barry Wright Corp. v. ITT Grinnell Corp., 724 F.2d 227, 230 (1st Cir.1983). Like many courts and commentators, we refer to improper methods of acquiring or maintaining monopoly power as "exclusionary conduct." Borrowing from Professors Areeda and Turner, we have defined "exclusionary conduct" as " 'conduct, other than competition on the merits or

restraints reasonably "necessary" to competition on the merits, that reasonably appears capable of making a significant contribution to creating or maintaining monopoly power.' " Barry Wright, 724 F.2d at 230 (quoting 3 P. Areeda & D. Turner, Antitrust Law p 626, at 83 (1978) [hereinafter Areeda & Turner].

16

We shall assume for now that Edison has monopoly power in a relevant market and begin by examining whether Edison's pricing practices amount to exclusionary conduct. Traditional antitrust principles will guide our analysis. We shall compare the challenged practice's likely anticompetitive effects with its potentially legitimate business justifications. See Clamp-All Corp. v. Cast Iron Soil Pipe Inst., 851 F.2d 478, 486 (1st Cir.1988), cert. denied, 488 U.S. 1007, 109 S.Ct. 789, 102 L.Ed.2d 780 (1989); Interface Group, Inc. v. Massachusetts Port Auth., 816 F.2d 9, 10 (1st Cir.1987); Meeks, 72 Colum.L.J. at 80. In doing so, we shall bear in mind that a practice is not "anticompetitive" simply because it harms competitors. After all, almost all business activity, desirable and undesirable alike, seeks to advance a firm's fortunes at the expense of its competitors. Rather, a practice is "anticompetitive" only if it harms the competitive process. See Brunswick Corp. v. Pueblo Bowl-O-Mat, Inc., 429 U.S. 477, 488-89, 97 S.Ct. 690, 697-98, 50 L.Ed.2d 701 (1977); Brown Shoe Co. v. United States, 370 U.S. 294, 320, 82 S.Ct. 1502, 1521, 8 L.Ed.2d 510 (1962); Clamp-All, 851 F.2d at 486. It harms that process when it obstructs the achievement of competition's basic goals--lower prices, better products, and more efficient production methods. See, e.g., Grappone, Inc. v. Subaru of New England, Inc., 858 F.2d 792, 794 (1st Cir.1988); Rothery Storage & Van Co. v. Atlas Van Lines, Inc., 792 F.2d 210, 228-29 (D.C.Cir.1986), cert. denied, 479 U.S. 1033, 107 S.Ct. 880, 93 L.Ed.2d 834 (1987); Olympia Equip. Leasing Co. v. Western Union Telegraph Co., 797 F.2d 370, 379 (7th Cir.1986), on denial of r'hrg, 802 F.2d 217, cert. denied, 480 U.S. 934, 107 S.Ct. 1574, 94 L.Ed.2d 765 (1987); Assam Drug Co. v. Miller Brewing Co., 798 F.2d 311, 315 (8th Cir.1986); Westman Comm'n Co. v. Hobart Int'l, Inc., 796 F.2d 1216, 1220 (10th Cir.1986), cert. denied, 486 U.S. 1005, 108 S.Ct. 1728, 100 L.Ed.2d 192 (1988); see also Reiter v. Sonotone, 442 U.S. 330, 343, 99 S.Ct. 2326, 2333, 60 L.Ed.2d 931 (1979) ("Congress designed the Sherman Act as a 'consumer welfare prescription.' ") (quoting R. Bork, The Antitrust Paradox 66 (1978)); Apex Hosiery Co. v. Leader, 310 U.S. 469, 500-01, 60 S.Ct. 982, 996-97, 84 L.Ed. 1311 (1940) ("Restraints on competition" do not constitute antitrust violations unless they "have or [are] intended to have an effect upon prices in the market or otherwise ... deprive purchasers or consumers of the advantages which they derive from free competition").

17

Finally, we shall take account of the institutional fact that antitrust rules are court-administered rules. They must be clear enough for lawyers to explain them to clients. They must be administratively workable and therefore cannot always take account of every complex economic circumstance or qualification. See Barry Wright, 724 F.2d at 234. (Indeed, the need for clarity and administrability sometimes leads to per se rules that prohibit inquiry into the actual harms and benefits of challenged conduct. See Northern Pac. R. Co. v. United States, 356 U.S. 1, 5, 78 S.Ct. 514, 518, 2 L.Ed.2d 545 (1958).) They must be designed with the knowledge that firms ultimately act, not in precise conformity with the literal language of complex rules, but in reaction to what they see as the likely outcome of court proceedings.

18

These last-mentioned administrative considerations are particularly important when courts apply antitrust law to a regulated industry. That is because "regulation" and "antitrust" typically aim at similar goals--i.e., low and economically efficient prices, innovation, and efficient production methods--but they seek to achieve these goals in very different ways. Economic regulators seek to achieve them directly by controlling prices through rules and regulations; antitrust seeks to achieve them indirectly by promoting and preserving a process that tends to bring them about. An antitrust rule that seeks to promote competition but nonetheless interferes with regulatory controls could undercut the very objectives the antitrust laws are designed to serve. Thus, where regulatory and antitrust regimes coexist, see, e.g., Otter Tail, 410 U.S. at 372, 93 S.Ct. at 1027 (electric industry); United States v. Philadelphia Nat'l Bank, 374 U.S. 321, 352, 83 S.Ct. 1715, 1735, 10 L.Ed.2d 915 (1963) (bank mergers), antitrust analysis must sensitively "recognize and reflect the distinctive economic and legal setting" of the regulated industry to which it applies. Watson & Brunner, Monopolization by Regulated "Monopolies": The Search for Substantive Standards, 22 Antitrust Bull. 559, 565 (1977) [hereinafter Watson & Brunner]; see generally Meeks, supra.

19

For reasons we shall now set out, these principles lead us to conclude that a price squeeze of the sort at issue here does not ordinarily violate Sherman Act Sec. 2 where the defendant's prices are regulated at both the primary and secondary levels. In so holding, we are not saying either that the antitrust laws do not apply in this regulatory context, or that they somehow apply less stringently here than elsewhere. Rather, we are saying that, in light of regulatory rules, constraints, and practices, the price squeeze at issue here is not ordinarily exclusionary, and, for that reason, it does not violate the Sherman Act. Cf. 1 Areeda & Turner p 223d, at 140 (noting that "antitrust courts can and do consider

the particular circumstances of an industry and therefore adjust their usual rules to the existence, extent, and nature of regulation"); Watson & Brunner, 22 Antitrust Bull. at 560 (warning that "the dogmatic transposition of monopolization concepts from conventional market settings to regulated industries" may produce results inimical to the goals of antitrust). We reach this conclusion by (1) analyzing the ordinary price squeeze, (2) comparing it to the "regulatory" price squeeze, and (3) noting that regulation makes a critical difference in terms of antitrust harms, benefits, and administrative considerations.

20

A. Price Squeeze Liability in an Unregulated Industry

21

To explain our conclusion, we must first discuss the ordinary price squeeze case, the case of an unregulated monopolist whose prices (at, say, the wholesale and retail levels) leave inadequate room for an independent competitor (at, say, the retail level) to survive. Only if the reader understands the major antitrust harms, benefits, and administrative considerations in that ordinary case, and sees how they are rather evenly balanced, will he understand why price regulation tips that balance so significantly towards legality. Cf. 3 Areeda & Turner p 729c, at 235-38 (arguing that Alcoa overstated the harms, understated the benefits, and disregarded significant administrative burdens associated with the ordinary price squeeze).

22

The antitrust argument against the price squeeze typically begins by observing that a "one-level" monopolist engaged in a prolonged price squeeze may drive independent competitors out of business and thereby extend its monopoly power to a second industry level. This fact alone, however, does not mean that a price squeeze is anticompetitive. Merely eliminating competitors is not necessarily anticompetitive, for, as we explained earlier, even legitimate business activity that succeeds in helping a firm will likely disadvantage the firm's competitors. See pp. 21-22, supra, and cases cited. Moreover, the extension of monopoly power from one to two levels does not necessarily, nor in an obvious way, give a firm added power to raise prices. As several members of the Supreme Court have pointed out, a "widely accepted" (albeit "counterintuitive") economic argument supports the conclusion of many commentators that "there is but one maximum monopoly profit to be gained from the sale of an end-product," 3 Areeda & Turner p 725b at 199. See Jefferson Parish Hosp. Dist. No. 2 v. Hyde, 466 U.S. 2, 36, 104 S.Ct. 1551, 1570, 80 L.Ed.2d 2 (O'Connor, J., concurring in the

judgment); see also R. Bork, The Antitrust Paradox 229 (1978) ("[V]ertically related monopolies can take only one monopoly profit") [hereinafter R. Bork]; R. Posner & F. Easterbrook, Antitrust 870 (2d ed. 1989) ("There is only one monopoly profit to be made in a chain of production."). (For an intuitive explanation of this "counterintuitive" argument, see Appendix A to this opinion.) What difference does it make, one might ask, whether the monopolist obtains this "maximum monopoly profit" by controlling two industry levels or just one?

23

At least two arguments have been made in answer to this question. The first is that a monopolist who extends his monopoly to a second industry level raises "entry barriers," thereby fortifying his monopoly position. Assume, for example, that Alcoa has monopolized only ingot production and has raised its prices somewhat above the competitive level. Alcoa's monopoly position in ingot will no doubt deter some new firms from entering ingot production, but it might also attract some new firms hoping to obtain a share of Alcoa's supercompetitive profits. Now assume Alcoa extends its monopoly to sheet production as well, the industry level, say, that consumes most of Alcoa's ingot. New firms might now be more reluctant to enter ingot production, fearing that Alcoa, in its capacity as "sheet fabricator-monopolist," would refuse to buy ingot from the new ingot producer, or that it would drop its own ingot prices and take the industry's entire "single monopoly profit" from the sale of sheet. The new firm might conclude that the only prudent way to challenge Alcoa would be to enter the industry at both levels at once. Insofar as it is more difficult for a firm to enter an industry at two levels than at one, the monopolist, by expanding its monopoly power, has made entry by new firms more difficult. And insofar as the monopolist previously set prices cautiously to avoid attracting a competitive challenge, the added security of a two-level monopoly could even lead that monopolist to raise its prices. See 3 Areeda & Turner p 725h, at 204-08; see also id. p 1011, at 243-54; Shepherd, Potential Competition Versus Actual Competition, 42 Admin.L.Rev. 5-34 (1990).

24

The second argument against permitting extension of monopoly power concerns "non-price" competition. The existence of competitors at a second "level," irrespective of their effects upon price, provides an added incentive for the monopolist to develop better products and better, more efficient ways to produce the product. See 2 Areeda & Turner pp 402b2-402b3, at 269-70; id. p 403c, at 272 & n. 3; 3 Areeda & Turner p 725g, at 204. Indeed, a "second-level" independent firm that develops better products or more efficient production methods may thereby obtain the strength needed to challenge the monopolist at the "primary level."

25

We recognize that there are other arguments one might make about the possible anticompetitive effects of a price squeeze, see generally 3 Turner & Areeda pp 725e-725f, at 201-04 (reviewing other arguments); and we realize, as well, that scholars dispute the practical significance of the two arguments just made. Compare 3 Areeda & Turner pp 725g-725h, at 204-08 with R. Bork at 240-43. Nonetheless, we believe that other arguments (such as those related to economic price discrimination) tend to be inconclusive in respect to anticompetitive effects. See 3 Areeda & Turner pp 725e-725f, at 201-04; R. Bork at 241-42. We therefore believe that the two arguments just given are the most important of those that favor prohibiting a price squeeze; and we are willing to assume they would justify such a prohibition were there nothing to be said in favor of permitting prices that could create such a squeeze.

26

There are, however, at least two traditional circumstances in which prices that create a squeeze might simultaneously bring about economic benefits. First, the primary-level monopolist might carry out its second-level activities more efficiently than its independent competitors. If so, prices that squeeze the less efficient second-level competitors, even to the point of forcing them from the business, could (by lowering costs) lower prices, or, in any event, save economic resources. See 3 Areeda & Turner p 725d, at 201 ("Whenever vertical integration produces new efficiencies, some of the cost savings ... [will] be translated into a price reduction and ... [will] save productive resources."); R. Bork at 243.

27

Second, prices that squeeze a "second-level" firm will benefit consumers whenever the "second-level" firm is itself a monopolist. See Fishman v. Estate of Wirtz, 807 F.2d 520, 563 (Easterbrook, J., dissenting) ("That successive monopolies injure consumers is a proposition on which there is unanimous agreement") (collecting authority). If, for example, ingot costs $40, the fabricating process costs $35, and the profit-maximizing price for sheet is $100, an ingot monopolist will charge $65 for the ingot, hoping that competition at the fabricating level will keep the total price at $100. If a different, independent monopolist dominates the fabricating level, however, that independent monopolist buying ingot at $65 will mark up the price by more than $35, because he wants to earn monopoly profits as well. The result will be a market price of more than $100, resulting in smaller monopoly profits overall (for the final price is too high), but greater profits for the second monopolist than if he sold sheet for only $100. (Appendix B to this opinion contains a numerical example that

may help the interested reader understand intuitively why this is so.) Under these circumstances, entry by the ingot monopolist into the sheet-fabrication level--even by means of a price squeeze--will help the consumer by limiting the final price of sheet to $100. Cf. Atlantic Richfield Co. v. USA Petroleum Co., --- U.S. ----, 110 S.Ct. 1884, 1891, 109 L.Ed.2d 333 (1990) ("When a firm, or even a group of firms adhering to a vertical agreement, lowers prices but maintains them above predatory levels, the business lost by rivals cannot be viewed as an 'anticompetitive' consequence of the claimed violation." (footnote omitted)); see generally 3 Areeda & Turner p 725c, at 200-01.

28

Again, the price squeeze may have additional benefits, see generally 3 Turner & Areeda pp 725e-725f, at 201-04, and one can debate the empirical significance of those we have mentioned. But we believe that the two examples we have given show that prices that create a squeeze can, at least sometimes, have important beneficial effects.

29

Finally, we note that it is not easy for courts to administer Judge Hand's price squeeze test. That test makes it unlawful for a monopolist to charge more than a "fair price" for the primary product while simultaneously charging so little for the secondary product that its second-level competitors cannot make a "living profit." See Alcoa, 148 F.2d at 437-38. But how is a judge or jury to determine a "fair price?" Is it the price charged by other suppliers of the primary product? None exist. Is it the price that competition "would have set" were the primary level not monopolized? How can the court determine this price without examining costs and demands, indeed without acting like a rate-setting regulatory agency, the rate-setting proceedings of which often last for several years? Further, how is the court to decide the proper size of the price "gap?" Must it be large enough for all independent competing firms to make a "living profit," no matter how inefficient they may be? If not, how does one identify the "inefficient" firms? And how should the court respond when costs or demands change over time, as they inevitably will?

30

We do not say that these questions are unanswerable, but we have said enough to show why antitrust courts normally avoid direct price administration, relying on rules and remedies (such as structural remedies, e.g., prohibiting certain vertical mergers) that are easier to administer. See United States v. Trenton Potteries Co., 273 U.S. 392, 397, 47 S.Ct. 377, 379, 71 L.Ed. 700 (1927) (observing that "[t]he reasonable price fixed today may through economic and business changes

become the unreasonable price of tomorrow"); United States v. Addyston Pipe & Steel Co., 85 F. 271, 283 (6th Cir.1898) (Taft, J.) (stating that to inquire into the reasonableness of prices is to "set sail on a sea of doubt"), aff'd, 175 U.S. 211, 20 S.Ct. 96, 44 L.Ed. 136 (1899); 3 Areeda & Turner p 701, at 148-50 ("The courts correctly regard as uncongenial and foreign to the Sherman Act the burden of continuously supervising economic performance.").

31

To reach Alcoa's conclusion that the price squeeze is exclusionary, one must believe that the anticompetitive risks associated with a price squeeze outweigh the possible benefits and the adverse administrative considerations. We do not question that conclusion here, but we qualify it by noting that the harms and benefits (including the more doubtful ones we have glossed over, see generally 3 Areeda & Turner pp 725e-725f, at 201-04), are closely balanced. In light of this close balance, we should not blindly and automatically apply Alcoa's price squeeze rules in the presence of significantly different economic circumstances such as full economic regulation.

32

B. Price Squeeze Liability in a Regulated Industry

33

Full price regulation dramatically alters the calculus of antitrust harms and benefits. First, regulation significantly diminishes the likelihood of major antitrust harm. In particular, it diminishes the likelihood of "entry barrier" harm, namely the risk that (1) prices will rise because (2) new firms will hesitate to enter a market and compete after (3) a squeeze has driven pre-existing independent competitors from the marketplace. All three propositions are made doubtful by regulation. For one thing, in a regulated industry, regulators control prices directly. Statutes typically require regulators to maintain prices at "reasonable" levels. See, e.g., 16 U.S.C. Secs. 824d(a), 824e(a) (an electric utility's rates must be "just and reasonable"); FPC v. Hope Natural Gas Co., 320 U.S. 591, 603, 64 S.Ct. 281, 288, 88 L.Ed. 333 (1944) (A "reasonable" rate is one that permits the firm to recover its costs and earn a reasonable profit); see generally J. Bonbright, A. Danielsen & D. Kamerschen, Principles of Public Utility Rates (2d ed.1988) (describing ratemaking principles) [hereinafter J. Bonbright].

34

For another thing, factors related to regulation, such as the economic ability of a market to support new entry, or the legal requirement that a firm secure permission to enter, are likelier to determine

new entry into a regulated industry than is a new entrant's fear of a two-level monopolist's enhanced retaliatory power. After all, should the regulator decide that new entry is warranted, it typically has the legal authority to prevent an existing "two-level" monopolist from improperly disadvantaging a new "second-level" competitor by, say, refusing to deal to with it or by charging unreasonably high prices. See, e.g., 16 U.S.C. Secs. 824f, 824i-824k (authorizing FERC to order interconnections, sales and wheeling of electricity among utilities); see also 2 A. Kahn, The Economics of Regulation 113-172 (discussing the concept of "natural monopoly") [hereinafter A. Kahn]; Meeks, 72 Colum.L.Rev. at 95 & n. 135 (noting that utilities in "all but a handful of states ... must obtain a certificate of public convenience and necessity to serve an area" with electricity) (collecting statutes).

35

Finally, regulation in the electricity industry makes it less likely that a price squeeze will actually drive independent distributors from the marketplace. Local electricity distribution, after all, is itself considered a "natural monopoly" with service typically provided by a single franchised company to relatively immobile customers. See Joskow at 178; cf. Otter Tail, 410 U.S. at 369, 93 S.Ct. at 1025. Higher wholesale electricity prices (or lower retail prices elsewhere) will not lead the entire towns of Wellesley and Concord to pull up stakes and move to Dover. And the record lacks convincing evidence that competition between Edison and the plaintiff towns for customers living on the towns' borders, or for mobile high-energy-consuming industrial customers, is sufficiently important to threaten the plaintiffs with bankruptcy should such users switch to Edison. Cf. Joskow at 209-13 (arguing that, as a rule, there is little actual competition between electricity distributors). Even if an integrated utility such as Edison managed to set prices that severely squeezed a distributor, it could not take over the municipality's distribution area without the regulator's permission. See, e.g., M.G.L. ch. 164 Sec. 68 (Massachusetts municipality may not sell its distributorship without regulatory approval).

36

Second, regulators try to set prices that reflect costs. To the extent that they succeed, an integrated utility's prices are likely to squeeze independent distributors who buy from it at wholesale only if those distributors operate less efficiently, i.e., at higher cost. Consequently, a rule preventing prices that create a squeeze will more likely discourage efficient operations and deprive consumers of prices that reflect lower costs.

37

Third, for institutional reasons, an antitrust rule making the price squeeze illegal threatens consumers. The plaintiffs say that Edison violated the antitrust laws by asking for rates (i.e., filing rate proposals) that would produce for Edison a "higher return" on its "wholesale sales than.... on its retail sales." See Appellee's Brief at 4-8, 35-36. Other courts considering price squeeze claims have also emphasized that the alleged antitrust violation consisted of proposing or requesting rates that would narrow the gap between wholesale and retail rates so as to squeeze existing retail competitors. See, e.g., City of Kirkwood v. Union Elec. Co., 671 F.2d 1173, 1179 (8th Cir.1981) ("[T]he question" in a price squeeze case is "whether the defendant utility acted illegally in proposing a certain anticompetitive combination of rates") (emphasis added), cert. denied, 459 U.S. 1170, 103 S.Ct. 814, 74 L.Ed.2d 1013 (1983); City of Newark v. Delmarva Power & Light Co., 467 F.Supp. 763, 768 (D.Del.1979) (chief allegation in price squeeze case was that defendant "used its power to initiate rates ... to attempt to obtain an anticompetitive rate structure") (emphasis in original). A rule that penalizes the filing of wholesale rate increases would not necessarily lead firms to abandon wholesale rate increases; it could, instead, simply lead them to seek a retail rate increase whenever they seek a wholesale rate increase. (Conversely, it would discourage a utility from conceding that a retail price cut is reasonable unless it is also prepared to concede that a wholesale price cut is reasonable.) In this case, for example, the rule might have permitted Edison to escape liability by filing (or perhaps obtaining) a rate increase in the 39 cities and towns where it distributes electricity itself. Indeed, the rule would encourage Edison to seek a retail rate increase so high that a jury could not find the wholesale/retail gap too narrow. Such a rule of law (which could lead to unnecessary regulatory proceedings and unnecessary rate increases) would seem to work at cross-purposes with basic antitrust objectives. Cf. Matsushita Elec. Indus. Co. v. Zenith Radio Corp., 475 U.S. 574, 594, 106 S.Ct. 1348, 1360, 89 L.Ed.2d 538 (1986) (warning courts against adopting rules or practices that "chill the very conduct the antitrust laws are designed to protect"); Barry Wright, 724 F.2d at 234 ("[W]e must be concerned lest a rule or precedent that authorizes a search for a particular type of undesirable pricing behavior end up by discouraging legitimate price competition.").

38

Fourth, an anti-price-squeeze rule, particularly the "equal profit earned at both levels" rule that plaintiffs advocate, presents special administrative difficulties in the regulatory context. For one thing, the Constitution guarantees a regulated firm a reasonable return on its investment, see Hope Natural Gas Co., supra; Bluefield Water Works v. Public Serv. Comm'n, 262 U.S. 679, 690, 43 S.Ct. 675, 678, 67 L.Ed. 1176 (1923); see also 16 U.S.C. Secs. 824d(a), 824e(a) (rates must be "just and reasonable"). May courts prevent a utility from asking the regulator for a rate increase should the utility believe that current wholesale returns are, constitutionally speaking, too low? Similarly, what if

the utility believes that different risks at different levels of the industry call for somewhat different rates of return?

39

For another thing, how will juries decide the amount of the firm's investment on which the "rate of return" is calculated? Profit, after all, is a return on a certain amount of investment. Much electricity investment is used to serve both wholesale and retail services. Accounting systems for allocating investment between different services or customer classes are notoriously complicated and sometimes arbitrary. See J. Bonbright at 495 (noting that dozens of schemes exist for allocation investment among customer classes) (collecting sources); Joskow at 180 ("Despite the legal distinction between 'retail sales' ... and 'wholesale sales,' ... essentially there is no functional distinction between generation and transmission facilities dedicated to retail and wholesale transactions."). Will juries be permitted to "second guess" the regulator's allocation rules or its specific investment allocation decisions? Is the jury (as may, in a sense, have occurred here) to be permitted to decide that the utility's allocation (following Commission allocation procedures) of, say, $100 million investment to distribution and $400 million to wholesale, should have been $200 million/$300 million instead (in which case, of course, the pre-existing 10% profit earned at each level would look like 5% earned at retail and 13% at wholesale). What antitrust benefit would be gained by permitting juries to speculate in this way? Would not the results likely be arbitrary? For that matter, what antitrust benefit would be gained by a rule that freezes prices at levels designed to produce "equal profit rates" using some historically based commission-approved system for allocating investment? Any such court involvement in the rate-setting process could easily discourage utilities from proposing, and commissions from considering, such innovative, economically based, energy-conserving pricing systems as off-peak pricing (charging higher prices for electricity used during peak periods), incremental-cost pricing (charging prices for certain services reflecting the higher costs of hypothetical new construction), or even traditional Ramsay pricing (tailoring rates to reflect the comparative likelihood that higher rates will force customers to discontinue service); for there is no particular reason to believe that such rates would (or would not) reflect "equal" profits according to some particular prior, historical, commission-approved system for allocating investment. See generally 1 A. Kahn 87-103. At best, permitting judges and juries to speculate in this area would force regulators continually to revisit prior investment-allocation rules and would discourage regulatory efforts to develop (often with a utility's help) a set of economically rational pricing practices--which, after all, is a common objective of antitrust law and regulation.

40

Fifth and finally, a distributor who disagrees with the utility's, or the commission's, cost-allocation methods or with its rates, or who believes that a price squeeze will harm it (whether or not the price squeeze also harms the competitive process), has an administrative remedy. It can challenge a utility's rates and practices before the Commission as unjust, unreasonable, or "discriminatory." See 16 U.S.C. Secs. 824d(a, b), 824e(a); 18 C.F.R. Sec. 2.17(a) (1990). As part of its challenge, it can ask the Commission to consider the alleged price squeeze harm. See FPC v. Conway, 426 U.S. 271, 96 S.Ct. 1999, 48 L.Ed.2d 626 (1976). If FERC permits the rates to take effect after suspension and later determines that a price squeeze exists, it can order an appropriate refund.

41

In sum, the relevant antitrust considerations differ significantly, in degree and in kind, when a price squeeze occurs in a fully regulated as opposed to an unregulated industry. Indeed, these considerations, which are closely balanced in the ordinary price squeeze, change so significantly when the squeeze takes place in a fully regulated industry that, in our opinion, the legal consequences of the squeeze change as well. That is to say, a price squeeze in a fully regulated industry such as electricity will not normally constitute "exclusionary conduct" under Sherman Act Sec. 2.

42

We are aware that reported opinions in three circuit courts of appeal and several district courts suggest that a price squeeze may be unlawful in the regulated electricity industry. See City of Kirkwood v. Union Elec. Co., 671 F.2d 1173 (8th Cir.1982), cert. denied, 459 U.S. 1170, 103 S.Ct. 814, 74 L.Ed.2d 1013 (1983); City of Groton v. Connecticut Light & Power Co., 662 F.2d 921 (2d Cir.1981); City of Mishawaka v. American Elec. Power Co., 616 F.2d 976 (7th Cir.1980) (Mishawaka II), cert. denied, 449 U.S. 1096, 101 S.Ct. 892, 66 L.Ed.2d 824 (1981); see also City of Newark v. Delmarva Power & Light Co., 467 F.Supp. 763 (D.Del.1979); Borough of Ellwood City v. Pennsylvania Power Co., 462 F.Supp. 1343 (W.D.Pa.1979), later proceeding, 570 F.Supp. 553 (1983); City of Batavia v. Commonwealth Edison Co., No. 76-C-4388, slip op. (N.D.Ill. Jan 16, 1984). In only one of those cases, however, did the court actually enter, or affirm, a judgment against the defendant; thus, there is extremely little case law holding that a public utility price squeeze is exclusionary. Moreover, in all of the cases we have cited, the price squeeze allegation was but one of several allegations of several different kinds of "exclusionary conduct," and the one court that affirmed a Sherman Act Sec. 2 judgment against a utility stressed that it might well not have done so had the price squeeze stood alone (as it does in the case before us). See Mishawaka II, 616 F.2d at 986 (suggesting that "no one aspect [of the defendant's conduct] standing alone is illegal"). Further, in every case, the price

squeeze allegation involved wholesale prices that exceeded retail prices, (a matter here in doubt), thereby eliminating some of the "administrative" problems we have found surrounding a jury's efforts to determine the reasonableness of the price "gap." See City of Mishawaka I v. Indiana & Michigan Electric Company, 560 F.2d 1314, 1316 (7th Cir.1977) (wholesale price to municipalities "substantially higher" than retail price to industrial customers, threatening municipalities with "financial extinction"); City of Kirkwood, 671 F.2d at 1176; City of Groton, 662 F.2d at 925; City of Newark 467 F.Supp. at 766; Borough of Ellwood City 570 F.Supp. at 557; City of Batavia, supra. Thus, technically speaking, our holding is consistent with these other cases. Indeed, since we cannot find significant evidence that Edison's price squeeze deprived plaintiffs of a "living profit" (for the plaintiffs earned substantial returns, and even increased their sales, during the price squeeze period), our holding is arguably a consistent application of the rule set forth in Alcoa itself.

43

Nevertheless, we recognize that our reasoning and analysis differs significantly from that in the opinions (other than Alcoa) just cited. We have carefully considered these differences and have found nothing to convince us that our analysis is incorrect. Even so, we have limited our holding by stating that "normally" a price squeeze will not constitute an exclusionary practice in the context of a fully regulated monopoly, thereby leaving cases involving exceptional circumstances for another day. And we have stressed that our reasoning applies with full force only when the monopolist who engages in the squeeze is regulated at both industry levels. We recognize that a special problem is posed by a monopolist, regulated at only one level, who seeks to dominate a second, unregulated level, in order to earn at that second level the very profits that regulation forbids at the first. See 3 Areeda & Turner p 726e, at 217-20.

III

Monopoly Power

44

A second, legally fatal problem with plaintiffs' case is that they failed to show that Boston Edison possessed monopoly power in the relevant market. As we have repeatedly pointed out, the traditional, Alcoa-type price squeeze occurs when a monopolist uses its control over the monopolized product (e.g., ingot) to obtain a monopoly at a second industry level (e.g., sheet). In this case, it is undisputed that Boston Edison has a (lawful) monopoly at one level of the electricity industry, namely transmission, for a given area normally can support only one transmission system, see, e.g., Meeks,

72 Colum.L.Rev. at 87, and Edison owns the transmission grid covering the area that includes Concord and Wellesley. The plaintiffs, however, cannot show (indeed, with one exception, they do not try to show) that Boston Edison used its control of electricity transmission to their disadvantage. To the contrary, the record clearly demonstrates that Boston Edison will "wheel" (i.e., transmit) electricity from other producers to any distributor upon request, all for a nominal transmission charge that no one claims is unreasonably high. See Record Appendix at 1442, 1451-52, 1704-05 [hereinafter App.]. Indeed, Concord and Wellesley buy some of their electricity from the New York Power Authority, which wheels that electricity over Edison's lines, and at least half a dozen municipal distributorships use Edison's lines to obtain all of their electric power from non-Edison sources. See App. at 1449-50. Concord is a member of NEPOOL, all of whose members have interconnected lines, used to transfer electricity freely among them; and the record suggests that Wellesley is free to join NEPOOL should it wish to do so. The only arguably contrary evidence in the record consists of Concord's claim that Boston Edison refused to connect it to a 115 kv line; but the record does not (1) suggest that Concord needs that connection to buy electricity elsewhere, (2) indicate that the dispute concerns anything other than the connection price, or (3) show that the demanded connection price was unreasonably high.

45

At the same time, the record fails to show that Edison had monopoly power at the generation level, the industry level at which plaintiffs claim Edison charged too high a rate. To the contrary, the record shows that more than a dozen different power generating companies sell wholesale power in New England and that this power flowed freely among NEPOOL members. It shows that municipalities near Concord and Wellesley, municipalities to the north, south, east, and west of Concord and Wellesley, bought their electricity from production companies other than Boston Edison. It shows that Boston Edison's electricity production accounted for less than 12% of the electricity produced by these New England utilities. These facts and figures suggest that Boston Edison is far from attaining the 70% to 90% of the sales in a well-defined marketplace that Judge Hand (and later commentators) thought indicative of monopoly power. See, e.g., Alcoa, 148 F.2d at 424 (stating that a 90% market share "is enough to constitute a monopoly" but that 60% is probably, and 30% is "certainly," not enough); City of Mount Pleasant v. Associated Elec. Coop., Inc., 838 F.2d 268, 279-80 (8th Cir.1988) (electricity suppliers accessible via wheeling must be included in wholesale market); Lopatka, 31 U.C.L.A.L.Rev. at 611 (arguing that the geographic market for wholesale electricity "include[s] all generation sources from which the municipality could obtain power at relatively insubstantial wheeling costs").

46

In their efforts to show monopoly power at the generation level, the towns point to three pieces of evidence. First, they say that Concord had a study made "which concluded that it was not economically feasible for the Town ... to obtain its power requirements from non-Edison sources...." The study in question, however, says that "the reason" the town will have lower power costs in the long run by buying from Edison is that Edison's "rate is based on embedded resources which are lower priced than the new units with which" the town would have to "meet its power supply requirements" if it stopped buying from Edison. This "reason" does not show that Edison has monopoly power. It shows only that Edison can charge lower prices because it bought its generating equipment at older, cheaper prices, (rather like a wheat farmer who, having purchased unusually fertile land long ago, could now sell wheat at lower prices or earn higher profits than other farmers, even in a highly competitive wheat market). In a competitive market a firm that bought older, cheaper generating equipment would simply charge the higher prices that the cost of new, expensive equipment requires its competitors to charge and it would pocket the difference. In a regulated market, the regulator forces the "historically lucky" firm to pass along some, or all, of this potential extra profit to its customers in the form of lower prices. In either case, the "extra" profit resulting from lower costs is not a monopoly profit. It is called an "economic rent." And its presence is consistent with a perfectly competitive marketplace. See 2 Areeda & Turner p 510, at 33-34; P.A. Samuelson, Microeconomics, at 213-214, 635 (13th ed. 1989); G. Stigler, The Theory of Price, at 258-63 (4th ed.1987). Concord's study, which relates Edison's superior performance to cost advantages, is not evidence that Edison has market power.

47

Second, plaintiffs point to a provision in Edison's tariff that states that, if they wish to buy power from other firms, they must either (1) give Edison five years advance notice, or (2) compensate Edison for any adverse impact Edison suffers in the subsequent three years. We need not explain all the reasons why this provision does not show that Edison possesses monopoly power, however, because, at a minimum, the record does not show that the tariff either unreasonably, or significantly, limits a customer's ability to buy from others. To the contrary, the plaintiffs have bought power from non-Edison sources, namely the New York Power Authority, despite the provision. And the plaintiffs remain free to challenge before FERC any unreasonably high charge that Edison might try to assess under this provision.

48

Third, the plaintiffs argue that the kind of electricity supply service they need--a firm commitment by a generating source to supply their full requirements--was simply unavailable from any firm other than Edison; hence, Edison had "monopoly power" over the price of this particular kind of service. Plaintiffs offered three pieces of evidence to support this contention: (1) John O'Neil, superintendent of the Concord light plant, in response to the question, "Was there any other firm requirements power available for Concord to purchase from utilities other than Boston Edison?," answered, "Not to my knowledge, no, sir." (2) Maurice Berdan manager of the Wellesley light plant, in response to the question, "Is there any other source of firm requirements wholesale power available to the Town of Wellesley other than Boston Edison Company?", answered (over an objection that the answer "appears to call for an expert opinion and/or for hearsay"), "We had consultants look at it, and I have also had discussions with other people in the industry that we have not found any other sources of power, firm requirements power, that would be available ..." and it was "public knowledge in the industry that firm requirements power is not available." (3) The minutes of the March 10, 1987 Concord Municipal Light Board meeting report,

49

A question was raised as to whether or not Concord could become a customer of New England Power. It was stated that the 'NEES Plan' does not include new large power loads; Princeton's request to return to all-requirements status was rejected.

50

We do not think this evidence sufficient to show that Edison possessed "monopoly power" with respect to "firm requirements" contracts to supply electricity. For one thing, we do not see how a reasonable juror could conclude from this testimony and minutes that no firm other than Edison would sell "firm requirements" electricity at a price that roughly reflected current production costs. After all, the record is filled with evidence that other municipal distributors bought electricity from other suppliers. Both town witnesses were testifying in the context of a report (discussed at p. 30, above) that said it was less economical for the towns to buy electricity elsewhere than from a utility that had especially low historical costs. Neither town witness had himself tried to obtain a "firm requirements" contract for electricity (indeed, Mr. O'Neill conceded that he had "never talked to another utility besides Boston Edison on behalf of Concord"). And neither town witness made clear either the specific basis for his conclusion, nor, for that matter, did either witness say that he meant it was not possible to find an alternative supplier at prices reasonably close to current costs of production. To the contrary, David Foote, an Edison witness, who worked for a group of New Hampshire utilities, stated

that he had contacted three New England utilities (other than Edison) during the period in question and found that one of them would enter into a "firm requirements" contract to supply electricity.

51

For another thing, as we explained above, a simple showing that Edison offered a "firm requirements" contract (or any other type of contract) for power more cheaply than alternative suppliers does not show that Edison had monopoly power. Monopoly power is the power to raise price well above competitive levels before customers will turn elsewhere. See generally 2 Areeda & Turner pp 501-506, at 322-29. The plaintiffs' evidence shows neither that Edison's rates were significantly lower than the rates of others, nor that any such lower rates reflected anything other than lower costs; and the latter, as we explained earlier, is evidence only of an economic rent, not of monopoly power. See p. 30, supra. In the face of these economic facts, and other evidence in the record, the three pieces of evidence cited are simply too skimpy to support a "monopoly power" conclusion.

52

Finally, even though classical price regulation rests upon the premise that the regulated firm maintains a "natural monopoly," we resist any temptation to permit a jury to infer monopoly power from the mere fact of wholesale rate regulation itself. Whatever the historical basis for such regulation, today there is a growing conviction among students of the subject that the electricity generation business (as opposed to the transmission and distribution business) may be able to support several competing firms. See, e.g., P. Joskow, Regulatory Failure, Regulatory Reform and Structural Change in the Electric Power Industry in Brookings Papers: Microeconomics 1989 at 189-90 (1989); R. Pierce, A Proposal to Deregulate the Market for Bulk Power, 72 Va.L.Rev. 1183, 1212-15 (1986).

53

For these reasons, we do not believe plaintiffs met their burden of providing evidence from which a jury could find that Edison had significant monopoly power in the relevant market, i.e., the power to raise price significantly higher than the cost-based level that competition would otherwise bring about.

The judgment of the district court is

54

Reversed.

APPENDIX A

55

We offer the following highly simplified intuitive explanation of the "single monopoly profit" argument, based upon an example given by Professors Areeda and Turner.

56

Assume that copper is used only to produce pipe. Assume that the industry has two "levels:" (1) refining ore into copper ingot and (2) fabricating ingot into pipe. Assume the profit-maximizing price for copper pipe is $100. (Charging more would reduce pipe sales to the point where the monopolist would earn less profit; charging less would increase pipe sales, but not enough to make up for lost potential revenue on each sale.) Assume the cost (including a reasonable profit) of refining ore into ingot is $40. Assume the cost of fabricating ingot into pipe is $35.

57

Suppose that the "ingot" stage of the industry is monopolized but the "pipe" stage is competitive. An ingot monopolist would charge $65 for its ingot; it would rely upon competition among pipe fabricators to keep the mark-up at $35 and the fabricated-pipe price at $100; and it would pocket $25 ($65 minus $40) of monopoly profit.

58

Suppose, instead, that the "ingot" stage of the industry is competitive but the "pipe" stage is monopolized. Competition at the "ingot" stage would produce an ingot price of $40. The pipe monopolist would buy the ingot for $40; it would sell the fabricated pipe for $100; and (after subtracting the $40 cost of ingot and $35 cost of fabrication) it would pocket $25 of monopoly profit.

59

Now suppose a single, integrated firm managed to achieve a monopoly at both "levels" of the industry. It would sell pipe at the profit-maximizing price of $100. After deducting its total costs ($40 plus $35), it would pocket $25 of monopoly profit.

60

In each case, the profit-maximizing price remains $100 and the monopoly profit is $25. This normally will remain so as long as the manufacturing costs associated with ingot and pipe remain the same.

61

See 3 Areeda & Turner p 728b, at 199.

APPENDIX B

62

This is a graphic and a numerical example offered to help provide an intuitive understanding of the possible price effect of having two separate monopolists at two different industry levels.

63

NOTE: OPINION CONTAINS TABLE OR OTHER DATA THAT IS NOT VIEWABLE

64

Price	Quantity	Total revenue	M's costs	M's profits	R's costs	R's profits
10	1	10	1	9	6	4
9	2	18	2	16	12	6
8	3	24	3	21	18	6
7	4	28	4	24	24	4
6	5	30	5	25	30	0
5	6	30	6	24	36	[negative]
4	7	28	7	21	...	...
3	8	24	8	16	...	...

2	9	18	9	9	…	…
1	10	10	10	0	…	…

65

It costs $1 to make a widget.

66

A single monopolist ("M") will maximize his profit by setting a price of $6 and selling 5 widgets. His profit is $25 (represented by the area "rstu").

67

If M sells widgets for $6 each to an independent retailer ("R") with monopoly power, R will maximize his profit by setting a price of $8 and selling 3 widgets. His profit is 6 (represented by the area "rhgf").

68

M's best strategy is enter the retailing level of the industry and to set a final price of (roughly) $6.

- 915 F.2d.

Tuesday

May 22, 2012

Using Experiments to Inform the Privatization/Deregulation Movement in Electricity

Stephen J. Rassenti, Vernon L. Smith, and Bart J. Wilson

At the University of Arizona, electronic trading (now commonly known as e-commerce) in the experimental laboratory began in 1976 when Arlington Williams conducted the initial experiments testing the first electronic "double-auction" trading system, which he had programmed on the Plato operating system. The term "double auction" refers to the oral bid-ask sequential trading system used since the 19th century in stock and commodity trading on the organized exchanges. This system of trading has been used in economics experiments since the 1950s, and is extremely robust in yielding convergence to competitive equilibrium outcomes (Smith 1962, 1982a). Since information on what buyers are willing to pay, and sellers are willing to accept, is dispersed and strictly private in these experiments, the convergence results have been interpreted (Smith 1982b) as supporting F.A. Hayek's thesis "that the most significant fact about

Cato Journal, Vol. 21, No. 3 (Winter 2002). Copyright © Cato Institute. All rights reserved.

Stephen J. Rassenti is Professor of Economics at the Interdisciplinary Center for Economic Science (ICES) at George Mason University. Vernon L. Smith is Professor of Economics and Law, and Bart J. Wilson is Associate Professor of Economics, at ICES.

They acknowledge the influence and support of the many people and organizations who made possible the research program on which this paper is based: the Arizona Corporation Commission (Commissioners: Richard Kimball, Junius Hoffman, and Marianne Jennings) who in 1984 had the vision to fund our first efforts to study electricity deregulation; Penelope Brooke, who hosted Smith's tenure as a C.S. First Boston/Victoria University Visiting Fellow consultant on electric power reform in New Zealand, 1991; Prospect Electricity (now Integral Energy), and our host, John McQuarrie, in Rassenti and Smith's first visit to Australia as consultants, 1993; the National Grid Management Council (Australia), Hugh Outhred and John Kaye (NSW School of Electrical Engineering), who hosted our second visit, 1996; Hugh Outhred, who has continued to provide inspiration to us and to many other U.S. nationals in this country's deregulation debate, and specifically for his many valuable comments and corrections on an earlier draft of this paper. We also thank Lewis Evans of Victoria University, New Zealand, for providing us with a recent update of electricity restructuring in that country, and for his helpful comments on this paper.

this (price) system is the economy of knowledge with which it operates, or how little the individual participants need to know in order to be able to take the right action" (Hayek 1945: 526–27).

As with all first efforts at automation, the software developed by Williams allowed double-auction trading experiments that previously had kept manual records of oral bids, asks and trades, to be computerized.[1] That is, it facilitated real-time public display of participant messages, recording of data, and greater experimental control of a process defined by preexisting technology. It did not modify that technology in fundamental ways. This event unleashed a discovery process commonplace in the history of institutional change: the joining of a new technology to an incumbent institution causes entirely new, heretofore unimaginable institutions to be created spontaneously, as individuals are motivated to initiate procedural changes in the light of the new technology. Electronic exchange made it possible to vastly reduce transactions cost—the time and search costs required to match buyers and sellers, to negotiate trades, including agreements to supply transportation and other support services. More subtly it enabled this matching to occur on vastly more complicated message spaces, and allowed optimization and other processing algorithms to be applied to messages, facilitating efficient trades among agents that had been too costly to be consummated with older technologies. Moreover, resource allocation problems thought to require hierarchical command and control forms of coordination, as in regulated pipeline and electric power networks, became easily susceptible to self-regulation by entirely new decentralized pricing and property right regimes. Coordination economies in complex networks could be achieved at low transactions cost by independent agents, with dispersed information, integrated by a computerized market mechanism. This realization then laid the basis for a new class of experiments in which the laboratory is used to test-bed proposed new market mechanisms to enable a better understanding of how such mechanisms might function in the field, and to create a demonstration and training tool for potential participants and practitioners who become part of the "proving" process. Of course, once adopted, this modification and proving process continues in light of field experience.

We provide a short history of the application of the conception of

[1]Williams (1980) reports comparisons of the oral and electronic auctions. He found that oral auctions converged more rapidly for inexperienced subjects, but for experienced subjects (one previous session) the two systems were indistinguishable.

smart computer assisted markets to the design of electricity markets here and abroad.

The Privatization/Deregulation Movement in Electricity

We use the term "privatization" to describe generically the process of reform of foreign government command forms of organization of the electric industry. In *all* cases major components of the industry have *not* had their ownership transferred from public to private entities. Reform has focused on the use of decentralized spot and futures markets to provide price signals to improve the short and longer term management of the industry. The term "deregulation" applies to electricity reform in the United States, where 50 state and one federal regulatory body have regulated an industry already predominantly owned privately, but *not* decentralized except through recent reforms in some regional transmission systems that are still very much in transition.

The Arizona Utility Study

In 1984 the Arizona Corporation Commission (ACC) contracted with the University of Arizona experimental economics group to study alternatives to rate-of-return regulation of the utilities, with particular emphasis on electric power. The study consisted of two parts: incentive regulation (Cox and Isaac 1986) and deregulation (Rassenti and Smith 1986; also see Block, et al., 1985). Only the second part will be discussed here since this was the study that led to a long and continuing research program, encouraged by the privatization/decentralization movement abroad, with applications first in New Zealand, then Australia, and most recently in the United States.

Recommendations

The deregulation portion of the study led to many detailed recommendations that can be briefly summarized in the following key points (see Rassenti and Smith 1986):

1. The energy (generation) and wires (transmission and distribution) businesses would be separated, with generator plants (gencos) spun off from parent integrated utilities through the issuance of separate ownership shares to form independent companies.
2. An economic dispatch center (EDC) would be formed that would operate a computerized spot auction market for deter-

mining prices and allocations based upon hourly location (node) specific offer price schedules submitted by gencos. The spot market would be constituted so as to facilitate and incentivize the eventual inclusion of demand side bidding by discos (distribution companies and any other commercial and industrial bulk or wholesale buyers). Thus, ultimately and ideally, prices would be determined in an hourly two-sided auction in which discos would submit location specific bids to buy energy delivered to their location just as gencos would submit offers to inject energy at their respective locations on the grid.

3. Discos and transcos (transmission companies) would not be protected by exclusive franchise permits, and would be subjected to the price discipline of potential, if not actual, entry.
4. Important functions of existing institutions would be preserved but operate through a computerized spot market bidding mechanism based on decentralized ownership of gencos.

By "existing institutions" we referred to optimization—historically, computerized dispatch based on the engineering cost characteristics of generators and the network of integrated utilities—joint ownership by utilities of shared transmission capacity, and power pooling rules for security (spinning) reserves. In the proposed competitive reorganization, optimization algorithms would not be applied to production and transmission "cost" as in the regulated, hierarchical, integrated utility, but to the offer supply schedules and bid demand schedules submitted to the computer-dispatch center. The algorithms would maximize the gains from exchange (rather than minimize engineering cost as under regulation) in response to the real-time decisions of all buyers and sellers in the wholesale market. This specification was motivated by the recognition that (1) supply cost is subjective and measured by the willingness to accept payment for energy produced on location, and (2) demand is subjective and measured by the willingness to pay for delivered energy, where both types of information express the particular *real-time circumstances* of individuals. Coordination was a consequence of a new form of property rights: (1) rules for processing messages generated by decentralized agents themselves empowered by rights to choose offers and bids; (2) contingency rules for accepting offers and bids based on their merit order (higher bids and lower offers have priority in the rank ordering of bids and of offers), but importantly *qualified* by technical and security constraints that are essential if each agent is to bear the *true opportunity cost* that the agent imposes on all others.

The term "property rights" as we shall use it, provides a guarantee

allowing action within the guidelines defined by the right. Such guarantees are against arbitrary reprisal in that they restrict punitive strategies that can be levied against actions taken by the rights holder. Such guarantees provide only limited certainty of protection. Most specifically, property rights, as a guarantee allowing action, do not guarantee outcomes, since outcomes depend upon the property rights of others, and in electricity markets, as we shall see, upon global constraints affecting local outcomes that must be honored if the system is to be efficient, dynamically stable, and to incentivize the direction and level of capital investment.

Defining Competitively Ruled Property Rights to Unique "Monopolistic" Facilities

It was the ACC project that alerted us to the existence of "cotenancy contracts" for the joint ownership and operation of some large generation and transmission facilities. For us this was an illuminating empirical discovery, since this institution, that we modified with competitive property right rules, offered the potential to render the concept of natural monopoly null and void. Thus, suppose a city demand center can be adequately served by a unique physical facility such as a pipeline or transmission line. Under American-style regulation it is decreed that an exclusive franchise will be awarded to a single owner of the facility, whose price will be set so as to regulate the owner's rate of return on investment. Alternatively, in our proposed competitively ruled joint ownership property right regime it is decreed that (1) the facility must have two or more co-owners each having an agreed share of the rights to the capacity of the facility (In practice a common cotenancy contract rule is for each cotenant to receive capacity rights in proportion to his contribution to capital cost). Two additional competitive rules would allow (2) rights to be freely traded, leased or rented, and (3) new rights to be created by agreement to invest in capacity expansion by any subset of the co-owners, through unilateral action by any co-owner, or by outsiders if the existing owners resist expansion to meet increased demand. In historical practice cotenancy contracts had prohibited sale by individual rights holders without the consent of the other cotenants, and capacity expansion was allowed to occur only by joint agreement. The proposed new property rights structure creates multiple rights holders to compete in marketing downstream services utilizing the unique facility, and encourages new investment in response to increased demand. Subsequent to the ACC study, new research uncovered other examples of cotenancy contracts, a common one being the joint ownership of specialized print-

ing facilities by a consortium of newspapers in a city. Clearly, who prints the newspapers is a production issue potentially separable from the competition of newspapers for subscribers and advertising services. The courts repeatedly affirmed this principle when such cotenancy contracts attempted to include marketing and pricing conditions in what was ostensibly a shared production agreement (Reynolds 1990). Thus, our conception of a joint venture property right regime had already been well articulated in court cases involving newspapers. There was no new principle, only the question of how it might be reformulated for application to network industries.

This model of cotenancy as an instrument of competition was further elaborated in Smith (1988, 1993) and tested experimentally in the context of a natural gas pipeline network funded by the Federal Energy Regulatory Commission (Rassenti, Reynolds and Smith 1994). The model would also play a facilitative role in our consulting on privatization in New Zealand. But such discussions are far from culminating in a completed instrument, with many practical implementation difficulties remaining.[2]

Aftermath of the Arizona Study

By 1985 when the study report was filed and presentations made to the ACC, the political composition of the commission had altered, and the immediate impact of the recommendation for deregulation on Arizona policy was nil. By the time our final report was completed the commission was composed of new elected office holders and they considered our proposal to be impractical, idealistic and politically infeasible. Of course the commission's actions made the last claim a self-fulfilling truth. Unknown to us at the time, subsequent developments would reveal that this experience was a minor battle in a wider war for institutional change that would begin abroad but would ultimately spread to the United States, but with less success, we believe, than abroad.

Contrary to the position of the new commission, we considered our proposal eminently feasible in the electronic age, though in need of far more fundamental research, and resolved to undertake controlled experimental studies of various issues in the deregulation debate. Progress on this objective, however, was slow due to inadequate funding, and the fact that the cost of software development for the laboratory study of electronic trading in the context of electric networks

[2]Hugh Outhred (2001) notes that there is ongoing work in Australia under the NECA code-review process to explore practical implementations of network property rights (see www.neca.com.au).

was higher than for traditional forms of experimental research. Nevertheless, by 1987 we had conducted several pilot experiments in a six-node electric power network with three fixed inelastic nodal demand centers, and nine gencos (described in Rassenti and Smith 1986). The gencos, located at various nodes, submitted sealed offer price schedules each trading period to supply power over transmission lines whose energy losses were proportional to the square of energy injected. A valuable lesson from this unpublished research was the ease with which gencos could push up prices against inelastic demands by bulk buyers using a mechanism that did not permit demand-side bidding to implement consumer willingness to have deliveries interrupted conditional on price. This was our first brush with the important principle that competition is compromised in supply-side auctions in which buyers are passive and are unable through the mechanism to enter demand-side bid schedules. The California electricity market is now experiencing this principle in spades, but it was foreshadowed in the experience with privatization in England, and in other spot markets abroad and in the United States. We report experiments below that provide a rigorous demonstration that when the spot auction mechanism in common use around the world is supplemented by demand-side bidding it provides a property right regime that is a remarkably effective antitrust remedy.

Domestically, through the 1980s and into the 1990s, electric power would remain subject to American style rate-of-return regulation, while abroad government owned electric (and other) utilities were under political pressure to explore the use of markets for the management electrical energy allocations. Industry performance was seen as abysmal in the 1980s, causing countries such as Chile, the United Kingdom, and New Zealand to think the unthinkable: decentralization might be preferable to either government planning or direct regulation. But how might it be done?

How Experiments Were Used to Inform Privatization: New Zealand and Australia

Beginning in 1986 we initiated software development and a series of experiments to study mechanism design, industry structure, pricing, transmission and market power issues in electricity markets. (Rassenti and Smith 1986; Backerman, Rassenti, and Smith 1997; Backerman et al. 1997; Denton, Rassenti, and Smith 1998; Rassenti, Smith, and Wilson, 2000.) While this research was proceeding, one of the authors (Smith) consulted for the New Zealand Treasury in 1991

and two of us (Rassenti and Smith) in 1993, and also for Australia's Prospect Electricity in 1993 and National Grid Management Council in 1994. The impetus in New Zealand was our 1985 ACC report that fell unceremoniously on deaf ears in Arizona, but attracted attention abroad.

What Were the Questions?

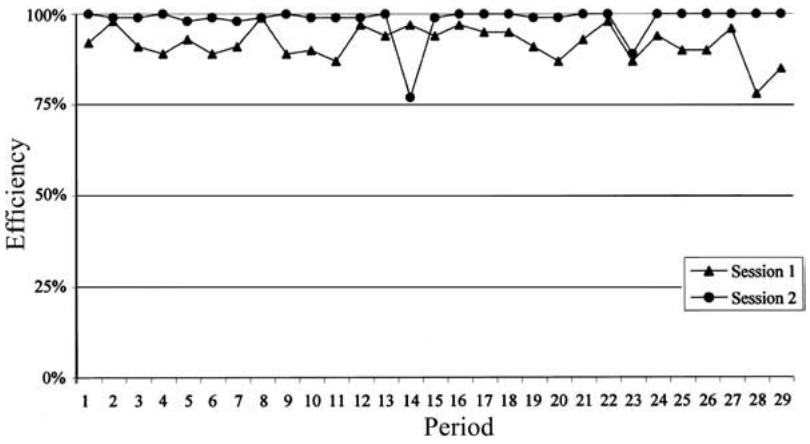
The two following research questions, addressed in laboratory electricity network experiments after 1986, and motivated by our ACC study, provided the primary information base for informing our contribution to the privatization process in electricity down under.

1. Is decentralization feasible and, if so, is it efficient to combine decentralized property rights in energy supply with a computer coordinated spot market and optimization schemes for dispatching generators?
2. How is the answer to question 1 affected by demand-side bidding?

Before the first experimental observations were made it was an open question whether it was feasible to replace engineering cost minimization in large integrated utility hierarchies with independent gencos submitting node-specific asking price schedules, bulk buyers submitting node-specific bid price schedules, and allocations determined by algorithms maximizing the gains from exchange implied by these marginal bid/ask schedules and the physical characteristics (loss characteristics and capacity constraints) of the grid. Engineers and managers to whom we made presentations were overwhelming skeptical—in fact were openly hostile—that such a system could be relied upon. ("You can't control electricity flows with markets—I know, because I'm an engineer.") The conventional wisdom of economists had been stated as follows:

> Generation and transmission are intimately and fundamentally related by the interconnections that the transmission system provides and the associated opportunities for area wide optimization . . . Because of these relationships, decisions either short-run or long-run, made at any point in a power system affect costs everywhere in the system. These effects raise potential externality problems. If a power system's components are owned by more than one firm, it is crucial for the efficiency of short-run and long-run decision making that all owners of parts of the system take into account all effects of their actions, not just the effects on the part of the system they own [Joskow and Schmalensee 1983: 63].

FIGURE 1
EFFICIENCY WITH EXPERIENCED SUBJECTS



Experimental markets, in which all energy sales and purchases were expressed as offers to sell and bids to buy so that allocations were determined simultaneously given the physical properties of the grid, demonstrated that energy market deregulation was eminently feasible. Furthermore, short-run efficiency was high—on the order of 90–100 percent of the maximum economic surplus, or gains from exchange were achieved in markets with very few participants. Figure 1 shows a plot of efficiency for two experimental sessions consisting of a series of 30 trading periods using experienced subjects in a 3-node radial network with 4 bulk buyers and 6 gencos (Backerman, Rassenti, and Smith 1997). Why are there no important efficiency losses due to short-run externalities? The answer resides in the condition that all allocations are determined simultaneously. Power loss on shared transmission lines varies as the square of total power injected. Therefore, genco A suffers higher costs of energy loss if genco B is using the same line. But if optimization is based upon every agent's marginal willingness to pay or to supply, with price and allocations determined simultaneously, then each agent bears the appropriate opportunity cost that his action imposes on all others at the margin. The problem is solved by the simultaneous submission of bid/ask schedules to which are applied algorithms for maximizing the implied gains from exchange taking account of system transmission losses.

But there are many other potential "external effects," besides shared system energy losses, that in principle are or can be internalized via mechanisms that link bid/ask schedules with system con-

straints through rule governed coordination: (i) voltage "constraints" (as they are so treated, technically, in all operating systems today), requiring "reactive power" to be produced, and therefore priced in the market if such constraints are to be incorporated into the market process;[3] (ii) intertemporal links on both the demand and generator sides of the market historically have implied the need for optimization over time, not just in the current spot market, but as shown by Kaye and Outhred (1989) and Kaye, Outhred, and Bannister (1990) the primary intertemporal coordination requirements can be met by forward markets; (iii) contingency provisions such as generator and transmission reserves to avoid blackouts from unscheduled equipment outages, and to avoid unstable cascades of outages that spread through the network.[4]

Turning to the second question, both regulation and government ownership have produced industries with a strong supply-side orientation. The politics of power yields a system in which (i) there are severe political repercussions if consumers "lose lights" too often, and (ii) consumers making decisions have no means of directly (or indirectly through wholesale markets) comparing the cost of new capacity with the cost of interruptions on peak or in emergencies. Consequently, adequate reserve capacity in generation and transmission requires supply-side investment sufficient to meet all demand, plus a large margin for security of supply. The regulatory and government-owned systems had no incentive to install technologies for relieving load stress by introducing time-of-demand pricing, and voluntary interruptible contracts for customers. For this to occur power users must have the real-time spot market capacity to either directly reduce

[3]Maintaining voltage to avoid "brownouts" requires generators, or special compensating devices, to provide local reactive power. Since generators can produce either reactive or active power (the latter is energy that does work) in variable proportions, (i) is a source of "externality" only if it is not priced, which is the universal practice inherited from centrally owned or regulated systems. We plan experimental designs to price reactive power as just another commodity.

[4]Generator (spinning) reserve can be supplied by a market for standby capacity in addition to the energy market. (See Olson, Rassenti, and Smith 2001 for an experimental study of such simultaneous markets). A simple such market (without network complications) is provided when you rent an automobile: if you use it you buy the gas in a separate energy market; if you do not use it then it is in standby reserve for contingent use. To maintain transmission reserves lines are typically constrained to carry much less than their thermal capacities by engineers whose zeal in minimizing the risk of losing a line, is not necessarily economical. A standard rule, based on $n-1$ analysis, is to set the capacity of each line in a network so that if any one line goes out the remaining $n-1$ lines can carry the peak load; if you want still more security $n-2$ analysis is applied and so on. Of course this approach begs the question of what price security. Can catastrophic insurance principles be applied with a variable premium that increases with monitored capacity utilization?

consumption in response to price increases, or indirectly by contract with the distributor to effect reduced deliveries in response to price increases. As we shall see below, the capability for interruption of energy flows must be expressible in the spot market if prices are to be adequately disciplined.

New Zealand

ESL's consulting work in New Zealand was directed entirely to questions of how a privatized NZ electrical industry, and a wholesale power market, might be structured. Intellectually, in the early 1980s, the sea change in issues of privatization versus government ownership and regulation was so drastic in the direction of economic liberalization that electricity reform seemed certain. The election of a new reform-committed Labour government was followed by a foreign exchange crisis the next day. All government enterprises had performed so poorly, and were such a drain on the Treasury that the country was soured on the "NZ (socialist) experiment." Everywhere in New Zealand, by 1991, were to be found people expressing the "user pays" principle as a slogan of reform.[5] This exuberance, strong in the late 1980s and early 1990s is now much abated, even reversed.

> New Zealand . . . retains large state-owned corporations that are suitable for privatization, but . . . its privatization activity has been muted for much of the 1990s. This decline reflects political perceptions of the privatization act as well as the resolution of property right issues, some of which arise from considerations of industry structure that is suitable for light-handed regulation, and some from the potential settlement of Maori claims on the crown [Evans 1998: 3].

ESL consulting for the New Zealand Treasury in 1991, and later for Transpower, NZ in 1993, created as the state-owned enterprise that maintained and operated the high voltage grid, emphasized privatizing transmission, transmission pricing, and demand-side bidding.

Privatizing Transmission

What might be the incentive and ownership structure that should be implemented for the New Zealand grid, and for the market dis-

[5]The impetus for reform was a drastic reduction in the performance of the NZ economy from 1953 to the late 1970s. New Zealand had the world's third-highest per capita income in 1953 (behind the United States and Canada but tied with Switzerland) and by 1978 had slipped to twenty-second (less than half the per capita income of Switzerland). See McMillan (1998).

patch center that would determine allocations of energy supply among decentralized generator owners who bid into the spot market?

Our recommendations had their genesis in our 1985 ACC study of cotenancy, but the basic idea—a cotenacy property right system—was substantially extended and tailored to fit the special physical properties of electric power flows in interconnected alternating current (AC) networks. Primarily these properties are twofold: (a) flows on individual links in the network cannot be precisely controlled because in AC networks there has not existed anything analogous to the valves on links in fluid and gas pipeline networks; (b) optimization in such networks requires knowledge of willingness-to-pay bid demand values at delivery nodes, offer supply terms at power injection nodes, and the physical properties (loss characteristics and capacity constraints) of all elements of the network. One can then solve simultaneously for the pattern of energy injections and deliveries that satisfy all demands and constraints while maximizing the short-run gains from exchange based on all such information. These two characteristics combined imply that it is not possible to specify well-defined path rights from any source node to and delivery node. The flow on a given path may be optimal at one time, but with a change in the supply and demand pattern, and with different transmission constraints binding, the flow on that path may be much different, even reversed at another time.

We proposed that these characteristics of the electricity industry be supported by a property right regime with the following commensurate features when the system is privatized as a joint (competitively ruled) venture, or cotenancy, owned by all users.

(a) At each energy injection node is connected a set of generators with some specified capacity that has occurred in history up to the time of privatization. That capacity is assumed to reflect the benefits, based on historical utilization rates, and site value of locating the capacity where it resides.

(b) Similarly, each delivery node will have associated with it a capacity to withdraw power.

(c) Rights to inject (or withdraw) power at each node can then be defined and certificated in capacity terms based on historical investment.

(d) Each generator has the right to submit a bid supply schedule indicating the various quantities the supplier is willing to inject at corresponding stated asking prices, where the schedule is restricted not to exceed a total offer of that generator's capacity rights at its connection node. How much of this offer is accepted by the dispatch center, depends on the offer terms of competing suppliers at the

same or other nodes, the nodal pattern of demand, and the physical properties of the grid at any time. Stability, security and voltage considerations may require certain key generator offers to be accepted in exception to the general merit order rule that the lowest priced generators have priority over higher priced ones. Such key generators are likely to change with the network load configuration. Thus, each generator merely has a right to offer up to its capacity in units of power, not the right for the offer to be accepted. Such uncertainties are inherent in the nature of the system, and property rights must reflect these contingencies. Technological and institutional innovation may alleviate exposure to these risks, and such developments must be allowed, and have an incentive, to happen.

(e) These capacity rights can be freely traded, leased or rented to others subject only to contract laws applicable to any industry; but as in other industries, electricity may leave its own footprints on the form of those contracts.

(f) Any individual user in this structure, or any group of users forming a consortium, is free to invest in increasing the capacity of any line or lines in the system. Those making the investment will acquire rights, as in (c)–(e) above, to any increase in capacity at individual nodes that is made possible by the investment. Any such increases in capacity will be uncertain, and based on imperfect engineering simulations that are commonly used to evaluate and site capacity expansions.

(g) Finally, since incumbent users may not be well motivated to expand capacity, the cotenants cannot prevent the entry of new investors who invest in line capacity expansion, and acquire rights to the consequent increase in nodal rights to inject (or withdraw) power.

Transmission Pricing

Given the joint ownership structure indicated above, all users share output-invariant operating and maintenance costs in proportion to their respective capacity rights. The primary variable cost of transmission is the energy lost in the transfer of power from source nodes to delivery nodes. This loss (per mile of line) in high voltage lines varies approximately as the square of energy injected—less energy is received than is sent. Hence, if the average loss per unit is A (usually a number between .02 and .2) for a given line, the marginal loss is $M = 2A$. This implies that if the price at an upstream injection node is P, then at any downstream node the price is $P' = P + PM$, i.e. the delivery price is the price at the injection point plus the marginal cost of energy lost in delivery. Note that PM is the true opportunity cost

of energy lost in transmission, and all buyers served by remote generators must pay this cost in an efficient energy supply network. On long lines where the average loss at peak demand can be up to 20 percent ($A = .2$), the nodal price difference, $P' - P = 2AP$ can be up to 40 percent of the delivered price.[6]

Demand-Side Bidding

Competition is greatly enhanced if wholesale buyers can bid into the spot market using discretionary demand steps that define price levels above which they are prepared to interrupt corresponding blocks of power consumed. As we shall see, demand-side bidding also reduces price spikes on peak. Moreover, interruptible flows can substitute for security reserves of generation capacity, while reducing the prospect that transmission lines will become constrained.

New Zealand deliberations on structuring the grid continue. However, the functions of the spot market, called the New Zealand Electricity Market (NZEM), have been structured as a ruled-governed joint venture. (For a detailed report see, Arnold and Evans, 2001; also see NZEM 1999). Only three countries have implemented policies requiring the grid users to fund investment expansions: Chile, Peru, and Argentina. In all three cases, however, the multiple owners operate under regulated prices (Kleindorfer 1998: 69). Thus, no country has implemented a completely privatized grid regulated only by property rights, nor is this likely to be achieved in the near future.

Although our fledgling proposals for structuring joint ownership of the grid have not been implemented, and indeed require a lot more intensive work to be operational, the New Zealand spot market implements both the marginal loss pricing of transmission and demand-side bidding. It is important, however, to note that nodal energy pricing in New Zealand does not provide ex ante real-time prices that can be avoided by action of buyers and sellers in the current period. Prices are an ex post cost recovery and distribution scheme, and effect decisions only insofar as events/conditions are repeated and anticipated by decision makers. The same is true for the systems implemented in California and the Middle Atlantic regions in the United States. This is partly the result of industry traditions in which people

[6]As a practical matter, because of the cost of metering and monitoring, network pricing always involves a certain amount of aggregation of subsystems into representative nodes or paths. Hence, the above principles are indeed conceptual, and only imperfectly captured in any actual operating system. Moreover, low voltage distribution systems do not follow the square loss law rule at all well, and losses are commonly averaged across the high density of users.

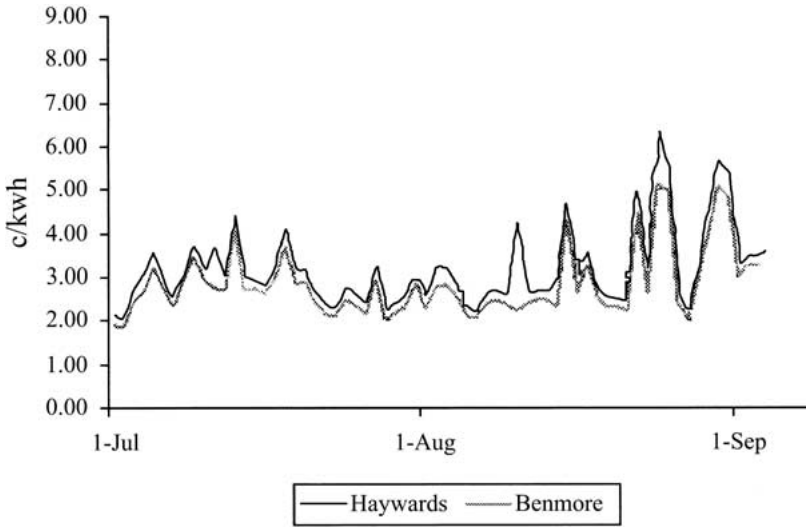
think of prices as cost recovery devices rather than signals of avoidable opportunity costs, and partly a consequence of implementing the appropriate technology and institutional arrangements. New Zealand, however, is moving to implement true avoidable cost pricing as used now in Australia (see below).

Marginal cost pricing of transmission is politically very difficult to implement in democratic regimes—three other countries (Chile, Peru, and Australia) have adopted it (Kleindorfer 1998: 69). Strong political pressures favor averaging transmission losses across all customers. This creates an incorrectly priced external effect that is avoidable by appropriate specification of property right rules, and illustrates one of the many externality problems *created, not solved* by collective action. With minor exceptions averaging losses over all customers was the universal practice in both state owned and American style regulatory regimes, and this practice dies very hard. People do not understand the opportunity cost/efficiency principle here: each agent pays the cost that his consumption imposes on others, thereby eliminating external effects. But collective agreement is necessary to implement the application of this principle to grid pricing. (Note that the principle creates no problem in the airline or accommodation industries, where on-peak prices emerge spontaneously in competition, à la Hayek's 1945 perceptive argument, and collective agreements are not needed. This illustrates one of the many hazards in decentralizing interdependent network industries using some collective agreement process.)

Most of the New Zealand population and electricity demand is on the North Island, while most of the generation capacity is on the South Island. It is some 900 miles from the bottom of the South Island, where the most remote generators are sited, to the top of the North Island, where the largest concentration of population is located (Auckland). Consequently, at peak demand, with no constrained lines causing a further price difference due to congestion, there is a price difference of approximately 33 percent between the two most remotely separated nodes. Figure 2 provides a chart of New Zealand electricity prices at the inter-island link, Haywards and Benmore in the South (not at the two extreme nodes), for the winter months of July and August, when the heating demand for energy is greatest.

Relevant to demand-side bidding the New Zealand Electric Market (NZEM) rules specify that "Each trading day, each Purchaser Class Market Participant will submit to the Scheduler the bids pursuant to which . . . (that Participant) . . . is prepared to purchase Electricity from the Clearing Manager for each trading period of the following trading day" (NZEM 1999: B.2.1). Such bids specify the relevant

FIGURE 2
AVERAGE ENERGY PRICES IN NEW ZEALAND, 2000



SOURCE: Chart was downloaded from www.m-co.co.nz/C2dPricesMonth/000905.htm.

trading periods, the grid exit node, must represent reasonable endeavors to predict demand, and specify up to 10 prices (price steps or "bands") and corresponding quantities. There are no upper or lower limits on prices "The highest price band for each bid will be deemed to start at a quantity of zero" (NZEM 1999: B.2.3). Note that this provision defines the strike price where the Marshallian bid demand schedule intersects the price axis. Since the technology for interrupting flows is limited, these provisions of the NZEM are currently little used (as reported to us in private conversation with Lewis Evans at Victoria University, NZ), but the institutional stage is set for more extensive demand-side bidding as the appropriate technology becomes more available and cheaper. They will become more significant when New Zealand implements real-time pricing.

Australia

We were invited to visit Australia in 1993 by Prospect Electricity (now part of Integral Energy) in New South Wales, the second largest distribution company in that state. Australia, unlike New Zealand

(initially), was not committed to privatizing electricity, although the political debate had begun. Rather, the commitment was to decentralization, setting up a national wholesale market. This was the charge of the National Grid Management Council (NGMC). (Privatization if it occurred was the providence of the states, which were the owners of existing power system assets. All generation, transmission, and distribution systems remain publicly owned even today, with the exception of Victoria where all are privately owned, while South Australia has executed 200 year leases of its assets to private entities.)[7] It was during this visit that we learned that the constituency for privatization was made up of bulk buyers—commercial, industrial, and distribution companies—who expressed the belief that the state government-owned electricity industries were producing power at exorbitant cost, and this was hampering the ability of Australia's energy intensive industries to compete in world markets. Primarily our sponsors consisted of the buyer side of the industry, and our task was to supply market information and deliver demonstration technology: give lectures, seminars and conduct experimental workshops with a wide spectrum of industry and government representatives who would participate in our prototype wholesale electricity experiments, demonstrating feasibility, efficiency, and possible structural features for a decentralized wholesale market, with the extent and form of decentralization yet to be determined. These lectures and workshops were well attended, but with understandably more enthusiasm coming from the demand side than the supply side. Such was the political environment as we saw it.

Subsequently, the central government created the National Grid Management Council to plan and oversee a wholesale energy market embracing the states, integrated by a national interconnected grid. This led to a controversial "paper trial" (cost, $2 million) in which participants walked through proposed procedures for bidding and clearing in a spot energy market. Our Australian contacts pressed, and won, approval to conduct laboratory experiments with a prototype for the proposed market. We were consultants on software specifications, and experimental design, but with all development and experiments to be conducted in Australia. This ultimately led to a two-week (7 hours per day) electronic trading experiment using nonindustry participants trained in the exchange procedures, and earning significant cash profits based on induced costs, and demands, and on Australian parameters and grid characteristics. We advised against using any

[7]Based on private correspondence with Hugh Outhred.

industry participants because of their known political biases for or against the impending market reforms.

On December 13, 1998, the National Electricity Market began trading Australian electricity. Prior to that period separate markets traded power in the States of Victoria and New South Wales as early as 1996.

In summary, experimental methods in economics served to facilitate the development of a wholesale electricity market in Australia in the following ways:

1. It provided a pre-1991 experimental database demonstrating the feasibility of using a smart market, price signals to coordinate production and transmission over huge geographical areas, and to help inform the political decision process.
2. Treatment results from specific experimental designs suggested that overall market efficiency, price volatility and the distribution of surplus among the buyers, sellers and the transmission system were significantly impacted by the following: transmission and auction market pricing rules, whether or not there was demand-side bidding, and whether or not transmission line constraints were binding.
3. As noted in communication with Hugh Outhred, the new experiments "at UNSW also demonstrated the importance of forward markets in containing market power" (see Outhred and Kaye 1996).
4. It provided hands-on experience and training for managers and technical staff, and alerted the principal agents involved in the wholesale market to some of the potential design issues in the process.
5. It enabled the Australians to go through the process of market prototype software development, to conduct experiments using Australian grid and generator cost parameters, and to learn much more about how their proposed market system might work prior to actual trading in Victoria and New South Wales.

The wholesale market in Australia has implemented features that make it among the most advanced anywhere from the perspective of reflecting good economic design principles, although it is important to emphasize that those principles are under ongoing review and modification in the light of changing experience and technology. We mention two features central to the issues discussed above that were in the National Electric Code *prior* to their experiments (quoted from personal correspondence with Hugh Outhred, February 2, 2001):

(a) "Network pricing in Australia does incorporate marginal network losses in the following manner: the 'notional interconnectors' between regions include . . . (adjustment for) . . . marginal losses . . . directly into the process for setting five-minute prices; inter-regional transmission loss factors are set annually on the basis of average marginal network losses (the averaging period may be shortened at some future time) . . ." Hence, the loss factors, as such, are not based on current real-time conditions, as are the flows to which the factors are applied.

(b) "The Australian National Electric Market Rules (NEM) . . . (also) . . . incorporate the demand side—both formally as bids . . . and informally as price elasticity. The latter option exists because: half-hourly prices are forecast at least 24 hours ahead and broadcast to all market participants (supply and demand side); participants can change their bids and offers from the time of their original submission (one day ahead) down to the half hour to which they apply; the actual spot price is set in 'real-time' and broadcast to all participants—a consumer can simply reduce demand in response to that price signal and thus avoid paying the price. That facility is now being used in practice, both by a consumer participating directly in the NEM and by retailers backed up by discretionary demand reduction contracts with final consumers." It is evident, however, that "much more development (is) needed" [Outhred 2001: 20].

The United States

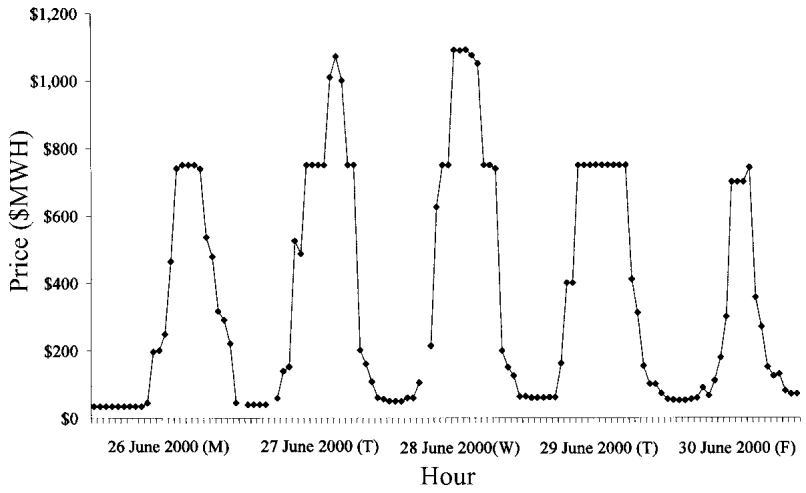
The deregulation of electricity did not impact the United States until privatization/decentralization reform was well advanced abroad. Viewed from the perspective of those of us interested in market design for deregulation, the U.S. experience has been disappointing, and the design details heavily politicized. At the start, the industry strongly opposed deregulation. Nothing new here, as the same was predominantly true for airline, gas, railroad, and trucking deregulation. But with electricity there was the need for state or regional *collective agreement* on how the industry would be restructured, and what rules would govern market operation since there was clear need for computer coordination of generator loads to meet instantaneous demand on highly interconnected networks. (No need for such agreement in the deregulated airline industry. The routes no longer had to be certificated, the industry was regulated by free entry and exit, and what emerged spontaneously in response to the demand for frequent low-cost service was the hub-and-spoke structure that was anticipated and deliberately planned by no one.) Originally, for example circa 1985 when we finished our ACC report, the industry had argued that deregulation was not technically feasible, but that proposition had

been shot down all over the world by decentralization programs none of which had followed American style rate of return regulation. There were various forms of "light-handed" regulation such as price caps on charges for the "wires" business—high voltage transmission or local low voltage distribution—but energy was being priced competitively limited only by technology and the state of learning. No one abroad wanted to use the American model, which was perceived to be broken just as badly as the state owned or dominated models that were being reformed.

In this environment, once the writing was on the wall, the utilities focused not on questions of market design and efficient spot markets, but on lobbying for fixed new monthly charges to cover their alleged "stranded costs." This was price design for revenue protection not *market* design for efficiency. Most economists seemed to accept the need for such compensation, either because it was "fair" for utilities to recover the cost of investments made in good faith under a regulatory regime that was being replaced (Baumol and Sidak 1995), or because it was considered the political price to be paid for utility support for deregulation (Block and Leonard 1998). Since the utilities were already privately owned, had long engaged in bilateral economy energy exchanges, and energy marketers, or intermediaries, had emerged to facilitate such contracts, there was opposition to the very idea of an open spot market. Bilateral interests wanted to report only origin and destination flows to schedulers, with prices remaining proprietary. Ironically, the bilateral special interest groups had been fostered by legislation intended to move the industry toward market liberalization: the Public Utility Regulatory Policies Act of 1978, and the Energy Policy Act of 1992. These initiatives were designed to facilitate transmission access by independent power producers as a step toward fostering the development of wholesale power markets. (Bear in mind that such access was being opposed by some utilities, and federal action was seen as necessary). The bilateral trading model was promoted, partly because of its perceived success in reforming the gas industry, but also because gas marketing intermediaries wanted to expand into electrical energy markets. California followed the bilateral model in restructuring electricity. We long regarded this model as misguided: bilateral bargaining in the electronic age could not provide the foundation for an efficient market model of interdependent (pipeline or transmission) networks.[8] California, however,

[8]For a critique of this trend see Smith (1987, 1996), and for studies of smart computer assisted markets in gas pipeline networks see McCabe, Rassenti, and Smith (1989, 1990), and Rassenti, Reynolds, and Smith (1994).

FIGURE 3
CALIFORNIA PX PRICES



did require the demand of the Investor owner Utilities to be processed through the CalPX, but these demand quantity bids were "at market" (pay whatever is the supply-side asking price that clears the market); they were not price contingent bids implemented by interruptible service contracts.

Thus, in California and elsewhere, the new "wires" utilities succeeded in instituting new fixed monthly charges to cover their stranded costs, and fixed per unit energy charges for retail customers, but no one was preparing for and investing in the technology for demand-side bidding as an instrument to discipline prices in the hourly spot market and to provide incentives for users to reduce demand or switch their time-of-day consumption from higher to lower cost periods. Imagine what would be the consequences to the airlines, and all of their passengers, if, in order to be licensed, airlines were required to charge all passengers an identical regulated monthly access fee and a fixed price per mile traveled, *independent of flight destination, time of day, time of week, season or holidays, and independent of the flier's willingness to pay*!

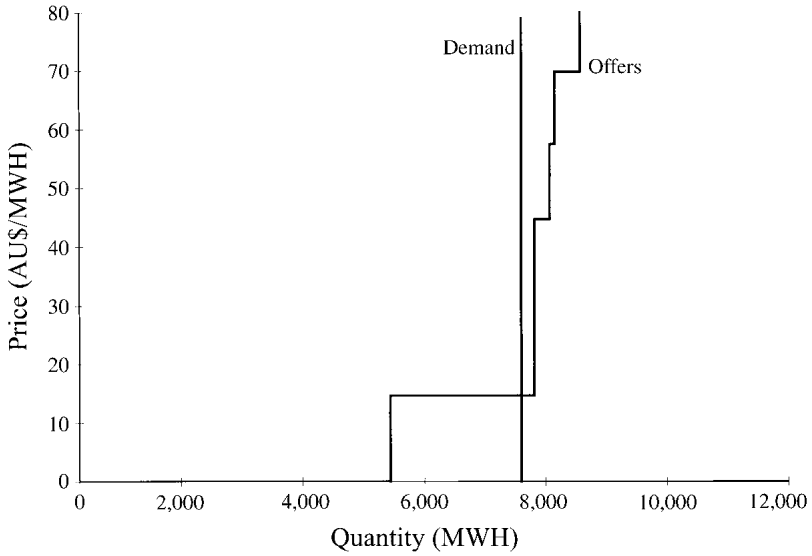
Figure 3 illustrates a typical 24-hour period of price variation on the California PX (their open spot market exchange). Since most of the power was either traded via bilateral contracts at secret prices, not part of the spot market, or through the PX as bids "at market," demand was not price responsive. Observe in Figure 3 that the peak

demand and most of the "shoulder" transition demand (between peak and off peak) are at prices above 10 cents per kilowatt ($100 per megawatt), and are therefore far in excess of what local distributors collect from their residential customers. There are numerous other examples of on-peak price spikes of up to 10 or more times the normal energy prices (in the $25–$30 per megawatt range). (See the Bloomberg Daily Power Report, online, Summer 1999 for a report on sharp price spikes in the Midwest and South.) These price differences imply an enormous rate of return on investment in contracts for voluntary selective interruption of energy deliveries, with gains shared by both the distributor and its customers.

Demand-Side Bidding Controls Market Power and Price Spikes

Earlier experimental market research, cited above, used demand-side bidding, and we observed very competitive results. New experiments study this issue much more systematically in the design reported by Rassenti, Smith, and Wilson (2000) comparing prices with and without demand-side bidding. Bulk buyers submit discretionary bid steps reflecting the prices above which they are prepared to reduce demand by invoking their contracts for interrupting deliveries. It is important in a competitive electricity market that bulk energy providers contract for *discretionary* interruption of (suitably compensated) consumers. Why? Because then their bids in the wholesale market cannot be known with certainty by the supply-side bidders, and demand-side bidding can better deter supply-side market power. The problem created by inadequate price responsive demand in a supply-side dominated auction can be illustrated with the chart shown in Figure 4, due to Outhred. In such a market, the clearing price is sensitive to the asking prices submitted by peaking generators in short supply, especially near peaks in demand. Thus, in Figure 4, the price is $15 per MW with demand 7,700 MW, but if demand had been 8,000 MW, the spot price would have been $45 per MW, and at a demand level of 9,000 MW, the price would have been indeterminate forcing the dispatch center to use security reserves or to involuntarily interrupt customers. Unquestionably, many consumers would have been prepared to reduce demand to avoid such a price spike, provided that they had been given the opportunity and incentives commensurate with the savings. In the United States are such conditions to be judged a problem in supply-side market power, or an institutional and incentive failure of the market mechanism to implement responsive demand? The tendency is to blame market power although in another industry—hotel/motel accommodations, or airline

FIGURE 4
PRICE DETERMINATION IN THE AUSTRALIAN ELECTRICITY MARKET

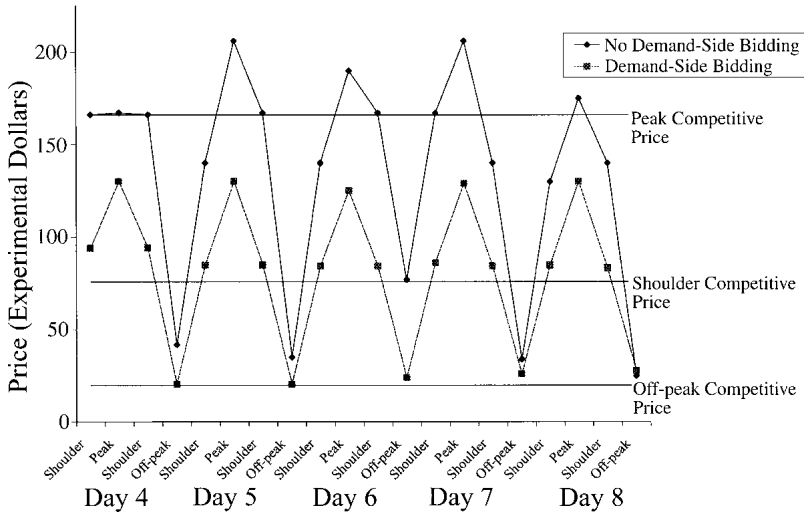


SOURCE: This graph is drawn from Hugh Outhred's presentation entitled "Australia: Spot Trading Results and Implications for Ancillary Services," 5 January 2000. The data are for the 17 May 1996, targeting 2000.

seat pricing, where the product also is nonstorable—demand is strongly responsive to time variable competitive prices.

Figure 5 plots experimental data comparing prices with and without demand-side bidding over the course of 5 "days" of trading. Each day in an experiment consists of a cycle of four demand pricing periods: shoulder, peak, shoulder, and off peak. Hence, the experiments consolidate the shoulder transitions, peak, and off-peak hours (shown in Figure 3) into four simpler time blocks for auction price determination. Note that when there is no demand-side bidding, prices are much increased, well in excess of the controlled experimental competitive prices, especially on the shoulder and peak demand periods. Both generator "market power" and upward price spikes are effectively controlled by the introduction of demand-side bidding leaving all other features of the market unchanged. In these experiments a very modest proportion (16 percent) of peak demand is interruptible by wholesale buyers; most of the on peak demand (84 percent) is what the industry calls firm or "must serve" demand.

FIGURE 5
AN EXAMPLE OF THE EFFECT OF DEMAND-SIDE BIDDING

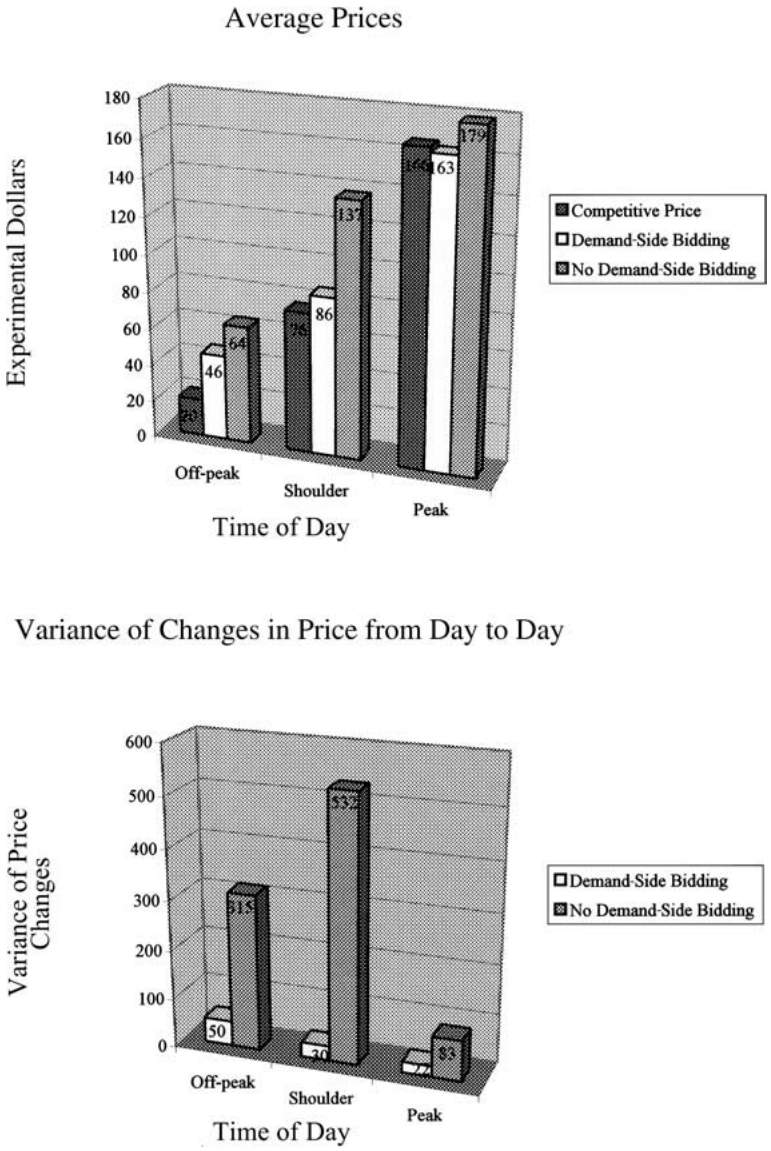


The chart in Figure 5 plots the data from just one of four independent experimental comparisons reported in Rassenti, Smith and Wilson (2000). Figure 6 provides a bar graph summarizing all of the experimental results. With demand-side bidding the average level of prices is reduced in all segments of the daily demand cycle, while the great variability in price changes is nearly eliminated.

Implications for Electricity Deregulation in the United States

The computerization of laboratory market experiments using profit-motivated human subjects in the 1970s unexpectedly revolutionized our thinking about the purpose and uses of experiments. In particular we soon came to recognize that the laboratory could be used to test-bed new electronic trading systems for application to industries traditionally perceived as requiring hierarchical organization and government regulation to achieve proper coordination and control over the resulting legally franchised monopolies. Electricity was a prime example, and we attempted to use our first experience with what we called "smart computer assisted markets" to inform Arizona's cautious and tentative interest in restructuring its electrical industry to rely on markets to regulate the energy segment of the industry. Failing at the time to influence policy, our effort was not

FIGURE 6
PRICES AND VOLATILITY WITH AND WITHOUT DEMAND-SIDE BIDDING



ignored abroad, and we participated as consultants in developing proposals and the use of experiments to help inform some of the key research issues in decentralization, and to serve as a hands-on training tool for those managing the transition. Decentralization required the creation of new property rights: a governance structure and efficient pricing for the grid, generator entry and exit rules, market rules governing messages and contracts in the context of computer controlled coordination, optimization, and communication, but with all outcomes driven by the decisions of dispersed agents whose circumstances of time and place were reflected in market bids to buy or offers to sell.

In the United States the industry was already privatized, but subject to centralized state and national price regulation based on a "fair" return on investment. With the proposed deregulation of electric utility prices and consumption each state or region needed to develop a plan for restructuring their industry and specifying the auction market rules for determining the real-time wholesale price of energy. Without exception, the resulting market designs, hammered out by regulators, consultants, industry representatives, and various power-marketing intermediaries, all employed supply-side bidding mechanisms for the hourly spot market. These spot markets were supplemented with wide ranging freedom for power users, producers, and intermediaries to engage in a variety of bilateral contracts outside of direct price discipline by the spot market. For the spot market this supply-side emphasis meant that any user, regardless of the individual circumstances of that consumer's need for an uninterruptible flow of energy, would be guaranteed that this demand would be served. Bilateral contractors could agree to allow various degrees of firmness of demand to impinge on contract terms. But in this longer-term contract market prices are negotiated and secret, and are not subject to the direct real-time opportunity cost constraints provided by the spot market.

The "must serve" demand policy in the spot market was inherited from a rigid regulatory regime that politicized the reliability of electricity flows to all consumers, whatever the cost. This cost was collectivized by averaging it across all users regardless of individual consumer differences in willingness-to-pay for keeping the lights on. The local utility was expected to maintain service, or restore it quickly, even in inclement weather, spreading the cost of this super-reliability thinly over all customers. This cost included the maintenance of substantial reserves in generation and transmission capacity. Thus system reliability and the capacity to satisfy all retail demand were exclusively a supply-side adjustment problem. In providing this superior service

to all, the supply side was always justified in claiming 100 percent cost recovery plus a fair profit. The consequence of this supply-side mindset was uncontrolled cost creep that increased to a gallop and ultimately became part of the political outcry for deregulation. Implicitly, however, the process of deregulation assumed that this built-in supply-side bias did not require fundamental rethinking when it came time to design spot markets for the new world of competition. As always in market institutions, the devil was in the details.

Beginning three years ago in Midwestern and Eastern markets peak prices hit short-run levels of 100 or more times the normal price level of $20–$30 per megawatt hour. This was the predictable direct consequence of *completely unresponsive spot demand impinging on responsive discretionary (bid) supply.* More recently the California spot market has been plagued by exorbitant increases in prices as illustrated in Figure 3. This has led to political action to impose price caps on this market, which, of course, can only discourage a positive supply response to the shortages. The move to replace American-style regulation with what may become known as American-style deregulation is in danger of being derailed by these interventions.

Controlled comparisons between markets with and without demand-side bidding, in which only 16 percent of peak demand can be voluntarily interrupted, show that the effect of demand-side bidding can dramatically lower both the level of prices and their volatility.

The public policy implications are evident: wholesale spot markets need to be strengthened institutionally by making explicit provision for demand-side bidding. Distributors need to incentivize more of their customers to accept contracts for voluntary power interruptions, or use time of day meters and load control systems to manage their own price responsivity. Industrial and commercial buyers who already have the capacity to handle interruptible energy supply, but who contract outside the spot market need adequate incentives to participate in the spot market where their more responsive demands can impact public prices. Distributors stand to gain by interrupting demand sufficiently to avoid paying higher peak and shoulder spot prices, and these savings can be used to pass on incentive discounts to customers whose demand, or portions of it, can be reduced or delayed to off-peak periods when supply capacity is ample. In California, news reports indicate that distributors have lost some $10 billion buying high (Figure 3) and selling at vastly lower residential rates.

The technology and capacity for implementing such a policy already exists and can be expanded. This policy recognizes that adjustment to the daily, weekly, and seasonal variation in demand, and to the need to provide adequate security reserves, is as much a demand-

side problem as it is a supply-side problem. The history of regulation has created an institutional environment that sees such adjustment as exclusively a supply responsibility, and views prices as an ex post means of cost recovery. The result is an inefficient, costly and inflexible system that has produced the recent price shocks and involuntary disruption of energy flows. Demand-side bidding and price feedback coupled with the supporting interruptible-service incentive contracts can eliminate unjustified price volatility, price increases and reduce the need for reserve supplies of generator and transmission capacity.

References

Arnold, T., and Evans, L. (2001) "The Law and Economics Basis of Enforcement of Governance of Private Joint Venture Networks: The Case NZEM." (Forthcoming in *The Antitrust Bulletin.)*

Backerman, S.; Rassenti, S.; and Smith, V. (1997) "Efficiency and Income Shares in High Demand Energy Network: Who Receives the Congestion Rents When a Line Is Constrained?" (Forthcomng in *Pacific Economic Review*.)

Backerman, S.; Denton, M.; Rassenti, S.; and Smith, V. (1997) "Market Power in a Deregulated Electrical Industry." (Forthcoming in *Journal of Decision Support Systems*.)

Baumol, W., and Sidak, J. G. (1995) *Transmission Pricing and Stranded Costs in the Electric Power Industry.* Washington: AEI Press.

Block, M.; Cox, J.; Isaac, R. M.; Pingry, D.; Rassenti, S.; and Smith, V. (1985) "Alternatives to Rate of Return Regulation." Final Report of the Economic Science Laboratory, University of Arizona, to the Arizona Corporation Commission (15 February).

Block, M., and Lenard, T. (1998) *Deregulating Electricity: The Federal Role*. Washington: Progress and Freedom Foundation.

Cox, J., and Isaac, R.M. (1986) "Incentive Regulation." In S. Moriarity (ed.) *Laboratory Market Research*, 121–45. Norman, Okla.: University of Oklahoma Press.

Denton, M.; Rassenti, S.; and Smith, V. (1998) "Spot Market Mechanism Design and Competitivity Issues in Electric Power." *Proceedings of the 31st International Conference on System Sciences: Restructuring the Electric Power Industry.* (Forthcoming in *Journal of Economic Behavior and Organization*.)

Evans, L. (1998) "The Theory and Practice of Privatization." Working Paper, Victoria University of Wellington (August).

Hayek, F. A. (1945) "The Use of Knowledge in Society." *American Economic Review* 35: 519–30.

Joskow, P., and Schmalensee, R. (1983) *Markets for Power.* Cambridge: MIT Press.

Kaye, R. J., and Outhred, H. (1989) "A Theory of Electricity Tariff Design for Optimal Operation and Investment." *IEEE Transactions on Power Systems* 4(2): 46–52.

Kaye, R. J.; Outhred, H.; and Bannister, C. H. (1990) "Forward Contracts for the Operation of an Electricity Industry under Spot Pricing." *IEEE Transactions on Power Systems* 5(1): 606–13.

Kleindorfer, P. (1998) "Ownership Structure, Contracting and Regulation of Transmission Services Providers." In H. Chao and H. Huntington (eds.) *Designing Competitive Electricity Markets.* Boston: Kluwer.

McCabe, K. S., and Smith, V. (1989) "Designing 'Smart' Computer Assisted Markets in an Experimental Auction for Gas Networks." *European Journal of Political Economy* 5: 259–83.

McCabe, K.; Rassenti, S.; and Smith, V. (1989) "Markets, Competition, and Efficiency in Natural Gas Pipeline Networks." *Natural Gas Journal* 6: 23–26.

———(1990) "Auction Design for Composite Goods: The Natural Gas Industry." *Journal of Economic Behavior and Organization* (September): 127-49.

McMillian, J. (1998) "Managing Economic Change: Lessons from New Zealand." Graduate School of International Relations and Pacific Studies, University of California, San Diego (July).

New Zealand Electricity Market (1999) *Rules of NZEM*. Wellington, NZ: The Marketplace Company, Ltd.

Olson, M.; Rassenti, S; and Smith, V. (2001) "Market Design and Motivated Human Trading Behavior in Electricity Markets." (Forthcoming in *IIE Transactions on Operations Engineering*.)

Outhred, H. (2001) "Electricity Industry Restructuring in California and Its Implications for Australia." Report for the National Electricity Market Management Company (February).

Outhred, H., and Kaye, R. J. (1996) "Structural Testing of the National Electricity Market Design: Final Report." Report for the National Grid Management Council, Unisearch (September).

Rassenti, S. (1981) "0–1 Decision Problems with Multiple Resource Constraints: Algorithms and Applications." Unpublished Ph.D. thesis, University of Arizona.

Rassenti, S.; Reynolds, S.; and Smith, V. (1994) "Cotenancy and Competition in an Experimental Auction Market for Natural Gas Pipeline Networks." *Economic Theory* 4: 41–65.

Rassenti, S., and Smith, V. (1986) "Electric Utility Deregulation." In *Pricing Electric, Gas and Telecommunication Services*, 53–74. The Institute for the Study of Regulation (February)

Rassenti, S.; Smith, V.; and Bulfin, R. (1982) "A Combinatorial Auction Mechanism for Airport Time Slot Allocation." *Bell Journal of Economics* 13(Autumn): 402–17.

Rassenti, S.; Smith, V.; and Wilson, B. (2000) "Controlling Market Power and Price Spikes in Electricity Networks: Demand-Side Bidding." Working Paper, Economic Science Laboratory, University of Arizona.

Reynolds, S. (1990) "Cost Sharing and Competition among Daily Newspapers." Department of Economics, University of Arizona (October).

Smith, V. (1962) "An Experimental Study of Competitive Market Behavior." *Journal of Political Economy* 70: 111–37.

———(1982a) "Microeconomic Systems as an Experimental Science." *American Economic Review* 72: 923–55.
———(1982b) "Markets as Economizers of Information: Experimental Examination of the 'Hayek Hypothesis." ' *Economic Inquiry* 20: 165–79.
———(1987) "Currents of Competition in Electricity Markets." *Regulation* 11(2): 23–29.
———(1988) "Electric Power Deregulation: Background and Prospects." *Contemporary Policy Issues* 6: 14–24.
———(1993) "Can Electric Power—A Natural Monopoly—Be Deregulated?" In H. H. Landsberg (ed.) *Making National Energy Policy*. Washington: Resources for the Future.
———(1996) "Regulatory Reform in the Electric Power Industry." *Regulation* 1: 33–46.
Williams, A. (1980) "Computerized Double Auction Markets: Some Initial Experimental Results." *Journal of Business* 53: 235–58.

Simple-Offer versus Complex-Offer Auctions in Deregulated Electricity Markets

Rimvydas Baltaduonis

Economic Science Institute, Chapman University, One University Drive, Orange, CA 92866, USA

Phone: (714) 628-7272

Fax: (714) 628-2881

E-mail: baltaduo@chapman.edu

Abstract Previous research of complex-offer auctions designed for deregulated electricity markets finds that offer complexity allows great deal of strategic behavior, which consequently leads to anti-competitive and inefficient outcomes. In these complex-offer auctions, the sellers submit not only quantities and minimum prices at which they are willing to sell, but also start-up fees that are designed to reimburse the fixed start-up costs of the electric power generators. Using an experimental approach, I compare the performance of two complex-offer auctions against the performance of a simple-offer auction, in which the sellers have to recover all their generation costs – variable and fixed – through a uniform market-clearing price. I find that the simple-offer auction significantly reduces consumer prices and lowers price volatility. It mitigates anti-competitive effects that are present in complex-offer auctions and achieves allocative efficiency more quickly.

Keywords *Complex-offer auction, Electricity, Procurement, Efficiency, Two-part pricing, Avoidable fixed cost*

JEL Classification *C72, D4, D61, L94*

1 Introduction

11 billion kilowatt-hours were traded daily in the U.S. wholesale electric power markets in 2006. Average price ranged between a tenth of a mill and 50 cents per kWh[1]. Many of these markets employ auctions that differ from other widely used quantity-price offer auctions in their offer complexity. Besides the quantities and the minimum prices, at

[1] Energy Information Administration, Form EIA-861, Annual Electric Power Industry Report.

which the electric power producers are willing to sell, the sellers may also declare their technical constraints and start-up fees that are designed to reimburse the fixed start-up costs of the generation plants. The start-up costs are avoidable fixed costs that create non-convex allocation problems. This paper investigates what value is gained from incorporating this complexity into deregulated electricity markets. The generation contracts are allocated daily by a sealed-offer auction that employs a computationally involved market-clearing algorithm. Besides applying a rule for offer selection, a market-clearing algorithm has to ensure that the system demand and reserve requirements are met over a particular time.

Baltaduonis (2007b) compares the performance of two such auctions with regard to consumer prices and efficiency by using a laboratory experiment. The major finding is that the sellers exploit the offer complexity to extract high payments from the buyers. Consequently, the outcomes result in substantial inefficiencies. In this paper, I use a laboratory experiment to contrast the performance of these complex-offer auctions (COAs) against the performance of a simple-offer auction (SOA) where the sellers can recover their generation costs – both variable and avoidable fixed – only through a uniform market-clearing price (MCP). The paper inquires if the SOA could mitigate the anti-competitive behavior that is present in the COAs.

Two COAs differ from each other in their market-clearing algorithms. An offer cost minimization (OCM) algorithm is currently used by independent system operators (ISOs) in the U.S. It relies on the traditional unit commitment approach.[2] The algorithm minimizes the total offered cost of electricity for a given demand as if all selected sellers

[2] For a bibliographical survey on the unit commitment problem see Padhy (2004).

would be paid their offered prices and fees. *Sequentially*, after the offers are selected, a uniform MCP is determined as the highest accepted price for that period. All selected sellers receive their individual start-up fees and the *uniform* MCP for the supplied electricity during that period.

Yan and Stern (2002) point out that the OCM algorithm does not ensure the lowest procurement cost of electricity to consumers for a given set of offers. This motivated Luh et al. (2005) to develop a payment cost minimization (PCM) algorithm that minimizes the actual procurement cost of electricity *simultaneously* determining a MCP as the highest accepted price during that period. As in the OCM auction, the selected sellers would receive their individual start-up fees and the *uniform* MCP for the supplied electricity. Electrical engineers have studied non-convex optimization problems, similar to the OCM and the PCM algorithms, for many years. Attempts to improve these mechanisms heavily depend on the assumption of complete information about the generation costs of electric power.[3] Baltaduonis (2007b) reports that in both the OCM and the PCM auctions, sellers significantly raise the start-up fees and prices over their true start-up costs and variable production costs even in an environment with many competitors. Such behavior leads to both allocative and production inefficiencies. Thus, the theoretical assumption of truthful production cost revelation seems to be unwarranted.

To analyze the performance of the SOA I hold constant all other characteristics of the system described by Baltaduonis (2007b). The SOA is a less computationally involved auction than the COAs and thus more transparent to market participants. On the other

[3] For a recent work on mechanisms for markets with non-convexities that is motivated by electric power markets see O'Neill et al. (2004).

hand, the exact revelation of production costs is impossible in the SOA. The sellers have to mark up their offered prices to account for the fixed start-up costs or they might incur losses. A higher risk of losses becomes a concern.

Van Boening and Wilcox (1996), hereafter VW, report an experiment in which a continuous double SOA fails to converge and stabilize on 100% efficient allocations in an environment with avoidable fixed costs. Durham et al. (1996), hereafter DRSVW, explore two-part pricing competition in a sealed-offer auction experiment as a means of improving efficiency in the VW environment. DRSVW find that in a setting with experienced sellers and simulated buyers, this institution is effective in promoting full efficiency, however, still not immune to efficiency collapses. In a different environment with both fixed sunk and fixed avoidable costs, Durham et al. (2004), hereafter DMORS, examine the price levels under a SOA by varying the demand elasticity and the experience level of sellers. They observe the pattern of price signaling and responses which despite the presence of fixed costs help to maintain above normal profits. The authors do not comment on the efficiency performance of the auction or the magnitude of observed losses in the market.

All aforementioned studies model the market demand as static. This paper simulates a cyclical nature of the daily demand for electricity. Baltaduonis (2007b) points out that cyclical market demand might be essential in shaping strategic behavior in the COAs. As in Rassenti, Smith & Wilson (2003a, 2003b), hereafter RSW, my experiment allows for strategic behavior, controls for the level of unilateral market power, simulates trading environments with minimal demand elasticity, cyclical demand uncertainties and an absence of significant excess production capacity.

Concerns about market power in the electric power industry abound. One might want to know which trading rules are more effective in suppressing the exercise of market power. In the context of capacity-constrained competitors, Holt (1989) defines market power as the ability to deviate *profitably* and *unilaterally* from the competitive outcome. Baltaduonis (2007b) reports that both the OCM and the PCM auctions produce noncompetitive outcomes even in the treatments with no unilateral market power. Since the SOA reduces the scope of possible strategic behavior, I hypothesize that ceteris paribus, the SOA should increase competitiveness in the market. The COAs' intention to account for the non-convex cost structures of generation plants also opens opportunities to strategize over the different parameters of complex offers. The opportunities are fewer on that regard in the SOA. Baltaduonis finds that in the COAs, the offer complexity and the cyclical nature of market demand create incentives to start-up plants during the higher demand periods. Consequently, the incentives to compete for baseload or shoulder demand units vanish even with the presence of cheap excess production capacity. Opting for a SOA should eliminate these anti-competitive incentives.

An intention of this experimental study is to complement theoretical research of auctions where avoidable fixed costs are an important production characteristic. The study sheds some light on possible strategic behavior in smart markets that are proposed for wholesale electric power markets. The remainder of the paper is organized as follows. Section 2 outlines the market environment in the experiment and describes three auctions. Section 3 presents the experimental design and procedures. Section 4 reports the findings. Section 5 concludes and discusses the implications for public policy.

2 Market Institution, Structure and Environment

To isolate the institutional effects of the strategically complex auctions, I examine a very simple environment relative to actual electric power systems: (i) transmission constraints are negligible; (ii) generators have no physical ramping rates; (iii) security reserves and other ancillary services to protect the system from outages are ignored; and (iv) a trading institution accepts flat offer curves for each generating unit. Such an environment is most comparable to day-ahead wholesale markets of observed power systems. The performance of the SOA is measured against the OCM and the PCM auctions in a stationary supply and cyclical demand environment, controlling for unilateral market power.

2.1 Auction Institution

The sellers privately submit a schedule of offers; that is, plant start-up fees and prices for their production capacity for each pricing period of a day. The buyers submit a schedule of bids. Since active demand-side bidding is often absent in the naturally occurring spot markets for electricity, a computer is used to submit bids that perfectly reveal the demand at any point in time in the experiment[4]. The offers and the computerized bids are then sent to a market-clearing algorithm to allocate the production contracts for the next day. Currently, the dominant practice in the electricity spot markets is to employ uniform price auctions where each seller receives the same market price for the sold megawatts. The market price is usually the highest accepted price per megawatt among all the sellers. I retain these institutional features and put aside the discussion about the "pay-as-offered"

[4] Same as in RSW, DRSVW and DMORS.

discriminatory price auctions.[5] In all experimental treatments, i.e. OCM, PCM and SOA, the sellers get paid *uniform* prices and their *individual* start-up fees. In the SOA, the start-up fees are simply constrained to be zero.

In case of a uniform price auction where sellers ask for fixed start-up fees, the mechanism of distributing these fees across consumers is important. One way to do that is to divide the borne fees equally over the units dispatched during the period for which the extra generation capacity was called. The markup on the highest accepted offered price creates a gap between a uniform price that all sellers receive and a uniform price that all buyers pay. In this experiment, both the OCM and the PCM algorithms employ this method to compute the buyer prices and to determine the corresponding levels of demand. Note that a uniform price that all sellers receive and a uniform price that all buyers pay are the same in the SOA due to the absence of start-up fees.

2.1.1 The OCM Auction

The OCM algorithm minimizes the total offer costs of electricity, as if all selected sellers would be paid their offer prices and fees:

$$\min_{q_i(t),x_i(t)} \sum_{t=1}^{T}\sum_{i=1}^{N}\left(c_i(t)q_i(t)+f_i(t)x_i(t)\right)$$

$$\begin{aligned}
\text{subject to } & \sum_{j=1}^{M} d_j(t) = \sum_{i=1}^{N} q_i(t) \quad \forall t = 1,...,T, \\
& q_i(t) \le x_i(t)k_i(t) \quad \forall i = 1,...,N, \\
& q_i(t) \ge x_i(t)l_i(t) \quad \forall i = 1,...,N, \\
& x_i(t) \in \{0,1\} \quad \forall i = 1,...,N.
\end{aligned}$$

[5] For experimental investigations of uniform price versus discriminatory price auctions SOAs see Mount, Schulze, Thomas & Zimmerman (2001), and Rassenti, Smith & Wilson (2003b).

where $i = 1,...,N$ indexes the generation plants;
$j = 1,...,M$ indexes the buyers;
$t = 1,...,T$ indicates the pricing periods during a day;

$$\text{Offers submitted by sellers:} \begin{cases} c_i(t) = \text{price per unit asked for plant } i; \\ f_i(t) = \text{start-up fee asked for plant } i; \\ k_i(t) = \text{max capacity of plant } i; \\ l_i(t) = \text{min capacity of plant } i; \end{cases}$$

$$\text{Decision variables:} \begin{cases} q_i(t) = \#\text{ units produced in plant } i; \\ x_i(t) = \begin{cases} 1 \text{ if plant } i \text{ is chosen to produce,} \\ 0 \text{ if plant } i \text{ is not chosen to produce.} \end{cases} \end{cases}$$

After the offers are selected, a uniform MCP is determined as the highest accepted price for each period *t*:

$$MCP(t) = \max\left\{c_i(t), \forall i \text{ such that } q_i(t) > 0\right\}.$$

All selected sellers receive their individual start-up fees and the uniform MCPs for the supplied electricity.

2.1.2 The PCM Auction

The PCM algorithm minimizes the actual procurement cost of electricity, *simultaneously* determining a MCP as the highest accepted price for each period *t*:

$$\underset{q_i(t),x_i(t)}{\text{Min}} \sum_{t=1}^{T}\sum_{i=1}^{N}\left(MCP(t)q_i(t) + f_i(t)x_i(t)\right)$$

$$\begin{aligned} \text{subject to } \sum_{j=1}^{M} d_j(t) = \sum_{i=1}^{N} q_i(t) \quad & \forall t = 1,...,T, \\ q_i(t) \le x_i(t)k_i(t) \quad & \forall i = 1,...,N, \\ q_i(t) \ge x_i(t)l_i(t) \quad & \forall i = 1,...,N, \\ x_i(t) \in \{0,1\} \quad & \forall i = 1,...,N. \end{aligned}$$

As in the OCM auction, the selected sellers receive their individual start-up fees and the uniform MCPs for the supplied electricity.

In the experiment, both the OCM and the PCM auctions are designed to sell the maximum amount of units where buyers' marginal willingness to pay is higher or equal to a buyer price. Tied offer combinations in the OCM auction are picked in a way that generates lower procurement cost. Tied offer combinations in the PCM auction are selected by giving priority to those sellers whose offer cost is lower. Such tie breaking mechanism gives the best performance chances to both COAs. To achieve similar tie breaking in real life applications would require additional costly computational power and time.

2.1.3 The Simple-Offer Auction

The sellers in the SOA can recover their production costs – both variable and avoidable fixed – only through a uniform MCP. Note that the SOA is a special case of the COAs, i.e. it is a COA where the start-up fees are constrained to be zero. The contract allocations in two COAs are identical when the start-up fees equal to zero. The offer cost minimization becomes equivalent to the payment cost minimization. Hence, either the OCM or the PCM algorithm could be used for the SOA by simply restricting all start-up fees to be zero, i.e. $f_i(t) = 0$ for $\forall i = 1,...,N$. The selected sellers receive the uniform MCPs for the supplied electricity.

In the complex-offer auctions, the suppliers are able to reveal their costs and be reimbursed in a way that the costs are incurred. In the SOA, the sellers have to think how to recover the fixed costs through the offered prices. See Appendix A for a simple

numerical example that demonstrates the principles of offer-selection rules for all three considered auctions.

2.2 Environment

2.2.1 Supply & Demand

Each day in the experiment consists of four pricing periods: off peak period (low demand/night), shoulder period (medium demand/morning), peak period (high demand/ afternoon) and shoulder period (medium demand/ evening). Four pricing periods during the day are a simplification of the naturally occurring day-ahead electricity markets where separate prices are instituted hourly. Nevertheless, the cyclical dynamics of the demand are preserved.

Tables 1 and 2 as well as Fig. 1 depict aggregate supply and demand in the experimental environment. The second and third steps of the demand in Table 1 represent interruptible units of demand whereas the units on the first step at 250 are the "must serve" units. The level of "must serve" demand varied among three levels: 1 unit in off-peak periods, 4 units during shoulder periods, and 14 units during peak periods.

The market is comprised of six sellers denoted by an "*S*" and an identification number. The sellers own 13 plants of nine types. The technical characteristics of each plant are presented in Table 2. Fig. 1 presents the ownership of the plants. *S*1 and *S*2 own two low cost (type *A*) plants and two high cost plants (type *H* and *G* respectively). *S*3 and *S*4 own two high cost (type *E*) plants and respectively, one baseload (type *B*) plant and one intermediate cost (type *C*) plant. *S*4 also owns a very high cost (type *I*) peak capacity plant with average total cost (ATC) exceeding even the resale value at the "must serve"

level. Each *S*5 and *S*6 own one intermediate cost (type *D*) plant and one high cost (type *G* and *F* respectively) peak capacity plant.

[a]Average total costs at the maximum capacity of a plant.

Fig. 1 Market Structure and Design

Table 1 Demand Schedules

Demand	Quantity (demand values)		
	Step 1	Step 2	Step 3
Off-peak	1 (250)	1 (80)	N/A
Shoulder	4 (250)	2 (230)	1 (160)
Peak	14 (250)	2 (230)	2 (160)

The types and the distribution of ownership of the plants are designed to create a Bertrand-like competition between the marginal plants during each period of a day. In other words, at least two plants with identical production costs on the supply margins exist for each level of demand. In a competitive bidding process, *S*1's plant *A* can be

easily replaced by *S*2's plant *A* during an off-peak period and *S*5's plant *D* can be easily replaced by *S*6's plant *D* during the shoulder periods. Five plants with 10 units of total capacity and identical ATC compete to supply six units of peak demand. Some plants have low start-up costs with high production costs per unit, while other plants have high start-up costs but lower production costs per unit. In a competitive equilibrium, the number of supplied units is 2 in off-peak periods, 7 in shoulder periods and 16 in peak periods. The lower quantities of supplied units would be the evidence of allocative inefficiencies. Note that an efficient allocation of production contracts would never include *S*4's plant *I* and *S*1's plant *H*.

Table 2 Average Total Costs (ATC) of Production at Maximum Capacity (Cap.) by Plant Type

Plant Type (Quantity)	Min Cap. Units	Max Cap. Units	Start-up Cost $	Per Unit Cost $/Unit	ATC at Max Cap. $/Unit	Total Cap. Units
A (2)	0	2	0	20	20	4
B (1)	1	1	10	15	25	1
C (1)	0	1	20	70	90	1
D (2)	0	2	6	93	96	4
E (2)	0	2	120	112	172	4
F (1)	0	2	80	132	172	2
G (2)	0	2	40	152	172	4
H (1)	0	2	0	225	225	2
I (1)	0	2	0	255	255	2
					Total	24

2.2.2 Unilateral Market Power

In the experiment, a seller is said to be able to exert market power if, for a given distribution of capacity ownership, a seller profitably and unilaterally can submit an offer schedule above his plants' costs (or equivalently withdraw some generating capacity) such that the market price rises above the competitive level.

The costs for the units in the marginal plants are 20, 99, 172 and 93 per unit in the off-peak, shoulder 1, peak and shoulder 2 periods respectively. In the SOA, these costs translate into the competitive market prices at which the marginal plants earn zero economic profits. None of the six sellers can benefit from a unilateral attempt to raise the market prices above the competitive level. In doing so, the sellers of intermediate units would jeopardize their profits and the sellers of marginal units would simply lose the contract to his Bertrand-like competitor. The competitive prices correspond to a pure strategy Nash equilibrium in the SOA.

In the COAs, the marginal generators also have incentives to submit offers that are equal to the actual production costs of the marginal units but only if we look at an isolated period of the demand cycle. The asked fees do not necessarily need to be the actual start-up costs but then the offered seller prices need to add up to the actual production costs. Consider the OCM auction for an illustration. Take a shoulder 1 period. Each *S*5 and *S*6 owns a marginal intermediate cost plant that competes to supply the marginal seventh unit to the market. Either plant can generate this marginal unit at a cost of 99 [6+93]. If a seller offers to supply the unit at a cost higher than 99, the other seller would be able to undercut the offer by either lowering the fixed fee or lowering the offered seller price. Therefore, a start-up fee and a price per unit would have to add up to 99 in a competitive offer of *S*5 or *S*6.

Similarly, the competitive marginal offers $[(\text{start-up fee} + \text{price per unit} \times 2) \div 2]$ would have to be 20 during the off-peak periods, 172 during the peak periods and 93 during the shoulder 2 periods.

This illustration disregards the incentives to withhold a plant's capacity during the lower demand periods due to the opportunities to extract bigger start-up fees during the higher demand periods. For instance, *S*6 can decide to delay the start-up of his plant by placing a high offer for shoulder 1 period. Assume that *S*6's peak offer is 340 for the start-up fee and 1 for price. This offer is still cheaper than the competitive offers of the peak plants for 172 per unit $\left(> \left(340 + 2 \times 1\right) \div 2\right)$. In this case, *S*6 makes a significant profit by abandoning competition in shoulder 1 period and by deliberately delaying the start-up of his plant *D* till the peak period. Consequently, *S*5 can now profitably raise her offer above her costs for shoulder 1 period as the competition from her Bertrand-like competitor *S*6 is absent. A competitive outcome becomes practically impossible. By allowing the sellers to submit complex-offers, the market mechanism automatically creates market power. This result holds for both the OCM and the PCM auctions. Note that the additional production capacity would not improve competitiveness in this environment because the suppliers of the additional capacity would have the same incentives to delay the start-ups.

The supply and demand are the same during the two shoulder periods of a day. However, since most plants are generating electricity during the peak period, they do not incur start-up costs and do not receive start-up fees to continue production during the shoulder 2 period, i.e. $f_i(\text{shoulder 2}) = 0 \quad \forall i$ such that $q_i(\text{peak}) > 0$. For this reason, all three examined auctions should perform similarly during the fourth period of a day. Theoretically, none of the six sellers can benefit from a unilateral attempt to raise the MCP above 93. The competitive price of shoulder 2 corresponds to a pure strategy Nash equilibrium in all three auctions.

3 Experimental Design and Procedures

To compare how the behavior and market performance differ in the complex- and simple-offer auctions, I conducted 12 experimental sessions using undergraduate students at George Mason University. The data from eight sessions of the OCM and the PCM treatments was previously presented by Baltaduonis (2007b). The reported data from four SOA sessions is new. Each session lasted 53 trading days. The dataset discussed in this paper includes a total of 636 trading days. Each session lasted approximately 90 minutes. The subjects in each market were provided with complete information about the market supply structure. Plants' minimum and maximum production capacity, start-up cost, cost per unit and the ownership of all plants were public information. Information about demand, however, was not available to the subjects. The situation was framed as a market for identical product to avoid the use of possibly intimidating or confusing electric power jargon. The instructions informed the subjects that the costs and production capacities for each seller would not change during the experiment, but that the purchased quantities of the product would vary over the course of a day. In particular, the instructions indicated that the computer will purchase "low" amounts of product for the first quarter of a day, "medium" amounts for the second quarter of a day, "high" amounts for the third quarter of a day and "medium" amounts for the fourth quarter of a day. The subjects did not know the number of trading days in advance. The instructions were read aloud in the beginning of each session.

A subject had 75 seconds to submit an offer for each day.[6] An exception was made for the first day offers. The sellers could take as much time as they needed to formalize their initial offers. Once the last seller submitted her offer for the first day, the following trading days were limited to 75 seconds. The offers were automatically filled in with the offer information from the previous trading day. However, a seller could revise her offer at any time within the 75 second period. An offer indicated the prices, start-up fees and quantities of the product that a seller was willing to supply from a particular plant over the course of the following day. The subjects could not alter the minimum and maximum quantities of the offer.[7] These quantities were set equal to the minimum and maximum capacities of a plant. The subjects could still effectively withdraw the capacity from the market by asking extremely high prices for those capacity units. Thus, in a COA, a seller had to decide on the price and the start-up fee for each plant and for each quarter of the upcoming day.[8] In the SOA, a seller had to decide only on the price for each plant for each quarter of the upcoming day, as all start-up fees were set equal to zero. The instructions pointed out that the actual market price may be higher than their offered price and that all sellers would receive the same market price if their offers were selected. The sellers received start-up fees only for the periods when their plant had to be started. In the beginning of each day all plants were idle.

---

[6] The chosen time frame is similar to one-minute trading days of the RSW electric power experiments and 75 s trading days of DMORS experiment.

[7] ISOs usually demand an explanation if generators change their offered generation capacity or technical constraints. Thus strategic behavior is somewhat limited with regards to these parameters of an offer.

[8] I am aware that there are various initiatives to regulate start-up cost reimbursement (e.g. limiting the ability to change the start-up fees freely; and partial start-up cost reimbursement) for electric power generators in naturally occurring markets. However, the purpose of the study is to investigate the performance of the two auctions when such interventions into a deregulated market are absent.

At the end of the trading day, all offers were sent to the computerized market coordinator. A market-clearing algorithm was applied and the results of a sealed-offer auction were sent back to the sellers. Each seller could see how many units she sold, what the MCP for each period was and what profit/loss she earned on every owned capacity unit during each period of a day. The screens also displayed a history of the market prices from the past 10 days and the sold quantities during each quarter of the last day. The amount of paid fees was not public information.[9]

Subjects were paid $7 for showing up on time for the sessions. In addition to this show-up payment, the average earnings per subject for the data reported here was $21.55.

4 Results

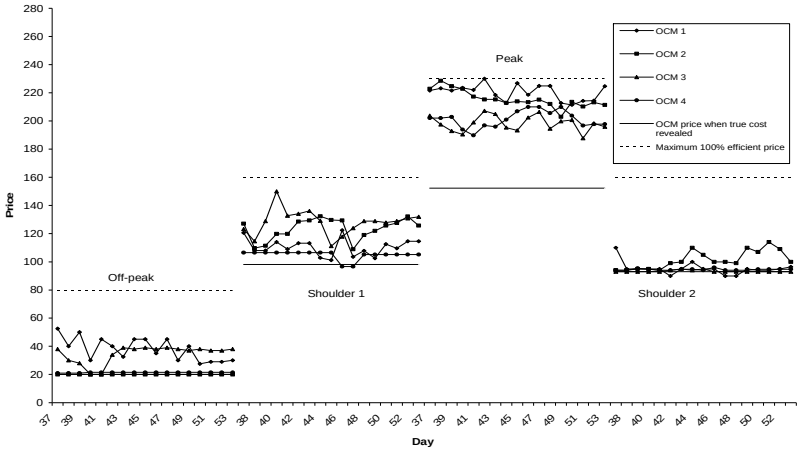
The SOA, OCM and PCM auctions respectively extract on average 93, 92 and 94 percent of maximum total surplus. All three auctions sell on average 32 units a day. Thus, considering that the demand side of the market is perfectly revealed in the experiment, lower levels of allocative efficiency must be attributed to higher degrees of production inefficiency. To present how the captured total surplus is allocated among buyers and sellers and how volatile the allocation is, Fig. 2 depicts the buyer prices in each session of the three treatments. The last 17 days of the data are grouped by level of demand (quarter) then sequenced by how the demand varied over a market day: off-peak, shoulder 1, peak and shoulder 2.

I evaluate the results with respect to a benchmark of true cost revelation. The outcome of true cost revelation is particularly interesting in electricity markets because the design

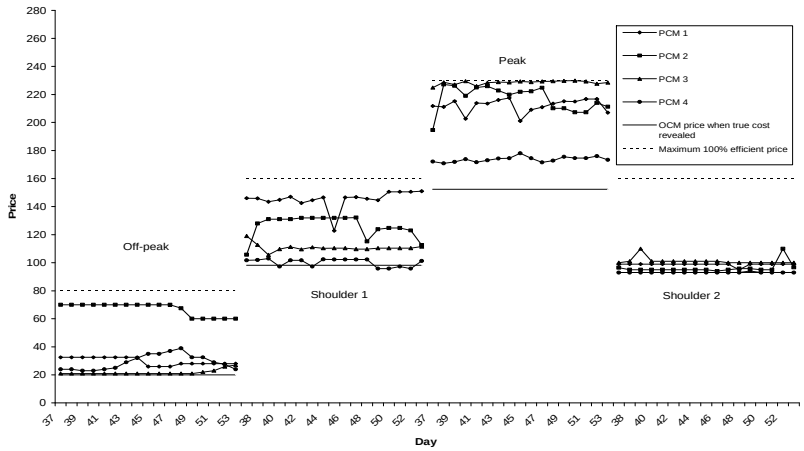
[9] See Appendix B for the experimental instructions and Appendix C for an example of a subject screen during the experiment.

and the engineering of these complicated market systems often start with the assumption of the true cost revelation. In Fig. 2, the outcome of the perfectly revealed costs is shown as a solid line. The dotted line represents the value of the nearest unit of interruptible demand. The prices up to the dotted line are 100% efficient with respect to allocation. As an attempt to control for the convergence of the bidding behavior, I focus on the last 17 market days (1/3 of all days) in each session unless referred otherwise.

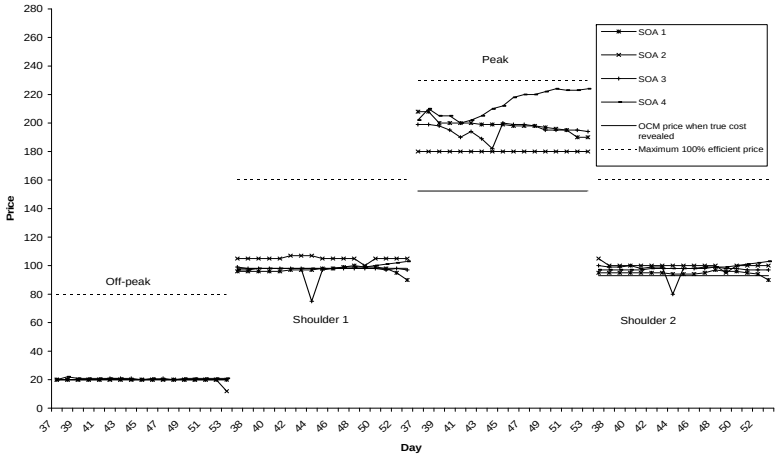
In Fig. 2, the SOA is more likely to approach the true cost revelation outcome than either of the two COAs. Both the OCM and the PCM auctions tend to deviate significantly from the outcome of perfectly revealed costs with shoulder 2 periods being an exception. In the SOA, the buyer prices substantially depart from the competitive outcome only during the peak periods. During shoulder 2 periods, all three auctions result in competitive outcomes. The conformity is not accidental since the fixed costs are absent in this period and, therefore, all three offer selection rules are identical as the start-up fees equal to zero.



(a) OCM



(b) PCM



(c) SOA

Fig. 2 Buyer Prices by Level of Demand for the Last 17 Market Days in Each Session

In what follows, the experimental results are summarized as a series of six findings. In addition to the qualitative results displayed in the figures, I analyze the data using a mixed-effects model for repeated measures on each of several sessions using different subjects. The results from estimating this model for the buyer prices by level of demand are given in Table 3. The dependent variable in this case is the difference between the observed buyer price (*Price*) and the buyer price from the OCM auction when production costs are perfectly revealed by the sellers, P^t. In the regressions, the SOA is used as a benchmark institution to allow for its straightforward comparison against the OCM and the PCM auctions. The treatment effects (*OCM* and *PCM*) are modeled as (zero-one) fixed effects, whereas the sessions are modeled as random effects, e_i. As mentioned above, the experimental days are divided into three equal groups to capture effects like learning over time. In the model, the data from the *First* and *Second* groups (days 1-18 and 19-36, respectively) are identified by (zero-one) dummy variables. Specifically, the estimated model is as follows:

$$Price_{ij}\text{-}P^t=\mu+e_i+\beta_1 OCM_i+ \beta_2 PCM_i +\beta_3 First_i + \beta_4 Second_i +\beta_5 OCM_i\times First_i+$$

$$\beta_6 OCM_i\times Second_i + \beta_7 PCM_i\times First_i+ \beta_8 PCM_i\times Second_i +\varepsilon_{ij};$$

where the sessions are indexed by i=1,…,12 and the repeated market days by j=1,…,53. $e_i \sim N(0,\sigma^2_1)$ and $\varepsilon_{ij} \sim N(0,\sigma^2_{2,i})$.

Finding 1. *Ceteris paribus, the SOA institution significantly lowers buyer prices relative to the COAs in the periods when start-up costs are relevant. Buyer prices are not significantly different in shoulder 2 periods when no new plants need to be started and no start-up fees need to be paid.*

Support: Fig. 2 clearly illustrates that both the OCM and the PCM auctions can produce higher buyer prices than the SOA in all three periods where new plants need to be started, that is, in off-peak, shoulder 1 and peak periods. Except for peak periods, buyer prices in the SOA settle very close to the competitive equilibrium, i.e. 20, 99, 172 and 93, during the respective quarters of a day. SOA prices for peak periods do not come close to the expected competitive level of 172. My speculation is that the incentives to undercut the competitors' offers are weaker in the peak periods because winning a marginal contract and setting a lower uniform market price also means smaller profits for the low or/and intermediate cost plants that the seller owns. On the other hand, no discernible separation exists among three auctions in shoulder 2 prices. Since most of the plants are operating during the peak periods, no new plants need to be started when market demand falls. The absence of start-up fees makes the three offer selection rules identical which consequently should lead to similar outcomes.

These qualitative observations are supported by estimates from the mixed-effects model in Table 3. The SOA significantly reduces prices by 19.5 (*p-value*=0.0077) and 17 (*p-value*=0.0744) experimental dollars in the shoulder 1 and peak periods when compared to the OCM auction. The SOA significantly reduces prices by 18.2 (*p-value*=0.0417), 17.4 (*p-value*=0.0127) and 20 (*p-value*=0.0418) experimental dollars respectively in the off-peak, shoulder 1 and peak periods when compared to the PCM auction. The prices in shoulder 2 periods are not significantly different across all three auctions (*p-values*=0.1615, 0.4404 for OCM and PCM, respectively).■

Table 3 Estimates of the Linear Mixed Effects Model of Treatment Effects for the Buyer Prices

$Price_{ij} - P^{I} = \mu + e_i + \beta_1 OCM_i + \beta_2 PCM_i + \beta_3 First_i + \beta_4 Second_i + \beta_5 OCM_i \times First_i + \beta_6 OCM_i \times Second_i + \beta_7 PCM_i \times First_i + \beta_8 PCM_i \times Second_i + \varepsilon_{ij}$, $e_i \sim N(0,\sigma^2_1)$ and $\varepsilon_{ij} \sim N(0,\sigma^2_{2,i})$

	Estimate	Std. Error	Degrees of Freedom	H_a	t-statistic	p-value
Off-peak						
μ	**-1.21**	**5.41**	**618**	**μ>0**	**-0.22**	**0.8227**
OCM	**8.58**	**7.70**	**9**	**$\beta_1 \neq 0$**	**1.12**	**0.2935**
PCM	**18.23**	**7.68**	**9**	**$\beta_2 \neq 0$**	**2.37**	**0.0417**
First	-1.66	0.40	618	$\beta_3 \neq 0$	-4.19	<.0001
Second	-0.36	0.40	618	$\beta_4 \neq 0$	-0.92	0.3584
OCM×First	-5.06	1.28	618	$\beta_5 \neq 0$	-3.95	0.0001
OCM×Second	-2.35	1.28	618	$\beta_6 \neq 0$	-1.84	0.0669
PCM×First	-0.81	0.89	618	$\beta_7 \neq 0$	-0.90	0.3664
PCM×Second	-1.31	0.89	618	$\beta_8 \neq 0$	-1.47	0.1434
Shoulder 1						
μ	**-1.13**	**3.99**	**618**	**μ>0**	**-0.28**	**0.7771**
OCM	**19.53**	**5.71**	**9**	**$\beta_1 \neq 0$**	**3.42**	**0.0077**
PCM	**17.36**	**5.60**	**9**	**$\beta_2 \neq 0$**	**3.10**	**0.0127**
First	-9.02	1.80	618	$\beta_3 \neq 0$	-5.02	<.0001
Second	1.96	1.80	618	$\beta_4 \neq 0$	1.09	0.2760
OCM×First	1.05	2.82	618	$\beta_5 \neq 0$	0.37	0.7102
OCM×Second	-3.57	2.82	618	$\beta_6 \neq 0$	-1.27	0.2054
PCM×First	16.74	2.11	618	$\beta_7 \neq 0$	7.93	<.0001
PCM×Second	-0.31	2.11	618	$\beta_8 \neq 0$	-0.15	0.8832
Peak						
μ	**40.15**	**5.98**	**618**	**μ>0**	**6.71**	**<.0001**
OCM	**17.02**	**8.44**	**9**	**$\beta_2 \neq 0$**	**2.02**	**0.0744**
PCM	**19.98**	**8.43**	**9**	**$\beta_2 \neq 0$**	**2.37**	**0.0418**
First	9.25	1.82	618	$\beta_3 \neq 0$	5.09	<.0001
Second	2.88	1.82	618	$\beta_4 \neq 0$	1.58	0.1137
OCM×First	-11.42	2.59	618	$\beta_7 \neq 0$	-4.41	<.0001
OCM×Second	1.47	2.59	618	$\beta_8 \neq 0$	0.57	0.5710
PCM×First	-17.87	2.48	618	$\beta_7 \neq 0$	-7.21	<.0001
PCM×Second	-6.24	2.48	618	$\beta_8 \neq 0$	-2.52	0.0121
Shoulder 2						
μ	**3.12**	**1.18**	**618**	**μ>0**	**2.65**	**0.0083**
OCM	**-2.29**	**1.50**	**9**	**$\beta_2 \neq 0$**	**-1.53**	**0.1615**
PCM	**-1.31**	**1.62**	**9**	**$\beta_2 \neq 0$**	**-0.81**	**0.4404**
First	-0.44	1.12	618	$\beta_3 \neq 0$	-0.40	0.6924
Second	2.29	1.12	618	$\beta_4 \neq 0$	2.04	0.0415
OCM×First	-0.07	1.16	618	$\beta_7 \neq 0$	-0.06	0.9539
OCM×Second	-2.61	1.16	618	$\beta_8 \neq 0$	-2.25	0.0248
PCM×First	10.87	1.65	618	$\beta_7 \neq 0$	6.57	<.0001
PCM×Second	1.07	1.65	618	$\beta_8 \neq 0$	0.65	0.5191

Note. The linear mixed-effects model is fit by maximum likelihood with 636 original observations and 12 sessions. For purposes of the brevity the session random effects are not included in the table.

Finding 2: *Ceteris paribus, markets in the SOA treatment quickly stabilize at the competitive equilibrium quantity at all levels of demand, whereas the COAs continue to interrupt market demand throughout the experiment, especially during the peak periods.*

Support: On only 18 occasions (out of possible 848 = 53 days × 4 quarters × 4 sessions) the SOA exchanged an allocative inefficient quantity. 17 of these occasions happened during the peak periods. The last inefficient allocation was observed during the twelfth trading day in session 2. The OCM (PCM) auction experienced 55 (24) allocative inefficient exchanges, with the latest observation being during the 44^{th} (53^{rd}) trading day. 44 (19) or 80% (79%) of these inefficient exchanges happened during the peak periods. In Fig. 2, the last 17 days in all sessions resulted in 100% efficient buyer prices. However, this does not necessarily mean that all OCM sessions supplied the efficient quantity to the market during all those days. In fact, the demand had to be interrupted on five occasions (out of possible 272 = 17 days × 4 quarters × 4 sessions), because the price for the efficient amount exceeded buyers' maximum willingness to pay. Similarly, the demand was interrupted on four occasions in the PCM sessions.■

Failure to supply the efficient amount of units is not the only source of possible inefficiencies. The total surplus might also be reduced by production inefficiencies, i.e. the situations when higher cost plants produce while lower cost plants are idle.

Finding 3: *Ceteris paribus, the COA and SOA treatments exhibit similar degrees of production inefficiency.*

Support: Fig. 3, the estimates from the mixed-effects model in Table 4 and the statistics of the non-parametric Mann-Whitney U test in Table 5[10] report evidence that in all periods, the COA and SOA treatments are not significantly different from each other. The dependent variable in the mixed-effects model is the difference between the observed production cost (*ProdCost*) and the minimum production cost for the exchanged quantity, *ProdCost**.[11] The treatment effects are insignificant for all periods[12].

Some of the production inefficiencies in the COAs stem from frequent occasions when the very high cost generators of H and I types are called to produce. The owners of the most inefficient plants (type H and I) are able to win contracts and profitably supply to the market by offering low prices and recovering their variable costs through high start-up fees. During the last 17 days of the OCM sessions, these plants are selected and make positive profits during 42 days [out of possible 68 = 17days× 4sessions]. The same plants sell profitably during six days in the PCM sessions and never in the SOA.■

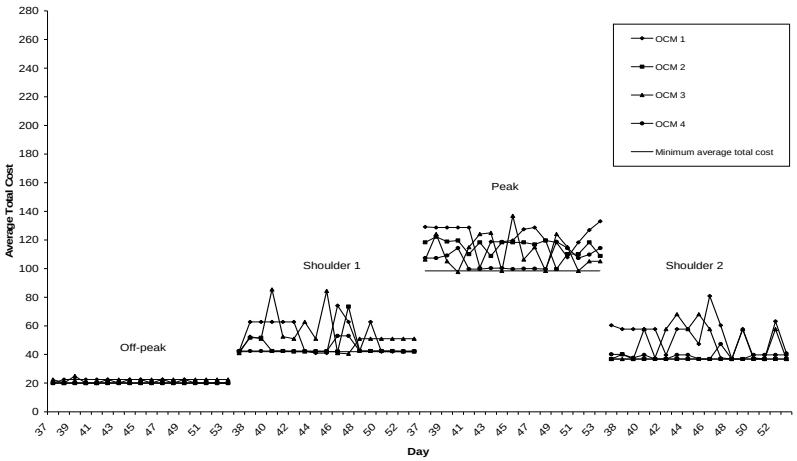
Since the deregulation of electricity markets, inflated and volatile wholesale electricity prices have been a concern.

---

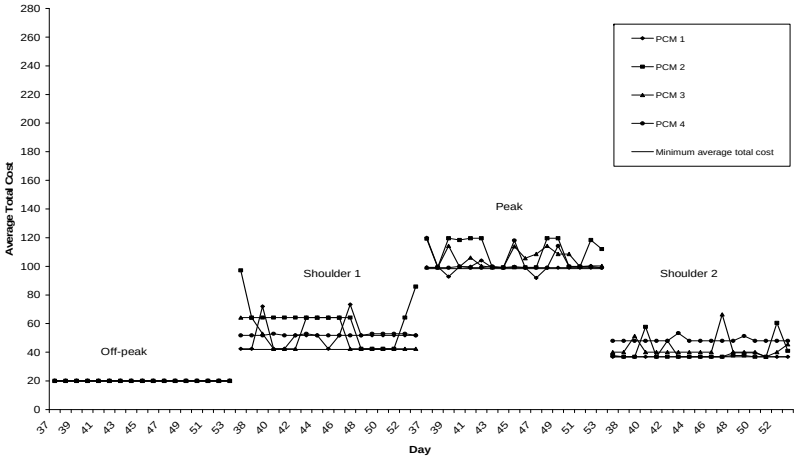
[10] Due to the lack of normality in the distribution of the estimated errors for the linear mixed effects model, I also applied the non-parametric Mann-Whitney U test to the average production costs for the last 17 market days.

[11] An interpretation of the regression results might be problematic if the exchanged quantity fluctuates across the days. However, this problem does not arise here since during the last 17 days of the experiment, the demand had to be interrupted only on 5 occasions (out of possible 272 = 17 days × 4 quarters × 4 sessions) in the OCM treatment, and on 4 occasions in the PCM treatment.

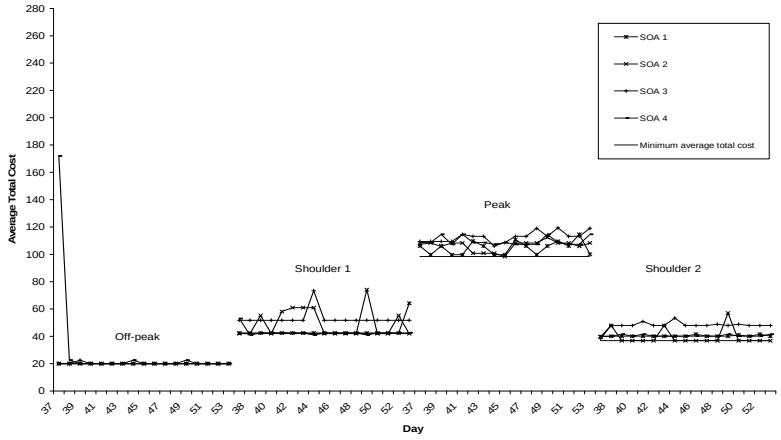
[12] The estimated mixed-effects model suggests that in peak periods, the SOA treatment raises production costs above the PCM level by 83.3 experimental dollars (*p-value*=0.0298); on the other hand, the Mann-Whitney U test finds this difference in costs statistically insignificant (*p-value*=0.1143).



(a) OCM



(b) PCM



(c) SOA

Fig. 3 Average Total Costs by Level of Demand for the Last 17 Market Days in Each Session

Table 4 Estimates of the Linear Mixed Effects Model of Treatment Effects for the Production Costs

$$ProdCost_{ij} - ProdCost^{*} = \mu + e_i + \beta_1 OCM_i + \beta_2 PCM_i + \beta_3 First_i + \\ + \beta_4 Second_i + \beta_5 OCM_i \times First_i + \beta_6 OCM_i \times Second_i + \beta_7 PCM_i \times First_i + \beta_8 PCM_i \times Second_i + \varepsilon_{ij}, \\ e_i \sim N(0,\sigma^2_1) \text{ and } \varepsilon_{ij} \sim N(0,\sigma^2_{2,i})$$

	Estimate	Std. Error	Degrees of Freedom	H_a	t-statistic	p-value
Off-peak						
Due to the lack of variability of the dependent variable, the model cannot be estimated for the off-peak periods. Treatment averages and standard deviations are presented instead.						
	Average	Std. Dev.				
SOA	**4.76**	**36.85**				
OCM	**1.69**	**2.38**				
PCM	**0.00**	**0.00**				
Shoulder 1						
μ	**47.96**	**17.63**	**618**	**μ>0**	**2.72**	**0.0067**
OCM	**-10.94**	**24.92**	**9**	**$\beta_2 \neq 0$**	**-0.44**	**0.6711**
PCM	**31.44**	**24.92**	**9**	**$\beta_2 \neq 0$**	**1.26**	**0.2389**
First	120.28	14.75	618	$\beta_3 \neq 0$	8.15	<.0001
Second	5.71	14.75	618	$\beta_4 \neq 0$	0.39	0.6986
OCM×First	-6.28	21.04	618	$\beta_7 \neq 0$	-0.30	0.7653
OCM×Second	31.00	21.04	618	$\beta_8 \neq 0$	1.47	0.1412
PCM×First	-64.65	20.70	618	$\beta_7 \neq 0$	-3.12	0.0019
PCM×Second	18.45	20.70	618	$\beta_8 \neq 0$	0.89	0.3731
Peak						
μ	**165.52**	**21.61**	**618**	**μ>0**	**7.66**	**<.0001**
OCM	**57.34**	**33.18**	**9**	**$\beta_2 \neq 0$**	**1.73**	**0.1180**
PCM	**-83.30**	**32.32**	**9**	**$\beta_2 \neq 0$**	**-2.58**	**0.0298**
First	-18.28	16.70	618	$\beta_3 \neq 0$	-1.09	0.2742
Second	-15.04	16.70	618	$\beta_4 \neq 0$	-0.90	0.3683
OCM×First	-47.65	29.11	618	$\beta_7 \neq 0$	-1.64	0.1022
OCM×Second	-35.60	29.11	618	$\beta_8 \neq 0$	-1.22	0.2218
PCM×First	72.55	27.92	618	$\beta_7 \neq 0$	2.60	0.0096
PCM×Second	61.60	27.92	618	$\beta_8 \neq 0$	2.21	0.0277
Shoulder 2						
μ	**58.95**	**18.92**	**618**	**μ>0**	**3.11**	**0.0019**
OCM	**-13.81**	**26.98**	**9**	**$\beta_2 \neq 0$**	**-0.51**	**0.6208**
PCM	**-17.44**	**26.66**	**9**	**$\beta_2 \neq 0$**	**-0.65**	**0.5293**
First	65.63	10.53	618	$\beta_3 \neq 0$	6.23	<.0001
Second	7.68	10.53	618	$\beta_4 \neq 0$	0.73	0.4663
OCM×First	51.16	16.11	618	$\beta_7 \neq 0$	3.18	0.0016
OCM×Second	21.09	16.11	618	$\beta_8 \neq 0$	1.31	0.1912
PCM×First	-4.42	16.06	618	$\beta_7 \neq 0$	-0.28	0.7830
PCM×Second	6.40	16.06	618	$\beta_8 \neq 0$	0.40	0.6903

Note. The linear mixed-effects model is fit by maximum likelihood with 636 original observations and 12 sessions. For purposes of the brevity the session random effects are not included in the table.

Finding 4: *Ceteris paribus, the variance of buyer prices in the SOA is same or lower than in the COAs.*

Table 5 Mann-Whitney U test on the Average Total Production Costs for the Last 17 Market Days

Period	SOA vs. OCM		SOA vs. PCM	
	$U_{4,4}$	p (two-tailed)	$U_{4,4}$	p (two-tailed)
Off peak	8	1.0000	12	0.3429
Shoulder 1	10	0.6857	13	0.2000
Peak	12	0.3429	14	0.1143
Shoulder 2	8	1.0000	8	1.0000

Support: Fig. 2 presents the dynamics of buyer prices in the auctions. Fig. 4 provides averages of the price variances for the 12 sessions presented here. Table 6 summarizes the results of the Mann-Whitney U test comparing the variances of the COAs against the variances of the SOA. The evidence suggests that in all periods the SOA attains at least as low volatility of prices as the COAs. Price volatility is significantly higher in shoulder 1 periods of the OCM auction (*p-value*=0.0571) and off-peak periods of the PCM auction (*p-value*=0.0571).■

Table 6 Mann-Whitney U test on the Buyer Price Variances for the Last 17 Market Days

Period	SOA vs. OCM		SOA vs. PCM	
	$U_{4,4}$	p (two-tailed)	$U_{4,4}$	p (two-tailed)
Off peak	9.5	0.6857	15	0.0571
Shoulder 1	15	0.0571	14	0.1143
Peak	12	0.3429	9	0.8857
Shoulder 2	9	0.8857	10	0.6857

Fig. 4 Buyer Price Variances by Treatment for the Last 17 Market Days

The above findings implicate that the SOA outperforms the COAs with respect to allocative efficiency, buyer prices and price volatility. On the other hand, a concern was raised that the SOA might increase a risk of short-term losses to the sellers. Finding 5 addresses this issue.

Finding 5: *Ceteris paribus, plants experience more short-term losses in the SOA than in the COAs; however, the relative size of occasional losses compared to accumulated profits is small.*

Support: The total amounts of experienced losses in the OCM, PCM and SOA sessions are respectively 24346, 11679 and 38192 experimental dollars. The losses substantially decline towards the end of the sessions. The amounts of losses during the last 17 days of the experiment are respectively 562, 973 and 1966 experimental dollars. These losses represent 0.3, 0.6 and 1.3 percent of market profits. Fig. 5 summarizes the total amounts of experienced losses by quarter of the day. Table 7 presents the results of the Mann-Whitney U test comparing the losses of the SOA against the COAs. One-tailed tests suggests that the SOA accumulates significantly higher losses than the COAs during the off-peak and shoulder 1 periods (*p-values*=0.0571). The differences during the peak and shoulder 2 periods are statistically insignificant.[13]■

[13] Interestingly, the amount of experienced losses in the SOA are almost perfectly correlated with the efficiency levels of the plants. The plants that are more costly are more likely to experience bigger short-term losses.

Fig. 5 Total Losses by Treatment for the Last 17 Market Days

Table 7 Mann-Whitney U test on the Total Experienced Losses for the Last 17 Market Days

Period	SOA vs. OCM		SOA vs. PCM	
	$U_{4,4}$	p (one-tailed)	$U_{4,4}$	p (one-tailed)
Off peak	14	0.0571	14	0.0571
Shoulder 1	14	0.0571	12	0.1714
Peak	10	0.3429	8	0.5571
Shoulder 2	9	0.4429	9	0.4429

Finding 6: *Ceteris paribus, outcomes in the SOA are more competitive than in the COAs during the periods when avoidable fixed costs are relevant.*

Support: The SOA always transacts competitive equilibrium amounts while as mentioned above, two COAs come short on number occasions especially during peak periods. Buyer prices in the SOA are significantly lower than the prices in two COAs (Finding 1). Average economic profits of the marginal plants are lower in the SOA than in two COAs during all but shoulder 2 periods (Fig. 6). The Mann-Whitney U test suggests that differences are significant during off-peak, shoulder 1 and peak periods (Table 8: one-tailed *p-values*=0.0571, 0.0286 and

0.0286 respectively for the OCM auction; *p-values*=0.0143 and 0.0571 for off-peak and peak periods in the PCM auction).■

Fig. 6 Average Profits of Marginal Plants by Treatment for the Last 17 Market Days

Table 8 Mann-Whitney U test on Average Profits of Marginal Plants for the Last 17 Market Days

	SOA vs. OCM		SOA vs. PCM	
Period	$U_{4,4}$	p (one-tailed)	$U_{4,4}$	p (one-tailed)
Off peak	13.5	0.0571	16	0.0143
Shoulder 1	15	0.0286	12	0.1714
Peak	15	0.0286	14	0.0571
Shoulder 2	13	0.9000	10	0.6571

5 Conclusions

In a dynamic trading environment that models wholesale electric power markets, the SOA reduces prices to consumers, lowers price volatility and achieves allocative efficiency more quickly than either of two COAs. These gains come at the cost of higher risk of short term losses. The short term losses, however, are rather small relative to the accumulated profits in the described environment.

A frequent critique is if we can learn anything about complex electric power markets from the undergraduates submitting offers in a computer laboratory over the course of 90 minutes. As Fig. 2 shows, prices collapse to the competitive

levels in the SOA. The so-called unprofessional undergraduates leave no room for professional commodity traders to be more competitive. On the other hand, the groups of random students in the COAs succeed in raising prices to levels observed in an environment with structural market power. Hence, undergraduates with no professional experience to a COA mechanism extract almost maximum profits during a 90 minute period of trading. The competitive forces are clearly weaker in two COAs relative to the SOA.

The SOA has less room for strategic behavior. Consequently, the SOA is able to mitigate anti-competitive effects that are present in the COAs, such as the incentive to withhold the lower cost production capacity for the higher demand periods and the ability to sell higher cost units by manipulating the combination of offered fees and prices. It seems that two noteworthy forces exist affecting the performance of COAs. First, the incentive to compete in a COA is weak. And second, a difficulty exists in identifying what offers could displace the offers of competitors since the information about relative structure of two-part priced offers is not public. Shoulder 2 periods are a good example how simpler and more transparent these markets could be if avoidable fixed costs did not exist. The outcomes are relatively competitive in all three auctions since most plants are not eligible for start-up fees in shoulder 2 periods. It is clear that allowing the sellers to recover their variable and avoidable fixed costs separately does not enhance the transparency and competition in the market.

These results implicate that auctions which adopt non-convex optimization mechanisms might be neither necessary nor constructive remedy dealing with non-convex production technologies. After all, many industries with fixed costs successfully operate in the competitive markets with price-per-quantity trades.

For policy makers the lesson is clear: keep market institutions simple. Allowing market participants to reveal more information and trying to make use of that information also creates more opportunities to act strategically. If a way to strike it rich exists within the institutional rules of trading, the market participants find it.

Acknowledgements I thank the National Science Foundation under grants ECS-0323685 and SES-0648937 and the International Foundation for Research in Experimental Economics for financial support of this research. Acknowledgements are extended to Vicki Knoblauch, Peter Luh, Stephen Rassenti, Vernon Smith, Bart Wilson, Artie Zillante and seminar participants at the University of Connecticut, George Mason University, Economic Science Association Meetings, Southern Economic Association Meetings and the State Seminar at the Institute for Regulatory Law and Economics for their comments, Jeffrey Kirchner for his superb programming.

References

Baltaduonis, R. (2007a). Efficiency in Deregulated Electricity Markets: Offer Cost Minimization vs. Payment Cost Minimization Auction. Mechanisms & Policies in Economics. Athens: ATINER, 173-188.

Baltaduonis, R. (2007b). An Experimental Study of Complex-Offer Auctions: Payment Cost Minimization vs. Offer Cost Minimization. University of Connecticut working paper.

Durham, Y., McCabe, K., Olson, M., Rassenti, S., Smith, V. (2004). Oligopoly Competition in Fixed Cost Environments. International Journal of Industrial Organization, 22, 147-162.

Durham, Y., Rassenti, S., Smith, V., Van Boening, M., Wilcox, N. (1996). Can Core Allocations Be Achieved in Avoidable Fixed Cost Environments Using Two-part Price Competition? Annals of Operations Research, 68, 61-88.

Holt, C. (1989). The Exercise of Market Power in Laboratory Experiments. Journal of Law and Economics, 32, S107-S130.

Knoblauch, V. (2005). Can a Newly Proposed Mechanism for Allocating Contracts in U.S. Electricity Wholesale Markets Lead to Lower Prices? A Game Theoretic Analysis. University of Connecticut working paper.

Luh, P., Blankson, W., Chen, Y., Yan, J., Stern, G., Chang, S., Zhao, F. (2005). Payment Cost Minimization Auction for Deregulated Electricity Markets Using Surrogate Optimization. IEEE Transactions on Power Systems.

O'Neill, R., Sotkiewicz, P., Hobbs, B., Rothkopf, M., Stewart Jr., W. (2004). Efficient Market-clearing Prices in Markets with Nonconvexities. European Journal of Operational Research, 164(1), 269-285.

Padhy, N. (2004). Unit Commitment – A Bibliographical Survey. IEEE Transactions on Power Systems, 19(3), 1196-1205.

Rassenti, S., Smith, V., Wilson, B. (2003a). Controlling Market Power and Price Spikes in Electricity Networks: Demand Side-Bidding. Proceedings of the National Academy of Sciences, 100(5), 2998-3003.
Rassenti, S., Smith, V., Wilson, B. (2003b). Discriminatory Price Auctions in Electricity Markets: Low Volatility at the Expense of High Price Levels. Journal of Regulatory Economics, 23(2), 109-123.
Van Boening, M., Wilcox, N. (1996). Avoidable Cost: Ride a Double Auction Roller Coaster. American Economic Review, 86(3), 461-477.
Yan, J., Stern, G. (2002). Simultaneous Optimal Auction and Unit Commitment for Deregulated Electricity Markets. The Electricity Journal, 72-80.

Appendix A An Example of a Simple Wholesale Electricity Market

To highlight the differences of the market-clearing rules in question, consider a three-supplier market examined by Knoblauch (2005) and Baltaduonis (2007a). Say we have an electricity market for one hour. The demand is inelastic and equal to 2 units. Supplier 1 ($S1$) and Supplier 2 ($S2$) are identical. They incur 6 dollars of fixed costs to start up their plants and 93 dollars of variable costs to generate one unit of electricity. Each of them can supply 0, 1 or 2 units of electricity. Supplier 3 ($S3$) has start-up cost of 20 dollars and variable cost of 70 dollars per unit. She can supply 0 or 1 unit of electricity.

For the purpose of this example suppose that all suppliers submit offers that reflect their true production costs. Since the fees are constrained to be zero in the SOA, the suppliers would incur losses unless they recover their fixed costs through the prices. Therefore, in the SOA, the fixed cost can be evenly distributed over the variable cost at the full capacity level of a plant. In this case, $S1$ and $S2$ would submit offers of 96 (=93+6÷2) dollars per unit and $S3$ would submit an offer of 90 (=70+20) dollars per unit. Given these offers the three auctions would generate the following outcomes.

A.1 The OCM Auction

The OCM algorithm minimizes the total offered cost of electricity, as if all selected sellers would be paid their offered prices and fees. Given the offers, an ISO calculates the minimum offered cost in two cases: 1) buying two units from $S1(S2)$ or 2) buying one unit from $S3$ and one unit from $S1(S2)$:

$$\text{Min}\{Price_{1,2}\times 2+Fee_{1,2}\ ,\ Price_3+Fee_3+Price_{1,2}+Fee_{1,2}\},$$

$$\text{Min}\{93\times 2+6,\ 70+20+93+6\}=70+20+93+6=189.$$

The auction chooses to buy 1 unit from $S3$ and 1 unit from $S1(S2)$. After the offers are selected, a uniform MCP is determined as the highest accepted price for that period; the MCP is 93 (=max{70, 93}). All selected sellers receive their individual start-up fees and the uniform MCP for the supplied electricity during that period; the total procurement cost of electricity is 212 (=93×2+20+6). The uniform market price that all buyers pay is 106 [=93+(20+6)÷2)]. Notice that this contract allocation is production efficient since no way exists to generate two units of electricity cheaper than the chosen suppliers do.

A.2 The PCM Auction

The PCM algorithm minimizes the actual procurement cost of electricity, simultaneously determining a MCP as the highest accepted price during that period. An ISO calculates the minimum procurement cost in two cases: 1) buying two units from $S1(S2)$ or 2) buying one unit from $S3$ and one unit from $S1(S2)$:

$$\text{Min}\{Price_{1,2}\times 2+Fee_{1,2}\ ,\ \max\{Price_3\ ,\ Price_{1,2}\}\times 2+Fee_3+Fee_{1,2}\},$$

$$\text{Min}\{93\times 2+6,\ \max\{70,\ 93\}\times 2+20+6\}=93\times 2+6=192.$$

The auction chooses to buy two units from $S1(S2)$. The MCP is 93. As in the OCM auction, the selected sellers receive their individual start-up fees and the uniform MCP for the supplied electricity. Both the total procurement cost and the total production cost are equal to 192 (=93×2+6). The market price for buyers is 96 (=93+6÷2). This contract allocation is not production efficient since $S3$'s plant with relatively lower average total cost is idle.

A.3 The Simple-Offer Auction

The sellers can recover their production costs – both variable and avoidable fixed – only through a uniform MCP in the SOA. Notice that the SOA is a COA where the start-up fees are constrained to be zero. The contract allocations in two COAs are identical when the start-up fees equal zero. Hence, either the OCM or the PCM algorithm could be used for the SOA by simply restricting all start-up fees to zero. In the discussed example, an ISO considers two options: 1) buying two units from $S1(S2)$ or 2) buying one unit from $S3$ and one unit from $S1(S2)$:

$\text{Min}\{Price_{1,2}\times 2\ ,\ Price_3+Price_{1,2}\}$,

$\text{Min}\{96\times 2,\ 90+96\}=90+96=186$.

The auction chooses to buy 1 unit from $S3$ and 1 unit from $S1(S2)$. The MCP is 96. The selected sellers receive the uniform MCP for the supplied electricity. The total procurement cost of electricity is equal to 192 (=96×2). This contract allocation is production efficient, since no way exists to generate two units of electricity cheaper than the chosen suppliers do. However, this outcome is problematic because $S1(S2)$ is not able to recover all production costs and incurs a

loss of -3 (=96-93-6). Since this outcome can not be sustained in the long run, *S*1(*S*2) would be forced to increase the offer in order to recover the fixed cost even when she sells only one unit of energy. The minimum sustainable offer is 99 dollars per unit. In this case, the outcome is as follows:

Min{99×2, 90+99}=90+99=189.

The auction chooses to buy one unit from *S*3 and one unit from *S*1(*S*2). The MCP for both buyers and sellers is 99. The total procurement cost of electricity is equal to 198 (=99×2). The contract allocation is production efficient.

In the presented example, given the assumption of truthful production cost revelation, the PCM auction produces the lowest procurement cost of electricity. It slightly outperforms the SOA and more significantly the OCM auction. On the other hand, the PCM auction is the only one to yield a production inefficient allocation. The SOA case shows that the sellers might face a risk of short-term losses.

Appendix B Experimental Instructions

<page 1>

Welcome

This is an experiment in the economics of decision-making. If you read the instructions carefully and make good decisions, you may earn a considerable amount of money that will be paid to you in CASH at the end of the experiment.

The experiment will take place through the computer terminals at which you are seated. If you have any questions at any time, please raise your hand and a monitor will come to assist you.

In this experiment, owners of plants sell an identical product to a computer buyer every day. Each day lasts 75 seconds. You are an **owner** of #yourNumberOfPlants# plants. There are #numberOfSellers# sellers and #numberOfPlants# plants including yours. Each seller owns between 1 and 4 plants.

<page 2>

Each day is divided into 4 quarters. Each quarter is represented by a line in the table at the top of your screen. The computer will purchase varying quantities of the product over the course of a day: Low, Medium, High and Medium amounts.

Sellers submit offers to sell. An offer indicates the prices and quantities of the product that you are willing to sell during the course of the following day. All quantities are measured in number of units.

<page 3 OCM and PCM>

You as a seller are able to decide:

Price/unit is the price per unit you are willing to sell at during that quarter from that plant. This is the minimum price at which you are willing to sell. The actual market price may be higher depending on the demand of the product. Each seller receives the same market price for sold units during the quarter. The market price

is the highest accepted Price/unit among all of the sellers. If you sell the product you also incur a cost per unit sold. This cost is listed on the right side under the table and must be paid for each unit you sell.

Start-Up Fee is a fee that is paid to you for turning on your plant. The fee is paid to you only if the plant was not operating during the previous quarter. When your plant is turned on, you also must pay the start-up cost, which is listed on the right side under the table.

You will be able to make this decision for each quarter of the upcoming day for each plant that you have.

<page 3 SOA>

You as a seller are able to decide:

Price/unit is the price per unit you are willing to sell at during that quarter from that plant. This is the minimum price at which you are willing to sell. The actual market price may be higher depending on the demand of the product. Each seller receives the same market price for sold units during the quarter. The market price is the highest accepted Price/unit among all of the sellers. If you sell the product you also incur a cost per unit sold. This cost is listed on the right side under the table and must be paid for each unit you sell.

You will be able to make this decision for each quarter of the upcoming day for each plant that you have.

<page 4 OCM and PCM>

To switch between plants click on the tabs at the top of your screen. To enter the values select the appropriate cell in the table and double click.

Some offer values are automatically filled in for you:

Min Qty is the minimum number of units you are willing to sell during that quarter from that plant. Min Qty must be ≥ Minimum Capacity, which is specified under the table. This will be filled with that plant's Minimum Capacity.

Max Qty is the maximum number of units you are willing to sell during that quarter from that plant. Max Qty must be ≤ Maximum Capacity, which is specified under the table. Max Qty must also be ≥ Min Qty. This will be filled with that plant's Maximum Capacity.

<page 4 SOA>

To switch between plants click on the tabs at the top of your screen. To enter the values select the appropriate cell in the table and double click.

Some offer values are automatically filled in for you:

Min Qty is the minimum number of units you are willing to sell during that quarter from that plant. Min Qty must be ≥ Minimum Capacity, which is specified under the table. This will be filled with that plant's Minimum Capacity.

Max Qty is the maximum number of units you are willing to sell during that

quarter from that plant. Max Qty must be ≤ Maximum Capacity, which is specified under the table. Max Qty must also be ≥ Min Qty. This will be filled with that plant's Maximum Capacity.

When your plant is turned on, you also must pay the start-up cost, which is listed on the right side under the table. You will not receive the **Start-Up Fee** for turning on your plant.

<page 5 PCM>

Offers are sent to the computerized market coordinator when you click the Submit button or when the day is over. Your offer from the previous day will be automatically submitted for you if you choose not to make any changes during the course of a day.

The computerized market coordinator accepts those offers that satisfy the market demand during the day at the *lowest total procurement cost*, simultaneously determining the market price as the highest accepted Price/unit for that quarter.

If your offer has not been accepted, it means that other offers were able to satisfy the market demand at a lower or equal cost. The results are displayed on the right side of the table; you may need to scroll to the right to see them. Once you have reviewed the results of the previous day enter your offers for the next day for each plant and submit.

The right side of the table is filled in after everyone has submitted their offers. Your profit during each quarter of a day is:

(Units Sold × market price + Start-Up Fees collected) – (Units Sold × Cost/unit + Start-Up Costs incurred)

<page 5 OCM>

Offers are sent to the computerized market coordinator when you click the Submit button or when the day is over. Your offer from the previous day will be automatically submitted for you if you choose not to make any changes during the course of a day.

The computerized market coordinator accepts those offers that satisfy the market demand during the day at the *lowest total offered cost*. After the offers are selected, the market price is determined as the highest accepted Price/unit for that quarter.

If your offer has not been accepted, it means that other offers were able to satisfy the market demand at a lower or equal cost. The results are displayed on the right side of the table; you may need to scroll to the right to see them. Once you have reviewed the results of the previous day enter your offers for the next day for each plant and submit.

The right side of the table is filled in after everyone has submitted their offers.

Your profit during each quarter of a day is:

(Units Sold × market price + Start-Up Fees collected) – (Units Sold × Cost/unit + Start-Up Costs incurred)

<page 5 SOA>

Offers are sent to the computerized market coordinator when you click the Submit button or when the day is over. Your offer from the previous day will be automatically submitted for you if you choose not to make any changes during the course of a day.

The computerized market coordinator orders offered Prices/unit from lowest to highest for each quarter of the day. Market's bids to buy the product are ordered from highest to lowest. These two sorted lists will cross. The offered Price/unit where these lists cross becomes the market price during the quarter. The market coordinator accepts all offers with Prices/unit lower than the market price. If there is more than one offer exactly equal to the market price, then as many of those offers will be accepted as it is enough to satisfy the market demand during that quarter of the day.

If your offer has not been accepted, it means that other offers were able to satisfy the market demand at a lower or equal cost. The results are displayed on the right side of the table; you may need to scroll to the right to see them. Once you have reviewed the results of the previous day enter your offers for the next day for each plant and submit.

The right side of the table is filled in after everyone has submitted their offers.
Your profit during each quarter of a day is:
(Units Sold × market price) – (Units Sold × Cost/unit + Start-Up Costs incurred)

<page 6>

A history of the prices from the past 10 days and the sold quantities during each

quarter of the last day are displayed in the bottom portion of your screen.

Information about all plants (including yours) is available to all sellers by clicking on the Technology and costs button.

Plants are restarted at the beginning of each day, meaning that during the first quarter of each day you receive your start-up fee and incur the start-up cost if you sell the product.

At the end of today's session, your 'computer dollars' will be converted into cash at a rate of #exchangeRate# computer dollars to US$1. If you have any questions please raise your hand. Press Start when you are ready to begin.

Even if you decide to keep your offer from the previous day, click the Submit button. The experiment will advance to the next day after everyone has clicked on the Submit button.

Appendix C Sample Screen Shot

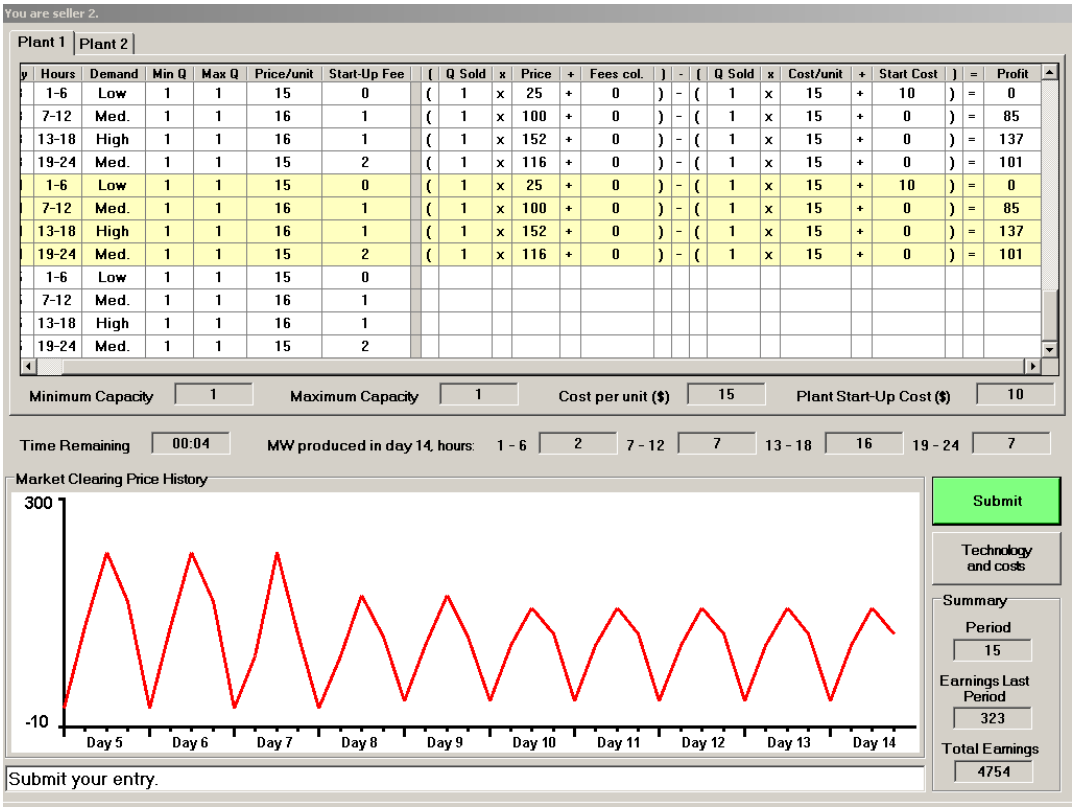


Fig. 7 Sample Screen Shot.