

## Contents

Preface to the Fourth Edition xvii

1 **Introduction**
   - The Rationale for Regulation and Antitrust Policies 2
   - Antitrust Regulation 3
   - The Changing Character of Antitrust Issues 4
   - Reasoning behind Antitrust Regulations 5
   - Economic Regulation 5
   - Development of Economic Regulation 6
   - Factors in Setting Rate Regulations 6
   - Health, Safety, and Environmental Regulation 8
   - Role of the Courts 9
   - Criteria for Assessment 9
   - Questions and Problems 11
   - Recommended Reading 11
   - Appendix 11

2 **The Making of a Regulation** 13
   - State versus Federal Regulation: The Federalism Debate 14
     - Advantages of Federalism 15
     - Advantages of National Regulations 16
     - Product Labeling Example 17
     - The Overlap of State and Federal Regulations 18
   - The Character of the Rulemaking Process 19
     - The Chronology of New Regulations 19
     - Nature of the Regulatory Oversight Process 24
       - The Nixon and Ford Administrations 24
       - The Carter Administration 25
       - The Reagan Administration 27
       - The Bush Administration 28
       - The Clinton Administration 28
       - The George W. Bush Administration 28
     - Regulatory Reform Legislation 28
     - Benefit-Cost Analysis 30
     - Discounting Deferred Effects 33
     - Present Value 34
# Contents

<table>
<thead>
<tr>
<th>Chapter/Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Criteria Applied in the Oversight Process</td>
<td>36</td>
</tr>
<tr>
<td>Regulatory Success Stories</td>
<td>36</td>
</tr>
<tr>
<td>Promotion of Cost-Effective Regulation</td>
<td>37</td>
</tr>
<tr>
<td>Distortion of Benefit and Cost Estimates</td>
<td>38</td>
</tr>
<tr>
<td>The Regulatory Role of Price and Quality</td>
<td>39</td>
</tr>
<tr>
<td>The Impact of the Oversight Process</td>
<td>39</td>
</tr>
<tr>
<td>The Cost of Regulation</td>
<td>40</td>
</tr>
<tr>
<td>Other Measures of the Size of Regulation</td>
<td>40</td>
</tr>
<tr>
<td>The Character of Regulatory Oversight Actions</td>
<td>45</td>
</tr>
<tr>
<td>What Do Regulators Maximize?</td>
<td>48</td>
</tr>
<tr>
<td>The Capture Theory</td>
<td>48</td>
</tr>
<tr>
<td>Other Theories of Influence Patterns</td>
<td>49</td>
</tr>
<tr>
<td>Comprehensive Models of Regulatory Objectives</td>
<td>49</td>
</tr>
<tr>
<td>Conclusion</td>
<td>51</td>
</tr>
<tr>
<td>Questions and Problems</td>
<td>51</td>
</tr>
<tr>
<td>Appendix: Trends in Regulatory Agency Budgets and Staff</td>
<td>52</td>
</tr>
</tbody>
</table>

## I ANTITRUST

### 3 Introduction to Antitrust

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Industrial Organization</td>
<td>61</td>
</tr>
<tr>
<td>Structure</td>
<td>62</td>
</tr>
<tr>
<td>Conduct</td>
<td>63</td>
</tr>
<tr>
<td>Performance</td>
<td>66</td>
</tr>
<tr>
<td>Government</td>
<td>68</td>
</tr>
<tr>
<td>Antitrust</td>
<td>69</td>
</tr>
<tr>
<td>Federal Antitrust Laws</td>
<td>69</td>
</tr>
<tr>
<td>Enforcement and Remedies</td>
<td>71</td>
</tr>
<tr>
<td>Exemptions from Antitrust</td>
<td>75</td>
</tr>
<tr>
<td>Summary and Overview of Part I</td>
<td>76</td>
</tr>
<tr>
<td>Appendix: Antitrust Statutes</td>
<td>76</td>
</tr>
<tr>
<td>Sherman Act</td>
<td>76</td>
</tr>
<tr>
<td>Clayton Act</td>
<td>77</td>
</tr>
<tr>
<td>Federal Trade Commission Act</td>
<td>78</td>
</tr>
</tbody>
</table>
4  Efficiency and Technical Progress 79
Economic Efficiency 79
  Partial Equilibrium Welfare Tools 80
  Monopoly-versus-Competition Example 82
  Oil Industry Application 84
Some Complications 85
X-Inefficiency 88
Monopoly-Induced Waste 89
Estimates of the Welfare Loss from Monopoly 90
Technical Progress 93
Importance of Technological Change 93
A Model of R & D Rivalry 95
Summary 98
Questions and Problems 99

5  Oligopoly, Collusion, and Antitrust 101
Game Theory 101
  Example 1: Advertising Competition 101
  Example 2: Compatibility of Standards 103
  The Strategic Form of a Game 104
  Nash Equilibrium 105
Oligopoly Theory 106
  The Cournot Solution 106
  Other Models of Oligopoly 112
  Product Differentiation 113
Collusion 116
  A Theory of Collusion 117
  Challenges to Collusion 121
  Collusion in Practice 128
Antitrust Law and Policy toward Price Fixing 134
  Economic Analysis of Legal Categories 135
  Per Se Rule Cases 137
  Tacit Collusion 140
  Enforcement Policy 144
Summary 150
Questions and Problems 151
Appendix
  Game Theory: Formal Definitions

6 Market Structure and Strategic Competition 155
Market Structure 155
  Concentration 155
  Scale Economies 156
  Entry Conditions 164
Dominant Firm Theory 174
  Static Analysis 174
  Dynamic Analysis: Limit Pricing 177
Strategic Competition 182
  Limit Pricing 183
  Investment in Cost-Reducing Capital 190
  Raising Rivals’ Costs 194
Preemption and Brand Proliferation 194
Summary 198
Questions and Problems 198

7 Mergers 203
Antitrust Laws and Merger Trends 204
Reasons for Mergers 207
  Monopoly 207
  Economies 207
  Reducing Management Inefficiencies 208
Horizontal Mergers 210
  Benefits and Costs 210
  Effects of Airline Mergers 219
  Cases 220
  U.S. Department of Justice Merger Guidelines 225
Conglomerate Mergers 229
  Potential Benefits 229
  Anticompetitive Effects and Cases 230
Summary 232
Questions and Problems 233
# 8 Vertical Mergers and Vertical Restraints

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vertical Mergers</td>
<td>235</td>
</tr>
<tr>
<td>Benefits</td>
<td>236</td>
</tr>
<tr>
<td>Anticompetitive Effects</td>
<td>237</td>
</tr>
<tr>
<td>Commitment and the Restoration of Market Power</td>
<td>241</td>
</tr>
<tr>
<td>Raising Rivals’ Costs</td>
<td>246</td>
</tr>
<tr>
<td>Antitrust Law and Policy</td>
<td>248</td>
</tr>
<tr>
<td>Historical Development</td>
<td>253</td>
</tr>
<tr>
<td>Time Warner and Turner</td>
<td>254</td>
</tr>
<tr>
<td>Vertical Restraints</td>
<td>255</td>
</tr>
<tr>
<td>Exclusive Dealing</td>
<td>257</td>
</tr>
<tr>
<td>Antitrust Law and Policy</td>
<td>258</td>
</tr>
<tr>
<td>Tying</td>
<td>263</td>
</tr>
<tr>
<td>Modern Theories of Leveraging</td>
<td>266</td>
</tr>
<tr>
<td>Manufacturer-Retailer Restraints</td>
<td>275</td>
</tr>
<tr>
<td>Summary</td>
<td>282</td>
</tr>
<tr>
<td>Questions and Problems</td>
<td>288</td>
</tr>
</tbody>
</table>

# 9 Monopolization and Price Discrimination

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Establishing Monopolization Claims</td>
<td>293</td>
</tr>
<tr>
<td>Measuring Monopoly Power</td>
<td>294</td>
</tr>
<tr>
<td>Assessing Intent to Monopolize</td>
<td>294</td>
</tr>
<tr>
<td>Development of Antitrust Case Law</td>
<td>298</td>
</tr>
<tr>
<td>1890–1940: Standard Oil and United States Steel</td>
<td>299</td>
</tr>
<tr>
<td>1940–1970: Alcoa and United Shoe Machinery</td>
<td>300</td>
</tr>
<tr>
<td>1970 to Present: Kodak, IBM, Microsoft, and Others</td>
<td>303</td>
</tr>
<tr>
<td>Predatory Pricing</td>
<td>305</td>
</tr>
<tr>
<td>Theories of Predatory Pricing</td>
<td>309</td>
</tr>
<tr>
<td>Efficiency Rationales</td>
<td>316</td>
</tr>
<tr>
<td>Antitrust Policy</td>
<td>317</td>
</tr>
<tr>
<td>The Areeda-Turner Rule and Other Single-Parameter Rules</td>
<td>317</td>
</tr>
<tr>
<td>The Brooke Case and the Two-Tier Rule</td>
<td>319</td>
</tr>
<tr>
<td>Recent Developments</td>
<td>321</td>
</tr>
<tr>
<td>Refusal to Deal and the Essential Facilities Doctrine</td>
<td>322</td>
</tr>
<tr>
<td>Essential Facilities Doctrine</td>
<td>323</td>
</tr>
<tr>
<td>Intellectual Property Rights</td>
<td>324</td>
</tr>
<tr>
<td>Kodak and Monopoly Power in Aftermarkets</td>
<td>326</td>
</tr>
</tbody>
</table>
### Contents

<table>
<thead>
<tr>
<th>topic</th>
<th>page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft Case</td>
<td>332</td>
</tr>
<tr>
<td>Network Externalities</td>
<td>332</td>
</tr>
<tr>
<td>Antitrust Case</td>
<td>336</td>
</tr>
<tr>
<td>Tying and Monopolization of the Browser Market</td>
<td>338</td>
</tr>
<tr>
<td>Maintenance of Monopoly in the Operating Systems Market</td>
<td>339</td>
</tr>
<tr>
<td>Remedies and Harm</td>
<td>342</td>
</tr>
<tr>
<td>Price Discrimination and the Robinson-Patman Act</td>
<td>343</td>
</tr>
<tr>
<td>Systematic Discrimination</td>
<td>344</td>
</tr>
<tr>
<td>Unsystematic Discrimination</td>
<td>349</td>
</tr>
<tr>
<td>Cases</td>
<td>350</td>
</tr>
<tr>
<td>Summary</td>
<td>352</td>
</tr>
<tr>
<td>Questions and Problems</td>
<td>352</td>
</tr>
</tbody>
</table>

## II ECONOMIC REGULATION

### 10 Introduction to Economic Regulation

<table>
<thead>
<tr>
<th>topic</th>
<th>page</th>
</tr>
</thead>
<tbody>
<tr>
<td>What Is Economic Regulation?</td>
<td>357</td>
</tr>
<tr>
<td>Instruments of Regulation</td>
<td>358</td>
</tr>
<tr>
<td>Control of Price</td>
<td>358</td>
</tr>
<tr>
<td>Control of Quantity</td>
<td>359</td>
</tr>
<tr>
<td>Control of Entry and Exit</td>
<td>359</td>
</tr>
<tr>
<td>Control of Other Variables</td>
<td>360</td>
</tr>
<tr>
<td>Brief History of Economic Regulation</td>
<td>362</td>
</tr>
<tr>
<td>Formative Stages</td>
<td>362</td>
</tr>
<tr>
<td>Trends in Regulation</td>
<td>364</td>
</tr>
<tr>
<td>The Regulatory Process</td>
<td>369</td>
</tr>
<tr>
<td>Overview of the Regulatory Process</td>
<td>369</td>
</tr>
<tr>
<td>Regulatory Legislation</td>
<td>371</td>
</tr>
<tr>
<td>Independent Regulatory Commissions</td>
<td>371</td>
</tr>
<tr>
<td>Regulatory Procedures</td>
<td>373</td>
</tr>
<tr>
<td>The Theory of Regulation</td>
<td>375</td>
</tr>
<tr>
<td>Normative Analysis as a Positive Theory</td>
<td>376</td>
</tr>
<tr>
<td>Capture Theory</td>
<td>379</td>
</tr>
<tr>
<td>Economic Theory of Regulation</td>
<td>380</td>
</tr>
<tr>
<td>Testing Theories of Regulation</td>
<td>392</td>
</tr>
<tr>
<td>Summary and Overview of Part II</td>
<td>396</td>
</tr>
</tbody>
</table>
Contents

Appendix
   A Theory of Interest Group Competition 397
Questions and Problems 399

11 Theory of Natural Monopoly 401
The Natural Monopoly Problem 401
   Permanent and Temporary Natural Monopoly 402
   Subadditivity and Multiproduct Monopoly 404
Alternative Policy Solutions 408
   Ideal Pricing 409
   Franchise Bidding 421
   Actual Solutions 421
Summary 423
Appendix 423
   The Troublesome Case of a Natural Monopoly 423
Questions and Problems 425

12 Natural Monopoly Regulation and Electric Power 429
Traditional Rate-of-Return Regulation 430
   The Rate Case 431
   Averch-Johnson Effect 433
Incentive Regulation 436
   Performance Standards 436
   Earnings Sharings 437
   Price Caps 439
   Yardstick Regulation 442
Rate Structure 443
   FDC Pricing 443
   Undue Discrimination 445
Peak-Load Pricing 447
   Costs of Power Production 447
   Peak-Load Pricing Model 449
Regulation and Restructuring of Electric Power 453
   Historical, Technological, and Regulatory Background 453
   Overview of Recent Legislation 455
   Restructuring in California 456
Summary 461
Questions and Problems 462
### Contents

#### 13 Franchise Bidding and Cable Television

- Theory of Franchise Bidding 465
- Competition at the Bidding Stage 467
- Contractual Arrangements for the Postbidding Stage 475
- Assessment of Franchise Bidding 478
- Cable Television 479
  - Historical/Regulatory Background 480
  - Cable Television as a Natural Monopoly 481
  - Franchising Process 485
  - Assessment of Franchise Bidding 487
- Rate Regulation 492
- Is There a Role for Government Intervention? 497
- Summary 499
- Questions and Problems 500

#### 14 Public Enterprise

- General Background 503
- Positive Theory of Public Enterprise 505
  - Managerial Model of a Firm 506
  - Managerial Model of a Private Enterprise 507
  - Managerial Model of a Public Enterprise 508
- Comparison of Public and Private Enterprise 510
- Municipal Electric Utilities 512
  - Pricing Behavior 512
  - Allocative Efficiency Comparison 513
  - Productive Efficiency Comparison 515
  - Assessment of Private versus Public Utilities 515
- Airlines 516
- Privatization 517
- Summary 521
- Questions and Problems 522

#### 15 Dynamic Issues in Natural Monopoly Regulation: Telecommunications

- Transformation of a Natural Monopoly 523
  - Basis for Natural Monopoly Regulation 524
  - Sources of Natural Monopoly Transformation 527
- Regulatory Response 529
- Intercity Telecommunications Market 534
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Telecommunications Act of 1996</strong></td>
<td>544</td>
</tr>
<tr>
<td><strong>Separation of Regulated Monopolies and Competitive Markets</strong></td>
<td>546</td>
</tr>
<tr>
<td>Benefits and Costs of Separation</td>
<td>547</td>
</tr>
<tr>
<td>Breakup of AT&amp;T</td>
<td>550</td>
</tr>
<tr>
<td><strong>Summary</strong></td>
<td>551</td>
</tr>
<tr>
<td><strong>Questions and Problems</strong></td>
<td>553</td>
</tr>
<tr>
<td><strong>16 The Regulation of Potentially Competitive Markets: Theory and Estimation Methods</strong></td>
<td>555</td>
</tr>
<tr>
<td>Theory of Price and Entry/Exit Regulation</td>
<td>556</td>
</tr>
<tr>
<td>Direct Effects of Price and Entry/Exit Regulation: The Competitive Model</td>
<td>556</td>
</tr>
<tr>
<td>Direct Effects of Price and Entry/Exit Regulation: The Imperfectly Competitive Model</td>
<td>560</td>
</tr>
<tr>
<td>Indirect Effects of Price and Entry Regulation</td>
<td>564</td>
</tr>
<tr>
<td>Some Indirect Effects of Price and Exit Regulation</td>
<td>567</td>
</tr>
<tr>
<td>Regulation and Innovation</td>
<td>568</td>
</tr>
<tr>
<td>Methods for Estimating the Effects of Regulation</td>
<td>572</td>
</tr>
<tr>
<td>Overview of Estimation Methods</td>
<td>572</td>
</tr>
<tr>
<td>Intertemporal Approach</td>
<td>572</td>
</tr>
<tr>
<td>Application: New York Stock Exchange</td>
<td>574</td>
</tr>
<tr>
<td>Intermarket Approach</td>
<td>575</td>
</tr>
<tr>
<td>Application: Advertising of Eyeglasses</td>
<td>576</td>
</tr>
<tr>
<td>Application: 44 Liquormart Decision</td>
<td>578</td>
</tr>
<tr>
<td>Counterfactual Approach</td>
<td>578</td>
</tr>
<tr>
<td>Application: State Usury Laws</td>
<td>579</td>
</tr>
<tr>
<td>Measuring the Return to Price and Entry Restrictions: Taxicab Regulation</td>
<td>583</td>
</tr>
<tr>
<td><strong>Summary</strong></td>
<td>585</td>
</tr>
<tr>
<td><strong>Questions and Problems</strong></td>
<td>587</td>
</tr>
<tr>
<td><strong>17 Economic Regulation of Transportation: Surface Freight and Airlines</strong></td>
<td>589</td>
</tr>
<tr>
<td>Transportation Industry</td>
<td>589</td>
</tr>
<tr>
<td>Surface Freight Transportation</td>
<td>591</td>
</tr>
<tr>
<td>Regulatory History</td>
<td>591</td>
</tr>
<tr>
<td>Description of Regulatory Practices</td>
<td>595</td>
</tr>
<tr>
<td>Effects of Regulation</td>
<td>597</td>
</tr>
</tbody>
</table>
### Contents

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airlines</td>
<td>609</td>
</tr>
<tr>
<td>Regulatory History</td>
<td>609</td>
</tr>
<tr>
<td>Description of Regulatory Practices</td>
<td>611</td>
</tr>
<tr>
<td>Effects of Regulation</td>
<td>612</td>
</tr>
<tr>
<td>Competition and Antitrust Policy after Deregulation</td>
<td>625</td>
</tr>
<tr>
<td>Lessons from Regulation and Deregulation</td>
<td>638</td>
</tr>
<tr>
<td>Summary</td>
<td>639</td>
</tr>
<tr>
<td>Questions and Problems</td>
<td>640</td>
</tr>
</tbody>
</table>

#### 18 Economic Regulation of Energy: Crude Oil and Natural Gas

- The Theory of Price Ceilings                                      | 642  |
- Price and Quantity Regulation of the Crude Oil Industry           | 646  |
  - Regulatory History                                              | 648  |
- Oil Prorationing                                                    | 651  |
  - Regulatory Practices                                             | 651  |
  - Rationale for Prorationing                                       | 651  |
  - Solutions to the Common Pool Problem                             | 656  |
  - Effects of Prorationing                                          | 657  |
- Mandatory Oil Import Program                                       | 659  |
  - Regulatory Practices                                             | 659  |
  - Effects of Regulation                                            | 659  |
- Crude Oil Price Controls                                           | 661  |
  - Regulatory Practices                                             | 661  |
  - Effects of Price Regulation                                      | 663  |
- Price Regulation of the Natural Gas Industry                       | 671  |
  - Regulatory History                                               | 672  |
  - Regulatory Practices                                             | 673  |
  - Effects of Price Regulation                                      | 675  |
- Transition from Regulation to Markets in the Transmission of Natural Gas | 683  |
| Summary                                                              | 685  |
| Questions and Problems                                               | 686  |

#### III HEALTH, SAFETY, AND ENVIRONMENTAL REGULATION

#### 19 Introduction: The Emergence of Health, Safety, and Environmental Regulation

- Risk in Perspective                                                | 691  |
  - Measuring Mortality Risks                                         | 694  |
The Infeasibility of a No-Risk Society 695
   Homeland Security 696
   Wealth and Risk 699
   Irrationality and Biases in Risk Perception 700
Policy Evaluation 703
   Regulatory Standards 704
   Benefit-Cost Analysis 704
   The Role of Heterogeneity 705
Uncertainty and Conservatism 707
   The Role of Risk Ambiguity 707
The Role of Political Factors 709
   Economic Models of Environmental Policies 709
Voting Patterns 711
Summary and Overview of Part III 713
Questions and Problems 714
Recommended Reading 715

20 Valuing Life and Other Nonmonetary Benefits 717
Policy Evaluation Principles 718
   Willingness-to-Pay versus Other Approaches 720
Variations in the Value of Statistical Life 723
The Labor Market Model 725
Empirical Estimates of the Value of Life 730
Value of Risks to Life for Regulatory Policies 731
Survey Approaches to Valuing Policy Effects 736
   Valuation of Air Quality 738
   Exploratory Nature of the Survey Approach 738
Sensitivity Analysis and Cost-Effectiveness 739
Risk-Risk Analysis 740
Establishing Prices for Health, Safety, and Environmental Regulation 742
Questions and Problems 742

21 Environmental Regulation 745
The Coase Theorem for Externalities 746
   The Coase Theorem as a Bargaining Game 747
   A Pollution Example 748
Long-Run Efficiency Concerns 750
Transaction Costs and Other Problems 750
Contents

22 Product Safety

Emergence of Product Safety Regulations
Current Safety Decisions
Changing Emphasis of Product Regulation
Premanufacturing Screening: The Case of Pharmaceuticals
Weighing the Significance of Side Effects
Drug Approval Strategies
The Behavioral Response to Product Safety Regulation
Consumer’s Potential for Muting Safety Device Benefits
The Costs of Product Safety Regulation: The Automobile Industry Case
Contents

Trends in Motor Vehicle and Home Accident Deaths 808
  Accident Rate Influences 809
  The Decline of Accident Rates 809
The Rise of Product Liability 810
  The Negligence Standard 812
  The Strict Liability Standard 813
  The Ford Pinto Case 813
  Escalation of Damages 815
Risk Information and Hazard Warnings 816
  Self-Certification of Safe Products 817
  Government Determination of Safety 818
Alternatives to Direct Command and Control Regulation 818
Regulation Through Litigation 820
  Breast Implant Litigation and Regulation 821
The Future of Product Safety Policy 823
Questions and Problems 825

23 Regulation of Workplace Health and Safety 827
  The Potential for Inefficiencies 829
How Markets Can Promote Safety 829
Compensating Wage Differential Theory 831
Risk Information 833
On-the-Job Experience and Worker Quit Rates 835
Inadequacies in the Market 836
  Externalities 838
OSHA’s Regulatory Approach 838
Setting OSHA Standard Levels 839
  The Nature of OSHA Standards 841
  The Reform of OSHA Standards 843
Regulatory Reform Initiatives 843
Changes in OSHA Standards 843
OSHA’s Enforcement Strategy 847
Inspection Policies 848
Trivial Violations 849
OSHA Penalties 849
Enforcement Targeting 850
Contents

24 Patents and Pharmaceuticals

Economics of Invention and Patents

Background on Patents

Incentives to Invent: Monopoly versus Competition

Welfare Analysis of Patents

Pharmaceuticals and the Role of Patents

Industry Structure

The 1984 Drug Price Competition and Patent Restoration Act

Other Policies That Affect R&D Incentives

Summary

Questions and Problems
One of the most exciting areas of economic policy is government regulation and antitrust. These efforts affect virtually all aspects of our lives, ranging from the food we eat to the prices we pay. This policy area has undergone dramatic changes in the past three decades. The traditional topics in this area would have included such issues as setting appropriate trucking rates as well as conventional antitrust issues. However, in many areas of economic regulation there has been substantial deregulation as market forces in a larger and more competitive economy have been given more rein. New areas of economic regulation have developed, such as those pertaining to the regulation of cable television rates and homeland security. In addition, there has been an entirely new wave of government regulation, chiefly relating to the environment and safety, which involves a commitment of economic resources that has continued to escalate.

The vibrancy of regulatory and antitrust policy is reflected in recent economic events. The settlement of the states’ lawsuits against the cigarette industry for more than $200 billion has spawned other litigation efforts that have blurred the boundary between regulation and litigation. Environmental policies have continued to dominate in terms of the economic costs of regulation, recently raising difficult ethical issues such as whether protecting the lives of the elderly is worth as much as protecting the lives of the young. Economic deregulation in areas such as cable television, natural gas, intrastate trucking, electric power, and telecommunications has continued the pattern of deregulation, sometimes generating profound changes. The aftereffects of the Telecommunications Act of 1996 are still being felt in markets ranging from cable television to long-distance telephone service. In spite of the debacle in California in 2000, state electric power markets continue to be a hotbed of regulatory restructuring. In enforcing our antitrust laws, the U.S. Department of Justice has aggressively pursued price fixing and has levied fines in the hundreds of millions of dollars and imposed prison sentences for many high-level executives. It waged a successful battle against Microsoft, from which the latter emerged mostly unscathed. At the same time, the past decade has witnessed megamergers, including Time Warner/America Online, Bell Atlantic/GTE, Daimler-Benz/Chrysler, Travelers/Citicorp, BPE/Amoco, and Exxon/Mobil.

The emerging character of antitrust policies and regulation has been accompanied by an intellectually vibrant economic literature. Using frontier tools such as game theory, economists have developed new theories to characterize firm behavior and to assess which market contexts warrant government intervention. Our view of which situations of apparently excessive market power warrant government interference has changed dramatically.

Economists have also developed new methodologies to deal with emerging health, safety, and environmental regulations. These regulatory efforts were largely nonexistent two decades ago, and the economic literature addressing these issues was similarly undeveloped. In this book we will attempt to convey the general character of the principles guiding economic regulation in this and other areas, as well as the most salient aspects of these policies. While the emphasis is on U.S. regulatory policies, the principles are quite general.
The traditional emphasis of economics textbooks on business and government is on the character of regulations and antitrust policies. This treatment is built around the question, What are these policy mechanisms, and how do they operate?

The orientation of *Economics of Regulation and Antitrust* is quite different. Rather than start with the institutional aspect of regulatory and antitrust policies, we begin with the economic issues at stake. What particular market failures provide a rationale for government intervention? How can economic theory illuminate the character of market operation, the role for government action, and the appropriate form of government action? What do formal empirical analyses of economic behavior and the effects of government intervention indicate about the direction that this intervention should take? To provide the most up-to-date answers to these important questions, we base our analysis on new developments in economic theory and empirical analysis that have been specifically devised to further understanding of regulations and antitrust policies.

Because this has been a fertile area of economic research for several decades, a large body of economic reasoning can be brought to bear in analyzing these issues. *Economics of Regulation and Antitrust* is the only economics textbook whose focus derives from the insights that economic reasoning can provide in analyzing regulatory and antitrust issues. This approach contrasts with previous treatments, which concentrate on the character of these policies and relegate the economic issues to a minor role.

This approach, which we established in earlier editions, has been carried forward in this edition as well. New topics have been added. In chapter 2 and wherever possible throughout the book, we have updated the statistics regarding the role of government regulation in the economy. Chapter 2 also includes a new discussion of regulatory oversight during the George W. Bush administration, as well as new information on regulatory cost trends updated to the current century.

Part I has undergone a major revision to reflect advance in theory of antitrust and the major antitrust cases of recent years. A section has been added to the chapter on price fixing that focuses on enforcement and recent policy innovations with the revision of the corporate leniency program and federal sentencing guidelines. We review the economics behind the corporate leniency program—how does it work in catching price fixers?—and current practice for determining penalties. Though the authorities have been active in fighting price fixing, the most controversial and significant cases have dealt with monopolization. To take account of these cases and to encompass new understanding about monopolization practices, chapters 8 and 9 have been substantially revised. Using simple examples, modern game-theoretic analyses of raising rivals' cost, tying, and exclusive dealing are reviewed and related to the merger of Time Warner and the Turner Broadcasting System and to the exclusionary practices used by Visa and MasterCard. Then, in chapter 9, we provide an analysis of how predatory pricing can work and flesh out the implications of recent judicial decisions. Attention is also given to the exclusionary practice known as “refusal to deal,” with a focus on its inter-
action with intellectual property rights—an issue that arose in the Federal Trade Com-
mmission’s case against Intel. Our coverage of Microsoft is significantly extended to include an
analysis of the economics of network externalities. Finally, we have added an in-depth exam-
ination of the economics of aftermarket and its relation to the historic Kodak decision by
the U.S. Supreme Court.

Our emphasis on economic principles in no way implies a neglect of the pertinent institutional
features. This text includes extensive case studies of major areas of regulation and
antitrust policy, including entire chapters devoted to such issues as government merger
policies, cable television regulation, and transportation regulation. Indeed, this book is unique
in its extensive coverage of several of these topics, as well as issues such as the role of the
White House regulatory oversight process. Although we discuss essential aspects of these
regulations and their performance, our intent is not to provide students with a list of case
names, government rules, and other institutional details. Rather, we hope to provide students
not only with insights pertinent today but also with the economic tools to analyze the impli-
cations of regulations and antitrust policies a decade from now. Future policies may have a
quite different structure from those currently in place, and it is the economic framework we
use to approach these issues that will be of lasting value.

Part II, dealing with economic regulation, updates our earlier coverage of the restructuring
of the telecommunications and electric power industries. An analysis of what went wrong
in the California energy market in 2000–2001 is provided, along with more extensive cov-
erage of how energy markets are being deregulated at the state level. Responding to an evolv-
ing technological landscape, the Telecommunications Act of 1996 was the first major piece
of legislation in this industry since the Federal Communications Commission was formed in
1934. The impact of this act is a “work in progress,” and we update how it is raising cable
television rates and lowering long-distance telephone rates. Incentive regulation—which is
designed to control prices while inducing regulated firms to be efficient—is increasingly used
in place of traditional regulatory practice. Our coverage has expanded to provide a more
detailed investigation of earning sharing, price caps, and yardstick regulation, with attention
to how they are applied to the electric distribution and local telephone markets. New case
studies of regulation are added, such as the 44 Liquormart decision, in which the U.S.
Supreme Court struck down a state prohibition against advertising of liquor prices. Finally,
appendices have been added that develop simple models to establish the rationale for
regulating a natural monopoly (chapter 10) and how interest group competition influences
regulation (chapter 16).

Part III, on social regulation, includes evidence on the cost per life saved for different gov-
ernment regulations through the first years of the George W. Bush administration. Chapter
20 now includes international evidence on the value of statistical life, as well as increased
discussion of risk-risk analysis. The most extensive changes took place with respect to the
environmental regulation discussion in chapter 21; extensive changes are appropriate because
new environmental regulation continues to be the most costly regulatory effort. That chapter now includes a discussion of the key generational issues that lie at the heart of current environmental debates. Should risks to the lives of the young be valued the same as risks to those with a very short life expectancy? How should the risks to future generations be valued as compared to our own well-being? More generally, how should we attach benefit values to environmental goods, which seldom are traded in markets and may be valued simply because of their existence, even if people will never use them? The environmental chapter also includes a new discussion of the siting of nuclear wastes and a more extensive treatment of conservative risk assessment practices. The most important new development with respect to the product safety issues treated in Chapter 22 has been the regulation through litigation movement. That chapter now discusses this phenomenon using the breast implant litigation as the principal case study. There is also a new presentation of the proper use of values of statistical life to determine the appropriate levels of safety in product liability cases. Chapter 23 now includes a discussion of the changes in the OSHA enforcement strategy enacted by the Clinton and George W. Bush administrations.

The minimum economics background needed for this book is usually an introductory price theory course. This background will enable students to grasp all of the empirical material and most of the theoretical developments. In some cases, the discussion advances to a level at which some background in intermediate microeconomic theory is desirable, but these more difficult sections can be omitted. A unique feature of this book is that it brings to bear on these issues new developments in industrial organization and game theory. The presentation of this more advanced material is self-contained, does not involve the use of calculus, and is incorporated in chapters in such a way that it can easily be omitted by an instructor with a different course emphasis.

We have used drafts of this book in our teaching at the undergraduate level, in business school curricula, and in teaching law students. Others have used this book in public policy schools. In no case did we use all of the book in any one course. Although the book’s coverage is nearly encyclopedic, it is still not all-inclusive. It is doubtful whether any single course can successfully cover all the material included in this book, except perhaps in an intensive two-semester sequence. Because instructors have a variety of different interests and instructional needs, we have structured the book in a manner that will facilitate its use in a variety of contexts.

Organization of the Book

*Economics of Regulation and Antitrust* consists of two introductory chapters, followed by three parts. The beginning of the book sets the stage and introduces some of the overriding issues, such as ascertaining what the objective is that government regulators maximize.
and considering the appropriate division of labor between the states and the federal government.

The following three parts of the book present the core of the analytical material. Part I focuses on antitrust policy, part II deals with economic regulation, and part III focuses on social regulation and patent policy. Each of these parts is structured in a similar manner. The first chapter of each part provides an overview of the key economic issues and the pertinent methodology that will be employed. We discuss the principal market failures in this context, and how economic analysis is used to address them. In every case, the first chapter of each part can be viewed as essential reading. The instructor can then select which of the subsequent case studies to use. Chapters that require the student to have read another chapter within that part, other than the introductory chapter, are noted in the following paragraphs. Otherwise, chapters within a part can be assigned in whatever order the instructor wishes. Any chapters that the instructor wishes to omit may be excluded.

Part I, which focuses on antitrust policy, includes a healthy dose of the analytical tools of modern industrial organization. Chapter 3 is an introductory overview of antitrust policy and of the other chapters in part I. Efficiency and technical progress are explained in chapter 4 as tools for evaluating policies. At least the first half of this chapter is probably necessary reading for understanding chapters 5–9.

Chapter 5, on oligopoly and collusive pricing, is novel in introducing oligopoly through a game-theoretic approach and then relating the theoretical models to antitrust cases. The discussion of market structure and entry deterrence (chapter 6) is mostly analytical; it can be skipped by instructors under time pressure in courses with a primary focus on antitrust cases. The remaining three chapters—horizontal and conglomerate mergers (chapter 7), vertical mergers and restrictions (chapter 8), and monopolization and price discrimination (chapter 9)—are stand-alone chapters that can be assigned or not, depending on the instructor’s preference.

Part II addresses the role of economic regulation. As evidenced by the dozen or so case studies in this part, economic regulation has been an integral part of the U.S. economy. Although there has been substantial deregulation of airlines, trucking, and long-distance telephone services, the debate over appropriate regulatory policies and reregulation is still very active.

An overview of economic regulation, including its historical development and a summary of regulatory practices, is provided in chapter 10. This chapter also provides the most in-depth textbook discussion of the efforts of social scientists to understand the extent of government regulation. The remainder of part II comprises two areas of interest. Chapters 11–15 cover the regulation of natural monopolies. The recent theory of natural monopoly is presented in chapter 11, while chapter 12 reviews actual regulatory practices with respect to electric utilities and local telephone companies. Although regulation is the standard U.S. government response to natural monopolies, alternatives are available; these are discussed in
chapters 13 and 14. Chapter 13 addresses a new and promising approach, franchise bidding, and provides a detailed case study of cable television. A more traditional alternative is that of government enterprise. It is reviewed in chapter 14, along with a comparative analysis of government ownership and regulation with respect to electric utilities. Then, in chapter 15, some dynamic issues related to monopoly regulation are explored in the context of the rapidly changing long-distance telecommunications market.

The regulation of markets that are potentially competitive receives in-depth treatment in the remaining three chapters of part II. A theoretical investigation of the effects of regulation is provided in chapter 16. These ideas are then applied to regulation in the transportation and energy industries. Chapter 17 closely examines airlines and surface freight transportation (in particular, trucking and railroads), while chapter 18 covers the crude oil and natural gas industries.

Part III focuses on the new forms of risk and environmental regulation that emerged primarily after the 1970s. Chapter 19 introduces the principal methodological issues, including market failures such as externalities and inadequate risk information, the primary economic test of benefit-cost analysis that applies in this area, and the rather daunting task that economists face in assigning dollar values to outcomes such as a five-degree temperature change caused by global warming.

The task of assigning market prices to outcomes that, by their very nature, are not traded in efficient markets is the focus of chapter 20. The primary case study concentrates on how economists attempt to assign a dollar value to risks to human life, which illustrates how economists have attempted to assess the pertinent trade-off rates that should be used in evaluating government policies. The next four chapters deal with various types of social regulation policies, including environmental protection regulation (chapter 21), product safety regulation (chapter 22), occupational safety regulation (chapter 23), and pharmaceutical regulation (chapter 24). Chapter 22 presents the greatest variety of social regulation issues that have been of long-term interest to researchers in industrial organization and in law and economics. A major strength of all these chapters is that they confront the current policy issues now under debate, including topics such as global warming, the role of product liability law, and the social consequences of smoking.

Chapter 24, on patents and pharmaceuticals, combines the theory of patents with a case study of their application to one of the most technologically progressive U.S. industries. It is a particularly timely chapter, given the current interest in health care reform and innovation.

Suggested Course Outlines

An intensive one-year course could cover this entire book. However, in most cases, instructors will be using the book in a context in which it is not feasible to cover all the material.
In this section we suggest course outlines that focus on entire chapters that are most appropriate for different course approaches. Most of the chapters include a series of sections that can be profitably assigned for student reading in separable units. This approach is especially useful for classes in which students may lack previous economic training. One of the authors, for example, has included extensive portions of part II of the book in table A as part of a more institutionally oriented course, where the focus is on the case studies and the most central economic principles.

In table A we have identified six different course approaches and the pertinent chapters that can be assigned for each one. The first type of course is the balanced one-quarter course. Such a course would include the introductory material in chapters 1 and 2 as general background; chapters 3–5, 7, and 9 from part I; chapters 10 and 12 from part II; and chapters 19–21 from part III.

The second course approach is a conventional antitrust course. It would place the greatest reliance on chapter 17 and part I of the book, which includes chapters 3–9. Instructors who wish to provide a broader perspective on some of the other topics in regulation might augment these chapters with the indicated chapters for the one-quarter course.

A course focusing on economic regulation would include primarily the introductory section and part II of the book, or chapters 1–2, 4, 10–18, 22, and 24. Similarly, a course focusing on social regulation would include the introductory section and part III of the book, or chapters 1–2, 4, and 19–24. In situations in which we have taught such narrowly defined courses, we have often found it useful to include the material from the balanced one-quarter course as well, to give the student a broader perspective on the most salient economic issues in other areas of government intervention.

Given the frontier treatment of industrial organization in part I, this book could also be used in a policy-oriented course on industrial organization. With chapters 3–6 providing the theoretical foundation in industrial organization, an instructor could select from the remaining chapters to cover a variety of policy issues. A suggestion is to use chapter 9 (its coverage of monopolization practices follows up on the theory of strategic entry deterrence in chapter 6), chapters 10, 13, and 16–18 (to examine how different types of economic regulatory structures can affect competition), and chapters 22 and 24 (to assess efforts such as product quality regulation).

The institutional course outline pertains to courses, particularly those in business schools, that wish to have a more institutional focus. For these courses, the objective is to focus on the empirical aspects of government regulation and antitrust policies, as well as the character of these policies. Moreover, these courses would require no advanced undergraduate economic methods.

The final course outline is a professional school survey, such as the one-semester course at Harvard Law School, where there is a mix of students’ economic backgrounds. Many chapters are included in their entirety: 1–4, 7, 10, 11, and 19–23. That course also includes all but
### Table A
Suggested Course Outlines

<table>
<thead>
<tr>
<th>Course Focus</th>
<th>Introduction</th>
<th>Part I</th>
<th>Part II</th>
<th>Part III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balanced one-quarter course</td>
<td>✓</td>
<td>✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
</tr>
<tr>
<td>Antitrust</td>
<td>✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
</tr>
<tr>
<td>Economic regulation</td>
<td>✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
</tr>
<tr>
<td>Social regulation</td>
<td>✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
</tr>
<tr>
<td>Industrial organization</td>
<td>✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
</tr>
<tr>
<td>Institutional</td>
<td>✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
</tr>
<tr>
<td>Professional</td>
<td>✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td>✓ ✓ ✓ ✓</td>
</tr>
</tbody>
</table>
some of the more technical material of chapters 5, 6, 8, 9, and 12. Much of the remaining chapters is also included in the course: one case study such as cable TV from chapter 13, the taxicab material from chapter 16, one example such as airlines from chapter 17, and the basics of price ceilings from chapter 18. Thus, many of the subsections of chapters are self-contained entities, so that instructors need not sacrifice substantive topics if the backgrounds or interests of students do not make it feasible to cover an entire chapter. The chapters in the book that meet these tests and can be readily grasped with an introductory economics background are also indicated in table A.

Acknowledgments

With this edition of *Economics of Regulation and Antitrust* the roster of authors has altered, as John Vernon has gone from second to third position. Joseph Harrington has taken over all the revisions of both the antitrust and economic regulation sections of the book. We remain indebted to John, who was instrumental in launching this venture over a decade ago and whose lucid expository style set a standard that we hope to maintain.

We would like to thank John Covell at MIT Press for his enthusiastic support, advice, and motivational comments.

Each of the authors has profited from word processing, editorial input, and research assistance at his institution. At Harvard Law School, we would like to thank Roxanne Halpine and C.W. Ross for research assistance, and the Olin Center for Law, Economics, and Business for financial support. At Johns Hopkins, we would like to thank Ron Taur for research assistance.

A number of individuals have provided comments on parts of the manuscript or advice along the way. These include the following:

Gilbert Becker, St. Anselm College
Klaus G. Becker, Texas Tech University
Myong-Hun Chang, Cleveland State University
R. Morris Coats, Nicholls State University
William Christie, Vanderbilt University
Christopher B. Colburn, Old Dominion University
Andrew R. Dick, University of California–Los Angeles
Susan Dudley, George Mason University
Henry Grabowski, Duke University
Preface to the Fourth Edition

Mark Kennet, Tulane University
Luther D. Lawson, University of North Carolina–Wilmington
Stanford L. Levin, Southern Illinois University at Edwardsville
Wesley A. Magat, Duke University
Matthew Marlin, Duquesne University
John Martin, Adelphi University
Sam Peltzman, University of Chicago
Robin A. Prager, Federal Reserve Board of Governors
Michael A. Salinger, Boston University
Daniel F. Spulber, Northwestern University
The government acts in many ways. The most familiar role of the government is the subject of public finance courses. The government raises money in taxes and then spends this money through various expenditure efforts. In addition, the government also regulates the behavior of firms and individuals. The legal system of the United States is perhaps the most comprehensive example of the mechanism by which this regulation takes place.

This book will be concerned with government regulation of the behavior of both firms and individuals within the context of issues classified as regulation and antitrust. Regulation of firms involves much more than attempting to deal with monopoly power in the traditional textbook sense. The setting of prices for public utilities, the control of pollution emitted in the firm’s production process, and the allocation of radio broadcast bands are all among the contexts in which government regulation plays a prominent role in influencing firm behavior.

The behavior of individuals has also come under increasing regulatory scrutiny. In some cases decisions are regulated directly, such as the requirement to wear seat belts. In addition, individuals are affected by regulations that influence either market prices or the mix of products that are available. Product safety standards, for example, serve to eliminate the high-risk end of the product-quality spectrum. The menu of products available to consumers and the jobs available to workers are subject to substantial regulatory influence.

To assess the pervasiveness of these efforts, consider a day in the life of the typical American worker. That worker awakes in the morning to the sound of his clock radio, where the stations he listens to and the wavelength they broadcast on are regulated by the Federal Communications Commission. Sitting down to breakfast, the worker is greeted by the label on the cereal box whose content is strictly regulated by the Federal Trade Commission and the Food and Drug Administration to avoid misleading consumers about the health benefits of breakfast cereals. The orange juice from concentrate can also no longer be labeled “fresh,” courtesy of a 1991 Federal Trade Commission action. The milk poured on the cereal is also regulated in a variety of ways, with perhaps the most important being through U.S. Department of Agriculture price supports (milk marketing orders). More recently, there has been substantial concern with the health risk characteristics of milk in terms of the presence of hormones (bovine somatotrophin), which has been the object of substantial regulatory debate. If one chooses to add fruit to the cereal, it is reassuring to know that the Environmental Protection Agency stringently regulates the pesticides that can be used on domestic produce. Unfortunately, imported produce that has been drenched in pesticides is not inspected with great frequency.

Before leaving for work, our typical American checks his e-mail messages and uses an Internet browser that has been the subject of the Microsoft antitrust litigation. While doing so, he may take prescription medicine manufactured by Glaxo Wellcome, which would have been manufactured by a larger company that also included SmithKline Beecham had not serious antitrust concerns been raised by their prospective merger.
Heading to work, our regulated individual climbs into a Japanese car that was successful in not violating any import quotas. The worker will be safer en route to work than in earlier years, thanks to extensive safety regulations by the National Highway Traffic Safety Administration. The fuel used by the car is also less environmentally damaging than would have been the case in the absence of U.S. Department of Transportation fuel economy standards and in the absence of EPA gasoline lead standards. The car will be more expensive as well due to these efforts.

Once on the job, the worker is protected against many of the hazards of work by occupational safety and health regulations. If injured, the worker will be insured through workers’ compensation benefits that the worker has in effect paid for through lower wages. A host of U.S. Department of Labor regulations, as well as Equal Employment Opportunity Commission stipulations, ensure that the worker will not be unduly discriminated against during the course of his employment.

Our worker’s phone calls are billed at telephone rates set by regulation, although increasingly these rates have been influenced by market forces. Visiting business associates travel on planes whose availability and fares have been greatly influenced by regulatory changes. The safe arrival of these associates is due in part to the continued vigilance of the Federal Aviation Administration and the safety incentives created by tort liability lawsuits following airplane crashes.

Even when our individual escapes from work for an evening of relaxation and recreation, government regulations remain present. If the worker eats dinner at a restaurant, there is a good chance that he or she will be forbidden to smoke cigarettes. The U.S. Consumer Product Safety Commission has regulatory responsibility for a wide range of sports equipment, ranging from all-terrain vehicles to baseball helmets.

While shopping over the weekend, the worker is asked by a political activist to sign a petition to force the local power company to reduce electricity rates. Lower electricity prices will surely save the worker money in the short run, but the worker wonders whether lower prices will deter this regulated monopoly from performing better in the future.

Although some deregulation has taken place in the past decade, the scope of government regulation remains quite broad. The role of regulation in American society remains pervasive. Various forms of government regulation touch almost every aspect of our activities and consumption patterns. The widespread impact of regulation is not unexpected, inasmuch as this represents a very potent mechanism by which the government can influence market outcomes.

The Rationale for Regulation and Antitrust Policies

If we existed in a world that functioned in accordance with the perfect competition paradigm, there would be little need for antitrust policies and other regulatory efforts. All markets would
consist of a large number of sellers of a product, and consumers would be fully informed of the product's implications. Moreover, there would be no externalities present in this idealized economy, as all effects would be internalized by the buyers and sellers of a particular product.

Unfortunately, economic reality seldom adheres very closely to the textbook model of perfect competition. Many industries are dominated by a small number of large firms. In some instances, principally the public utilities, there may even be a monopoly. Consumers who use hazardous products and workers who accept risky employment may not fully understand the consequences of their actions. There are also widespread externalities that affect the air we breathe, the water we drink, and the viability of the planet for future generations.

The government has two types of mechanisms at its disposal to address these departures from the perfectly competitive model. The first mechanism is price incentives. We can impose a tax on various kinds of activities in order to decrease their attractiveness. There is some attempt to have taxes that are product specific, as in the case of alcohol taxes and cigarette taxes, but there the notion has largely been that we should be taxing products perceived as luxuries. The tax on cars that fail to meet fuel economy standards, known as the gas-guzzler tax, perhaps best represents the notion of utilizing the price mechanism to influence economic behavior. Gasoline taxes, which remain below their optimal level, serve a similar function.

An alternative to taxes is to try to control behavior directly. We make this effort in the field of antitrust when the government takes explicit action to block mergers that might threaten the competitive character of a market. In the area of utility regulation, a complex web of regulations prevents public utilities from charging excessive rates for their electricity, which is a commodity for which the electric companies have a captive market. Much health, safety, and environmental regulation similarly specifies the technological requirements that must be met or the pollution standards that cannot be exceeded. This book will consequently be concerned primarily with various forms of government action that limit behavior related to the kinds of market failures discussed earlier.

Not all market failures stem from actions by firms. In some cases, individuals also may be contributing to the market failures. If we dispose of our hazardous waste products in a reckless manner, then there will be a need for government regulation to influence our activities. Although the preponderance of regulatory policies are directed at business, the scope of regulation is sufficiently comprehensive to include all economic actors.

---

**Antitrust Regulation**

The first of the three parts of the book deals with antitrust policy. Beginning with the post–Civil War era, there has been substantial concern with antitrust issues. This attention was stimulated by a belief that consumers were vulnerable to the market power of
monopolies. Because of the potential economic losses that result from monopolies, a number of states enacted antitrust laws at the end of the nineteenth century. The U.S. Congress also was particularly active in this area in the early part of the twentieth century, and many of the most important pieces of legislation governing the current antitrust policy date back to that time. The major federal statute continues to be the 1890 Sherman Act.

The Changing Character of Antitrust Issues

The scope of antitrust issues is quite broad. It encompasses the traditional concerns with a monopoly, but these issues are less prominent now than they once were. Several decades ago, major topics of debate concerned whether IBM, AT&T, General Motors, and other major firms had become too powerful and too dominant in their markets. Debates such as these would seem quaint today—perhaps useful as an exercise in an economic history course. Today these once-dominant companies are now humbled giants, weakened by the effects of foreign competition. In many respects we have a global market rather than a U.S. market for many products, so that some of the earlier concerns about monopolies have been muted.

Indeed, in the 1980s we even witnessed a merger that would have been totally unthinkable three decades earlier. The merger of General Electric with RCA created a powerful electronics corporation of unprecedented size. The rationale for the merger was that a large scale was necessary to support the innovation needed to meet the threat of foreign competition. The competitive threat was certainly real. Whereas several decades ago these companies produced the great majority of all electronics items used in the United States, by the 1990s it was difficult to find a TV or VCR not made in Japan.

In much the same vein, one wonders what the attitude toward the growing market power of Microsoft will be a quarter century from now. Will it continue to dominate the computer software market in much the same way that IBM did initially for mainframe computers, or will we observe the same kinds of inroads that were made in other highly concentrated markets? The presence of market power is not the only pertinent characteristic, as the source of this power and the potential for new entrants to be economically viable vary across different contexts. The network externalities that give rise to Microsoft’s influence are quite different from the nature of the market power of General Motors, which formerly made more reliable and more stylish automobiles.

The current structure of antitrust policies is diverse in character and impact. The overall intent of these policies has not changed markedly over the past century. Their intent is to limit the role of market power that might result from substantial concentration in a particular industry. What has changed is that the concerns have shifted from the rise of single monopolies to mergers, leveraged buyouts, and other financial transactions that combine and restructure corporations in a manner that might fundamentally influence market behavior.
Reasoning behind Antitrust Regulations

The major concern with monopoly and similar kinds of concentration is not that being big is necessarily undesirable. However, because of the control over the price exerted by a monopoly, there are economic efficiency losses to society. Product quality and diversity may also be affected. Society could potentially be better off if limitations were imposed on the operation of a monopoly or a similar kind of concentrated industry.

Recent research has greatly changed how we think about monopolies. For example, one major consideration is not simply how big a firm currently is and what its current market influence is, but rather the extent to which there is a possible entry from a competitor. If firms fear the prospect of such entry, which has been characterized through the theory of contestable markets, then the behavior of a monopolist will be influenced in a manner that will promote more responsible behavior.

One of the reasons concentrated industries emerge is that some firms may have exclusive rights to some invention or may have been responsible for a technological change that has transformed the industry. Coca-Cola and Pepsi Cola are much more successful soft drink products than their generic counterparts because of their perceived superior taste. If their formulas were public and could be generally replicated, then their market influence would wane considerably.

Once a firm has achieved a monopolistic position, perhaps in part due to past innovation, we want it to continue to be dynamic in terms of its innovative efforts. A substantial controversy has long been waged by economists as to whether monopoly promotes or deters innovation. Will a monopolist, in effect, rest on its laurels and not have any incentive to innovate because of the lack of market pressure, or will monopolists be spurred on by the prospect of capturing all of the gains from innovation that a monopoly can obtain, whereas a firm in a perfectly competitive market would lose some of the benefits of innovation as its innovation is copied by the competitors? We will explore the relative merits of these arguments and the dynamics of monopolies but will not draw any general conclusions indicating the desirability of monopolies. The relative merits of monopolistic power tend to vary across market contexts.

Economic Regulation

In many contexts where natural monopolies have emerged, for reasons of economic efficiency it is desirable to have a monopolistic market structure. Nevertheless, these economic giants must be tamed so that they will not charge excessive prices. We do not wish to incur all of the efficiency and equity problems that arise as a result of a monopoly. Prominent examples include public utilities. It does not make sense to have a large number of small firms providing households with electricity, providing public transportation systems, or laying
phone lines and cable TV lines. However, we also do not wish to give single firms free rein in these markets because the interests of a monopoly will not best advance the interests of society as a whole. What’s good for General Motors is not necessarily good for America.

Other kinds of regulation affect energy prices and minimum wage levels. In some instances the focus of economic regulation is to control product price. This may be indirectly through profit regulation by, for example, limiting public utilities to a particular rate of return. In other cases, there are complex rules governing prices, as in the case of U.S. energy regulations and long-distance telephone rate regulation.

Development of Economic Regulation

The genesis of these various kinds of economic regulation can be traced back to the late 1800s, as in the case of antitrust. Before the turn of the century, the U.S. Congress had created the Interstate Commerce Commission to regulate railroad rates, and the early twentieth century saw a surge in the number of regulatory agencies in the transportation, communication, and securities fields. It was during that period, for example, that the U.S. Congress established the Federal Communications Commission and the Securities and Exchange Commission. In the case of antitrust policy, the main thrust of these efforts has been to prevent the development of the kinds of market concentration that threaten the competitive functioning of markets. In contrast, economic regulation generally recognizes that market concentration not only is inevitable but in many cases is a superior structure for the particular market. The intent is then to place limits on the performance of the firms in this market so as to limit the losses that might be inflicted.

Factors in Setting Rate Regulations

Establishing a rate structure that will provide efficient incentives for all parties is not a trivial undertaking. Consider the case of an electric power company. The objective is not to minimize the rate to consumers, inasmuch as very low rates may affect the desirability of staying in business for the electric company. In addition, it may affect the quality of the product being provided in terms of whether power is provided at off-peak times or whether power outages are remedied quickly. A series of complex issues affects the role of the dynamics of the investment process in technological improvements. We want the electric power company to innovate so that it will be able to provide cheaper power in the future. However, if we capture all the gains from innovation and give them to the consumers through lower prices, then the firm has no incentive to undertake the innovation. We cannot rely on market competition to force them to take such action, for there is little competition within this market structure. Thus we must strike a delicate balance between providing sufficient incentives for firms to undertake cost-reducing actions while at the same time ensuring that the prices for consumers are not excessive.
Key concerns that have arisen with respect to economic regulation pertain to the differing role of marginal costs and fixed costs. When the electric company provides service to your house or apartment, there are specific identifiable costs that can be attributed to the product that is delivered to you—the marginal costs. However, the electric company also incurs substantial fixed costs in terms of its plant and equipment that also must be covered. How should the electric company allocate these fixed costs? Should it simply divide them equally among the total number of customers? Should it allocate the costs proportionally to the total bills that the customers have? Should it distinguish among different groups depending on how sensitive they are to price? If businesses are less price-sensitive than are consumers, should the major share of these costs be borne by firms or by individual consumers?

Over the past several decades, economists have developed a very sophisticated series of frameworks for addressing these issues. The overall object of these analyses is to determine how we can best structure the price and incentive schemes for these firms so that we protect the interests of electricity customers while at the same time providing incentives and a reasonable return to the firms involved.

In the case of both antitrust and economic regulation, it is seldom possible to replicate an efficient market perfectly. There is generally some departure from the perfect competition situation that cannot be glossed over or rectified, even through the most imaginative and complex pricing scheme. However, by applying economic tools to these issues, we can obtain a much more sensible market situation than would emerge if there were no regulation whatsoever.

It is also noteworthy that economic analysis often plays a critical role in such policy discussions. Economic analyses based on the models discussed in this book frequently provide the basis for ratemaking decisions for public utilities. A prominent regulatory economist, Alfred E. Kahn, was responsible for the deregulation of the airlines, in large part because of his belief that competition would benefit consumers and create a more viable market structure than the previous system, in which airline market entry was dictated by a government bureaucracy. In contrast, economic analysis often does not play such a central role in the operation of a perfectly competitive market. The paradigmatic firm in a competitive market is a small enterprise operating in a sea of other small enterprises. Firms in this market do not routinely draw demand curves, marginal revenue curves, and marginal cost curves. Yet few economists are disturbed by this failure to apply economic tools explicitly, as economists since the time of Milton Friedman have argued that they implicitly apply the laws of economics, much as the billiard player applies the laws of geometry even though he may not have had any formal training in the subject. In the case of economic regulation, the application of economic reasoning is quite explicit. Economists play a prominent role in these regulatory agencies. Much of the policy debate turns on economic analyses and consideration of the merits of the kinds of economic issues that we will address in the course of this book.
Health, Safety, and Environmental Regulation

The newest form of regulation is the focus of part III of the book. In the 1970s the U.S. Congress created a host of agencies concerned with regulating health, safety, and environmental quality. These new regulatory agencies included the U.S. Consumer Product Safety Commission, the Occupational Safety and Health Administration, the Environmental Protection Agency, the Nuclear Regulatory Commission, and the National Highway Traffic Safety Administration. Although these forms of regulation are often referred to as being social regulation policies, the exact dividing line between economic regulations and social regulations is unclear. As a result, we will use the more specific designation of health, safety, and environmental regulation to encompass these forms of (social) regulation.

The chief impetus for the health, safety, and environmental regulations is twofold. First, substantial externalities often result from economic behavior. The operation of businesses often generates air pollution, water pollution, and toxic waste. Individual consumption decisions are also the source of externalities, as the fuel we burn in our cars gives rise to air pollution. Informational issues also play a salient role. Because of the peculiar nature of information as an economic commodity, it is more efficient for the government to be the producer of much information and to disseminate the information broadly. Individual firms, for example, will not have the same kind of incentives to do scientific research unless they can reap the benefits of the information. As a result, it is largely through the efforts of government agencies that society has funded research into the implications of various kinds of hazards so that we can form an assessment of their consequences and determine the degree to which they should be regulated.

Many government policies in the safety and environmental area deal with aspects of market behavior that by their very nature do not involve voluntary bargains. We all suffer the effects of air pollution from local power plants, but we did not agree to consume this air pollution. No transaction ever took place, and we are not explicitly compensated for these losses. In the absence of such a market transaction, we do not have explicit estimates of the price. No specific price has been set for the loss in visibility, or for that matter the various kinds of health effects and materials damages that will result from air pollution. Thus the first task that must be undertaken is to assess the worth of these various kinds of policies, inasmuch as the benefit values do not necessarily emerge from market behavior. A case study that will be explored in part III is how we attach a value to risk of death, which is perhaps the most difficult and most sensitive of these fundamental trade-offs that we face.

The three dimensions of health, safety, and environmental regulation arise with respect to risks in our environment, risks in the workplace, and risks from the products we consume. Most of our regulatory influence over these risks is through direct government regulation. Several federal agencies promulgate detailed requirements on workplace technologies as well as overall performance requirements.
Role of the Courts

An increasingly prominent player in this regulatory area has been the courts. Whereas in the case of antitrust regulations the courts have been enforcing laws passed by Congress, in the case of these social regulations the obligations that courts have been assessing pertain to the common-law requirements that have developed through decades of judicial decisions and precedents regarding how various kinds of accidents and other externalities are handled.

The incentives generated by the courts in many instances dwarf those created by regulatory agencies. The court awards for asbestos-related claims have been so substantial that the asbestos industry in the United States has been all but eliminated by the financial burdens. Liability costs have led the pharmaceutical industry largely to abandon research on contraceptive devices, and many vaccines have also been withdrawn from the market because of high liability burdens. Visitors at motels will notice that diving boards have disappeared—a consequence of the added liability insurance costs associated with this form of recreational equipment. The 1998 settlement of the state attorneys’ general cigarette lawsuits for over $200 billion launched a new phenomenon of regulation through litigation. There has been a steadily increasing reliance on the courts to foster changes in products, including lead paint, guns, cigarettes, breast implants, and fast food. The lines between regulation and litigation have become blurred, making it increasingly important to understand the broader set of social institutions that create incentives that serve to regulation behavior. To understand the role of the government within the context of this type of regulation, one must assess not only how the regulatory agencies function but what doctrines govern the behavior of the courts. These matters will also be addressed in part III.

Criteria for Assessment

Ideally, the purpose of antitrust and regulation policies is to foster improvements judged in efficiency terms. We should move closer to the perfectly competitive ideal than we would have in the absence of this type of intervention. The object is to increase the efficiency with which the economy operates, recognizing that we may fall short of the goal of replicating a perfectly competitive market, but nevertheless we can achieve substantial improvements over what would prevail in the absence of such government intervention.

Put somewhat differently, our task is to maximize the net benefits of these regulations to society. Such a concern requires that we assess both the benefits and the costs of these regulatory policies and attempt to maximize their difference. If all groups in society are treated symmetrically, then this benefit-cost calculus represents a straightforward maximization of economic efficiency. Alternatively, we might choose to weight the benefits to the disadvantaged differently or make other kinds of distinctions, in which case we can incorporate a broader range of concerns than efficiency alone.
For those not persuaded of the primacy of efficiency-based policy objectives, economics can nevertheless play an important role. Understanding how regulations function in our market economy will help illuminate who wins and who loses from regulatory policies, and to what extent. Economic analyses of corporate mergers, for example, can trace through the effects on prices, corporate profits, and consumer welfare in a manner that will promote more informed regulatory policies irrespective of one’s policy viewpoint.

Although maximizing economic efficiency or some other laudable social objective may be touted by economists as our goal, in practice it is not what the regulators choose to maximize. Regulators respond to a variety of political constituencies. Indeed, in many instances the same kinds of market failures that led to the regulation also may influence the regulations that are undertaken. As a society, for example, we overreact to low-probability risks that have been called to our attention. We fear the latest highly publicized carcinogen, and we cancel our New York vacation plans after the 9/11 terrorist attack. These same kinds of reactions to risk also create pressures for regulatory agencies to take action against these hazards.

Moreover, even in instances in which government agencies do not suffer from irrationality or from irrational pressures, they will not necessarily maximize social welfare. The actions taken by government agencies will influence the fortunes of firms and particular groups in society in substantial ways. The granting of a cable TV franchise may make one a millionaire, and restrictions on foreign competition will greatly boost the fortune of firms in highly competitive international markets. There is a strong private interest in regulatory outcomes, and we will explore the economic foundations and mechanisms by which this private interest becomes manifest.

The net result of these private interests is that regulatory policies frequently do not perform in the manner that economists would intend in an ideal world. As Nobel laureate George Stigler demonstrated, economic regulation often advances private interests, such as increasing the profits of the industry being regulated. The apparent object is not always to maximize social welfare but rather to provide transfers among particular groups in society. Moreover, these transfers may be provided in an inefficient way, so that regulatory policies may fall far short of our ideal.

The successive disappointments with regulatory policy have given rise to the terminology “government failure” to represent the governmental counterpart of market failure. In much the same way as markets may fail because some of the idealized assumptions fail to hold, the government too may fail. Our task is not always to replace a situation of market failure with government action, for governmental intervention may not yield a superior outcome. We should always assess whether the particular kinds of intervention that have been chosen will actually enhance market performance and improve our welfare to as great an extent as possible. As we examine the various forms of regulation, we will consider the merits of the regulation as well as the test that we should use in assessing their adequacy.
Questions and Problems

1. Why should the government intervene in situations of market failure? Should the government intervene if a market is fully efficient in the sense of being perfectly competitive? What additional rationales are present if there is an inadequacy in the market?

2. Discuss some of the kinds of instances in which the government has an advantage in terms of informational capabilities as well as superior expertise to make decisions that consumers would not have.

3. Economists frequently use the yardstick of economic efficiency in judging the merits of alternative policies. What value judgments are implicit in the economic efficiency doctrine?

Recommended Reading


Useful links regarding regulatory activities and research include the Office of Management and Budget Office of Information and Regulatory Affairs website (http://www.whitehouse.gov/omb/inforeg/regpol.html), and the Harvard Program on Empirical Legal Studies (http://www.law.Harvard.edu/programs/pels/).

Appendix

Abbreviations for Key Regulatory Agencies

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Agency</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLS</td>
<td>Bureau of Labor Statistics</td>
</tr>
<tr>
<td>CAB</td>
<td>Civil Aeronautics Board</td>
</tr>
<tr>
<td>Acronym</td>
<td>Full Name</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>CEA</td>
<td>Council of Economic Advisors</td>
</tr>
<tr>
<td>CFTC</td>
<td>Commodity Futures Trading Commission</td>
</tr>
<tr>
<td>CPSC</td>
<td>Consumer Product Safety Commission</td>
</tr>
<tr>
<td>DOD</td>
<td>Department of Defense</td>
</tr>
<tr>
<td>DOJ</td>
<td>Department of Justice</td>
</tr>
<tr>
<td>DOT</td>
<td>Department of Transportation</td>
</tr>
<tr>
<td>EEOC</td>
<td>Equal Employment Opportunity Commission</td>
</tr>
<tr>
<td>EPA</td>
<td>Environmental Protection Agency</td>
</tr>
<tr>
<td>FAA</td>
<td>Federal Aviation Administration</td>
</tr>
<tr>
<td>FAO</td>
<td>Food and Agricultural Organization</td>
</tr>
<tr>
<td>FCC</td>
<td>Federal Communications Commission</td>
</tr>
<tr>
<td>FDA</td>
<td>Food and Drug Administration</td>
</tr>
<tr>
<td>FDIC</td>
<td>Federal Deposit Insurance Corporation</td>
</tr>
<tr>
<td>FEC</td>
<td>Federal Election Commission</td>
</tr>
<tr>
<td>FERC</td>
<td>Federal Energy Regulatory Commission</td>
</tr>
<tr>
<td>FHA</td>
<td>Federal Housing Administration</td>
</tr>
<tr>
<td>FMC</td>
<td>Federal Maritime Commission</td>
</tr>
<tr>
<td>FSLIC</td>
<td>Federal Savings and Loan Insurance Corporation</td>
</tr>
<tr>
<td>FTC</td>
<td>Federal Trade Commission</td>
</tr>
<tr>
<td>ICC</td>
<td>Interstate Commerce Commission</td>
</tr>
<tr>
<td>FTC</td>
<td>International Trade Commission</td>
</tr>
<tr>
<td>MSHA</td>
<td>Mine Safety and Health Administration</td>
</tr>
<tr>
<td>NHTSA</td>
<td>National Highway Traffic Safety Administration</td>
</tr>
<tr>
<td>NIH</td>
<td>National Institutes of Health</td>
</tr>
<tr>
<td>NIOSH</td>
<td>National Institute of Occupational Safety and Health</td>
</tr>
<tr>
<td>NLRB</td>
<td>National Labor Relations Board</td>
</tr>
<tr>
<td>NRC</td>
<td>Nuclear Regulatory Commission</td>
</tr>
<tr>
<td>OIRA</td>
<td>Office of Information and Regulatory Affairs</td>
</tr>
<tr>
<td>OMB</td>
<td>Office of Management and Budget</td>
</tr>
<tr>
<td>OSHA</td>
<td>Occupational Safety and Health Administration</td>
</tr>
<tr>
<td>SEC</td>
<td>Securities and Exchange Commission</td>
</tr>
<tr>
<td>USDA</td>
<td>United States Department of Agriculture</td>
</tr>
</tbody>
</table>
The Making of a Regulation

A stylized account of the evolution of regulation and antitrust policies is this: A single national regulatory agency establishes the government policy to maximize the national interest, where the legislative mandate of the agency defines its specific responsibilities in fostering these interests. The reality of regulatory policymaking differs quite starkly from this stylized view. The process is imperfect in that some observers claim that “government failure” may be of the same order of importance as market failure.¹

One important difference is that not all regulation is national in scope. Much regulation occurs at the state and local levels. Recent political concern with the importance of reflecting the preferences and economic conditions at the local level has spurred an increased interest in regulatory activity other than at the federal level. It is noteworthy that from a historical standpoint most regulation, such as the rate regulations for railroads, began at the state level. These regulations were subsequently extended to the national level.

Even in situations in which it is a national regulatory body that is acting, this group may not be fostering the national interest. Special interest groups and their diverse array of lobbyists also have an influence on regulatory policy. Moreover, the legislative mandates of the regulatory agencies are typically specified much more narrowly than simply urging the agency to promote the national interest.

Another difference from the stylized model is that typically the regulatory agency is not the only governmental player. Congress and the judiciary provide one check, and, more important, the regulatory oversight process within the White House has substantial input. Each of these groups has its own agenda. Few observers would claim that any one of these agendas coincides exactly with the national interest.

The final possible misconception is that it is a simple matter for the government to issue a regulatory policy or to make a decision regarding antitrust policy. There are explicit steps that government agencies must take before instituting regulations. At each of these stages, several governmental and private players have an input into the process and can influence the outcome. The nature of this process and the way it affects the regulatory outcomes is the subject of this chapter.

The underlying principles governing antitrust and regulation policies must be consistent with the legislative mandates written by Congress. Actions taken with these legislative stipulations in turn are subject to review by the courts. These two sets of influences are pertinent to all policy actions discussed in this book.

Other aspects of the character of these policies differ considerably. The U.S. Department of Justice’s vigilance in pursuing antitrust actions varies with political administrations, in part because of differences in interpretation of the law. Although the U.S. Department of Justice occasionally issues formal regulations to guide industry behavior, such as procedures for implementing civil penalties, for the most part the main policy mechanism of influence is

litigation against firms believed to be violating the antitrust statutes. This threat of litigation also produces many out-of-court settlements of antitrust cases.

Many of the economic regulation agencies are independent regulatory commissions, such as the Federal Trade Commission and the Federal Communications Commission. In addition to initiating legal action, these agencies place extensive reliance on issuance of regulations to guide business behavior. The steps that must be taken in issuing these regulations follow the procedures discussed later in this chapter, except that there is no review by executive authority over regulatory commissions.

The final group of agencies consists of regulatory agencies within the executive branch. These agencies rely primarily on issuing formal regulations pursuant to their legislative mandates. For example, the Environmental Protection Agency (EPA) has issued lead emission standards in implementing the Clean Air Act. This regulatory activity is subject to review by the Office of Management and Budget (OMB) and the full rulemaking process detailed later in this chapter.

Because the regulatory procedures for executive branch agencies are most complex, this chapter will focus on them as the most general case. The issues are of greatest pertinence to the policies that are considered in part III of the book. However, the economic lessons involved are quite general. Government policies should not be regarded as a fixed object to be treated reverentially within courses on business and government. Rather, they are generated by a complex set of political and economic forces, not all of which produce desirable outcomes. Part of the task of the subsequent chapters is to ascertain which policies are beneficial and which are not.

State versus Federal Regulation: The Federalism Debate

Although regulation is frequently viewed as being synonymous with federal regulation, not all regulation is at the federal level. Restrictions on cigarette smoking in restaurants are determined at the local level, as are drinking ages. State regulatory commissions set utility rates and often are involved in complex legal battles over appropriate jurisdiction. Almost all insurance regulation occurs at the state level as well. Some states regulate insurance rates quite stringently, whereas in other states these insurance rates have been deregulated. The terms under which there are payouts under insurance schemes also vary with locale, as some states have adopted no-fault rules in accident contexts. States also differ in terms of the factors that they will permit insurance companies to take into account when setting rates. In some instances, the states prohibit the insurance company from factoring in the driver’s age, sex, or race when setting automobile insurance rates. Finally, states differ in terms of whether they make automobile insurance mandatory and, if it is mandatory, the extent of the subsidy that is provided to high-risk drivers by the lower-risk drivers.
Advantages of Federalism

The existence of state regulations of various kinds is not simply the result of an oversight on the part of federal regulators. There are often sound economic reasons why we want regulation to take place at the state level. Indeed, particularly in the Reagan and Bush administrations there was an emphasis on transferring some of the control over the regulatory structure and regulatory enforcement to the states—an emphasis that comes under the general heading of “federalism.” The extent of the impact of federalism principles has, however, been less than advocates of this approach intended. In recognition of this emphasis, the OMB issued the following regulatory policy guideline:

Federal regulations should not preempt State laws or regulations, except to guarantee rights of national citizenship or to avoid significant burdens on interstate commerce.²

A number of sound economic rationales underlie this principle of federalism. First, local conditions may affect both the costs and the benefits associated with the regulation. Preferences vary locally, as do regional economic conditions. Areas where mass transit is well established can impose greater restrictions on automobiles than can states where there are not such transportation alternatives.

The second potential advantage to decentralized regulation is that citizens wishing a different mix of public goods can choose to relocate. Those who like to gamble can, for example, reside in states where gambling is permitted, such as Nevada or New Jersey. The entire theory of local public goods is built around similar notions whereby individuals relocate in an effort to establish the best match between the local public policies and their preferences. The diversity of options made possible through the use of state regulation permits such choices to be made, whereas if all regulatory policies and public decisions were nationally uniform, there would be no such discretion.

A third advantage of local regulation is that it can reflect the heterogeneity of costs and benefits in a particular locale. Ideally, we would like to set national standards that fully reflect benefit and cost differences across areas. We want to recognize, for example, the need to regulate pollution sources more stringently when there are large exposed populations at risk. Federal regulations seldom reflect this diversity. In contrast, state regulations are seldom structured in a way to meet the needs in other states rather than their own.

A related advantage stemming from the potential for heterogeneity with state regulation is also the potential for innovation. Many states have embarked on innovative regulatory policies. California has been a leader in this regard, as it has instituted labeling requirements for hazardous chemicals as well as efforts to drastically roll back automobile insurance rates.

Being innovative does not necessarily imply that these innovations are beneficial, but there is a benefit that other states derive from these experiments, since they can see which regulatory experiments work and which ones do not. Experimentation at the local level will generally be less costly than at the national level, should the regulatory experiments prove to be a mistake. Moreover, if the experiment proves to be successful, then other states can and typically will follow suit.

Advantages of National Regulations

Although the benefits of local regulation are considerable, one should also take into account the potential advantages of national regulatory approaches as well. First, the national regulatory agencies often have an informational advantage over the local agencies. The U.S. Food and Drug Administration (FDA), for example, administers a regulatory structure for pharmaceuticals that entails substantial product testing. Duplicating this effort at the local level would be extremely costly and inefficient. Moreover, most local regulatory agencies have not developed the same degree of expertise as is present at the national level in this or in many other scientific areas.

A second rationale for national regulations is that uniform national regulations are generally more efficient for nationally marketed consumer products. If firms had to comply with fifty different sets of safety and environmental pollution standards for automobiles, production costs would soar. Labeling efforts as well as other policies that affect products involved in interstate commerce likewise will impose less cost on firms if they are undertaken on a uniform national basis.

The efficiency rationale for federal regulation is often more general, as in the case of antitrust policies. If the product market is national in scope, then one would want to recognize impediments to competition in the market through federal antitrust policies rather than relying on each of the fifty states to pursue individual antitrust actions.

A third rationale for federal regulation is that many problems occur locally but have national ramifications. Air pollution from power plants in the Midwest is largely responsible for the problems with acid rain in the eastern United States and Canada. Indeed, many of the environmental problems we are now confronting are global in scope, particularly those associated with climate change. Policies to address global warming will affect all energy sources. There is a need not only for national regulation but also for recognition of the international dimensions of the regulatory policy problem.

A final rationale for national regulations is that we view certain policy outcomes as being sufficiently important that all citizens should be guaranteed them. A prominent example is civil rights regulations. We do not, for example, permit some states to discriminate based on race and sex even if they would want to if not constrained by federal affirmative action requirements.
Product Labeling Example

An interesting case study that illustrates the competing merits of national versus state regulation is the 1986 California initiative known as Proposition 65.³ That ballot measure required the labeling of all products that are carcinogenic or reproductive toxicants. In the case of carcinogens, the safe harbor warning was “WARNING: This product contains a chemical known to the state of California to cause cancer.” The cancer risk threshold for such a warning requirement was a lifetime cancer risk of 1/100,000. The regulation exempted naturally occurring carcinogens, and alcoholic beverages would be addressed by point-of-purchase warnings rather than on product labels. These more lenient provisions were in response to the pressures exerted by the California agriculture industry and wine industry rather than any underlying risk-based rationale for treating natural carcinogens and alcoholic beverages differently.

Producers and grocery manufacturers were initially fearful of the prospect of a myriad of state regulations. Products labeled as carcinogenic in California might end up in stores in Oklahoma, possibly causing consumer confusion and alarm. As other states also adopted warnings, the prospect of not matching up the product and its state-specific warning with the correct market seemed substantial. About 45 percent of national retail sales of food products are produced and distributed nationally or regionally, so that differences in state labeling requirements affect about half of all food products sold. To avoid products labeled in one state from being shipped elsewhere, there would be additional costs for transportation, plant warehousing, field warehousing, and inventory control that would total $0.05 for a product costing $0.50.

The prospect of these additional costs imposed by a myriad of state regulations led the food manufacturing and grocery industry to lobby the Reagan administration for a single national regulation. Companies that initially opposed the individual state regulations sought a national uniform standard to reduce their compliance costs. No national regulation was adopted, and the anticipated crisis for firms never materialized. Companies reformulated most of their products subject to the warnings so as to avoid the stigmatizing effect of the labels. Also, the feared proliferation of conflicting state warnings never occurred. Nevertheless, the product risk labeling experience illustrates how the compliance costs associated with a multiplicity of state regulations can lead firms to support a national variant of regulations that they oppose on an individual state basis. Similar concerns in 2004 regarding state regulations led the U.S. automobile companies to oppose state-specific fuel economy standards and to favor uniform federal standards.

The Overlap of State and Federal Regulations

Because national regulations tend to have a preemptive effect, even if there is no specific legal provision providing for preemption, the prevention of substantial encroachment on the legitimate role of the states requires some restraint on the part of federal regulators. In recent years there have been several attempts to recognize the legitimate state differences that may exist.

Many of the examples of policies providing for an increased role of the states pertain to the administration of federal regulation. Beginning in 1987, the Department of Health and Human Services gave the states more leeway in their purchases of computers and computer-related equipment for the Aid to Families with Dependent Children program. Previously, the states had to undertake substantial paperwork to get approval for their computer needs. Similarly, the Department of Transportation has eased the paperwork and reporting procedures associated with subcontract work undertaken by the states, as in their highway construction projects.

On a more substantive level, the U.S. EPA has delegated substantial authority to the states for the National Pollutant Discharge Elimination System. This program establishes the water pollution permits that will serve as the regulatory standard for a firm’s water pollution discharges. Many states have assumed authority for the enforcement of these environmental regulations, and the EPA has begun granting the states greater freedom in setting the permitted pollution amount for the firms. The Occupational Safety and Health Administration (OSHA) has undertaken similar efforts, and many states are responsible for the enforcement of job-safety regulations that are set at the national level but are monitored and enforced using personnel under a state enforcement program.

Although the states continue to play a subsidiary role in the development and administration of antitrust and regulatory policies, there has been increased recognition of the important role that the states have to play. This increased emphasis on the role of the states stems from several factors. Part of the enthusiasm for state regulation arises from the natural evolution of the development of federal regulation. If we assume that the federal government will first adopt the most promising regulatory alternatives and then will proceed to expand regulation by adopting the less beneficial alternatives, eventually we will reach a point where there will be some policies that will not be desirable nationally but will be beneficial in some local areas. The states will play some role in terms of filling in the gaps left by federal regulation.

Another force that has driven the expanding role of state regulation has been the recognition that there are legitimate differences among states. In many instances, the states have taken the initiative to recognize these differences by taking bold regulatory action, particularly with respect to insurance rate regulation.
Finally, much of the impetus for state regulation stems from a disappointment with the performance of federal regulation. Indeed, it is not entirely coincidental that the resurgence of interest in federalism principles occurred during the Reagan administration, which was committed to deregulation. There has consequently been an increased emphasis on the economic rationales for giving the states a larger role in the regulatory process and in ascertaining that federal intervention is truly needed. The main institutional player in promoting this recognition of federalism principles has been the U.S. OMB within the context of the regulatory oversight process, which we will consider in later sections.

The Character of the Rulemaking Process

Although federal regulatory agencies do have substantial discretion, they do not have complete leeway to set the regulations that they want to enforce. One constraint is provided by legislation. Regulations promulgated by these agencies must be consistent with their legislative mandate, or they run the risk of being overturned by the courts. In addition, regulatory agencies must go through a specified set of administrative procedures as part of issuing a regulation. These procedures do not provide for the same degree of accountability as occurs in situations where Congress votes on particular pieces of legislation. However, there are substantial checks in this process that have evolved substantially over time to provide increased control of the actions of regulatory agencies.

The Chronology of New Regulations

Figure 2.1 illustrates the current structure of the rulemaking process. The two major players in this process are the regulatory agency and the OMB. The first stage in the development of a regulation occurs when the agency decides to regulate a particular area of economic activity. Once a regulatory topic is on the agency’s regulatory agenda, it must be listed as part of its regulatory program if it is a significant regulatory action that is likely to have a substantial cost impact. The OMB has the authority to review this regulatory program, where the intent of this review is to identify potential overlaps among agencies, to become aware of particularly controversial regulatory policies that are being developed, and to screen out regulations that appear to be particularly undesirable. For the most part, these reviews have very little effect on the regulations that the agency pursues, but they do serve an informational role in terms of alerting the OMB to potential interagency conflicts.

The next stage in the development of a regulation is to prepare a Regulatory Impact Analysis (RIA). The requirements for such RIAs have become more detailed over time. At present they require the agency to calculate benefits and costs and to determine whether the benefits of the regulation are in excess of the costs. The agency is also required to consider potentially more desirable policy alternatives.
Figure 2.1
The Regulatory Management Process

Figure 2.1 (continued)
After completing the RIA, which is generally a very extensive study of the benefits and costs of regulatory policies, the agency must send the analysis to the OMB for its review, which must take place sixty days before the agency issues a Notice of Proposed Rulemaking (NPRM) in the Federal Register. During this period of up to sixty days, the OMB reviews the proposed regulation and the analysis supporting it. In the great majority of the cases, the OMB simply approves the regulation in its current form. In many instances, the OMB negotiates with the agency to obtain improvements in the regulation, and in a few rare instances the OMB rejects the regulation as being undesirable. At that point the agency has the choice either to revise the regulation or to withdraw it.

This OMB review is generally a secret process. Later in this chapter we will present overall statistics regarding the character of the regulatory decisions in terms of the numbers of regulations approved and disapproved. However, what is lacking is a detailed public description of the character of the debate between the OMB and the regulatory agency. The secretive nature of this process is intended to enable the regulatory agency to alter its position without having to admit publicly that it has made an error in terms of the regulation it has proposed. It can consequently back down in a face-saving manner. Keeping the debate out of the public forum prevents the parties from becoming locked into positions for the purpose of maintaining a public image. The disadvantage of the secrecy is that it has bred some suspicion and distrust of the objectives of the OMB’s oversight process, and it excludes Congress and the public from the regulatory policy debate. Moreover, because of this secrecy, some critics of the OMB may have overstated the actual impact the review process has had in altering or blocking proposed regulations. Under the Clinton administration, the OMB made a major effort to open up more aspects of this review to public scrutiny.

If the regulation is withdrawn, there is also one additional step that the agency can pursue. In particular, it can attempt to circumvent the OMB review by making an appeal to the president, or to the vice president if he has been delegated authority for this class of regulatory issues.

After receiving OMB approval, the agency can publish the NPRM in the Federal Register. This publication is the official outlet for providing the text of all proposed and actual regulatory policies, as well as other official government actions. As a consequence, it serves as a mechanism for disseminating to the public the nature of the regulatory proposal and the rationale for it. Included in the material presented in the Federal Register is typically a detailed justification for the regulation, which often includes an assessment of the benefits and costs of the regulatory policy.

Once the regulatory proposal has been published in the Federal Register, it is now open to public debate. There is then a thirty- to ninety-day period for public notice and comment. Although occasionally the agency receives comments from disinterested parties, for the most part these comments are provided by professional lobbying groups for business, consumer, environmental, and other affected interests.
After receiving and processing these public comments, the regulatory agency must then put the regulation into its final form. In doing so, it finalizes its regulatory impact analysis, and it submits both the regulation and the accompanying analysis to the OMB thirty days before publishing the final regulation in the *Federal Register*.

The OMB then has roughly one month to review the regulation and decide whether to approve it. In many cases, this process is constrained even further by judicial deadlines or by deadlines specified in legislation that require the agency to issue a regulation by a particular date. In recent years regulatory agencies have begun using these deadlines strategically, submitting the regulatory proposal and the accompanying analysis shortly before the deadline so that the OMB will have little time to review the regulation before some action must be taken. Rejected regulations are returned to the agency for revision, and some of the most unattractive regulations may be eliminated altogether.

The overwhelming majority of regulations are, however, approved and published as final rules in the *Federal Register*. Congressional review is a very infrequent process, and the typical regulation goes into effect after thirty days. The regulation is still, of course, subject to judicial review in subsequent years.

Despite the multiplicity of boxes and arrows in figure 2.1, there are very few binding external controls on the development of regulations. The OMB has an initial chance at examining whether the regulation should be on an agency’s regulatory agenda, but at that stage so little is known that this approval is almost always automatic. Moreover, the OMB review process became less stringent in the Clinton administration than in the Reagan and Bush administrations. The only two reviews of consequence are those of proposed rules and final rules. The OMB’s approval is required for these stages, but this approval process is primarily influential at the margin. OMB review activities alter regulations in minor ways, such as by introducing alternative methods of compliance that agencies might have that will be less costly but equally effective. Moreover, as we will see in chapter 20, the OMB is also successful in screening out some of the most inefficient regulations, such as those with costs per expected life saved well in excess of $100 million.

Although many of the other steps, particularly those involving public participation, are not binding in any way, the agency still must maintain its legitimacy. In the absence of public support, the agency runs the risk of losing its congressional funding and the support of the president, who appoints regulatory officials and, even in the case of commissioners to organizations such as the Interstate Commerce Commission, is responsible for periodic reappointments. Thus, the public comment process often has a substantive impact as well.
Nature of the Regulatory Oversight Process

The steps involved in issuing a regulation did not take the form outlined in figure 2.1 until the 1980s. In the early 1970s, for example, there was no executive branch oversight. After the emergence of the health, safety, and environmental regulatory agencies in the 1970s, it became apparent that some oversight mechanism was needed to ensure that these regulations were in society’s best interests. For the most part, these agencies have been on automatic pilot, constrained by little other than their legislative mandate and potential judicial review as to whether they were adhering to the mandate. Congress can, of course, intervene and pass legislation requiring that the agency take a particular kind of action, as it did with respect to the lawn mower standard for the Consumer Product Safety Commission. However, the routine regulatory actions seldom receive congressional scrutiny. Most important, there is no need for congressional approval for a regulatory agency to take action provided that it can survive judicial review. Proponents of the various types of “capture theories” of regulation would clearly see the need for such a balancing review. If a regulatory agency has, in effect, been captured by some special interest group, then it will serve the interests of that group as opposed to the national interest. There are those who have speculated, for example, that labor unions exert a pivotal influence on the operation of OSHA and that the transportation industry wields considerable influence over the U.S. Department of Transportation.

The Nixon and Ford Administrations

The first of the White House review efforts was an informal “quality of life” review process instituted by President Nixon. The focus of this effort was to obtain some sense of the costs and overall economic implications of major new regulations.

This review process was formalized under the Ford administration through Executive Order No. 11821. Under this order, regulatory agencies were required to prepare inflationary impact statements for all major rules. These statements required that agencies assess the cost and price effects that their new regulations would have. Moreover, President Ford established a new agency within the White House, the Council on Wage and Price Stability, to administer this effort.

Although no formal economic tests were imposed, the requirement that agencies calculate the overall costs of their new regulations was a first step toward requiring that they achieve some balancing in terms of the competing effects that their regulations had. Before the institution of this inflationary impact statement requirement, regulatory agencies routinely under-

took actions for which there was no quantitative assessment of the costs that would be imposed on society at large. Clearly, the costs imposed by regulation are a critical factor in determining its overall desirability. Knowledge of these cost effects ideally should promote sounder regulatory decisions.

The review process itself was not binding in any way. The Council on Wage and Price Stability examined the inflationary impact analyses prepared by the regulatory agencies to ensure that the requirements of the executive order had been met. However, even in the case of an ill-conceived regulation, no binding requirements could be imposed provided that the agency had fulfilled its obligations to assess the costs of the regulation, however large they may have been.

The mechanism for influence on the regulatory process was twofold. First, the Council on Wage and Price Stability filed its comments on the regulatory proposal in the public record as part of the rulemaking process. Second, these comments in turn provided the basis for lobbying with the regulatory agency by various members of the Executive Office of the President. Chief among these participants were members of the President’s Council of Economic Advisors and the president’s domestic policy staff.

The Carter Administration

Under President Carter, this process continued, with two major additions. First, President Carter issued his Executive Order No. 12044, which added a cost-effectiveness test to the inflationary impact requirement. The regulatory impact analyses that were prepared by regulatory agencies now had also to demonstrate that the “least burdensome of the acceptable alternatives have been chosen.” In practical terms, such a test rules out clearly dominated policy alternatives. If the government can achieve the same objective at less cost, it should do so. Reliance on this principle has often led economists, for example, to advocate performance-oriented alternatives to the kinds of command and control regulations that regulators have long favored.

In practice, however, the cost-effectiveness test affects only the most ill-conceived regulatory policies. For the most part, this test does not succeed in enabling one to rank policies in terms of their relative desirability. Suppose, for example, that we had one policy option that could save ten lives at a cost of $1 million per life, and we had a second policy option that could save twenty lives at a cost of $2 million per life. Also assume that these policy options are mutually exclusive: if we adopt one policy, we therefore cannot pursue the other. The first policy has a higher cost-effectiveness in that there is a lower cost per life saved. However, this policy may not necessarily be superior. It may well be in society’s best interest to save an additional ten lives even though the cost per life saved is higher, because overall the total net benefits to society of the latter option may be greater. Comparison of total benefits and costs of regulatory impacts was a common focus of Carter’s regulatory oversight program, but no formal requirements had to be met.
The other major change under President Carter was the establishment of the Regulatory Analysis Review Group. The primary staff support for this effort came from the Council on Wage and Price Stability and the President’s Council of Economic Advisors. However, the impact that reviews by this group had was enhanced by the fact that it also included representatives from the President’s Domestic Policy Staff, the OMB, and various cabinet agencies. The establishment of this group was a recognition that the executive oversight process had to be strengthened in some way, and the mechanism that was used for this strengthening was to bring to bear the political pressure of a consensus body on the particular regulatory agency. Moreover, the collegial nature of this group served an educational function as well in that there was a constant effort to educate regulatory officials regarding the proper economic approach to be taken within the context of regulatory analyses. For example, EPA officials present during a discussion of a proposed regulation by the National Highway Traffic Safety Administration could participate in a debate over the merits of the regulation and the appropriate means for assessing these merits, where the same kinds of generic issues were pertinent to their own agency as well. The reports by this group were not binding, but because they reflected the consensus view of the major branches of the Executive Office of the President as well as the affected regulatory agencies, they had an enhanced political import.

Even with these additional steps there was no binding test other than a cost-effectiveness requirement that had to be met. Moreover, the effectiveness of the informal political leverage in promoting sound regulatory policies was somewhat mixed. One famous case involved the OSHA cotton dust standard. OSHA proposed a standard for the regulation of cotton dust exposures for textile mill workers. The difficulty with this regulation, in the view of the regulatory oversight officials, was that the cost of the health benefits achieved would be inordinately high—on the order of several hundred thousand dollars per temporary disability prevented. The head of the Council of Economic Advisors, Charles Schultze, went to President Carter with an assessment of the undue burdens caused by the regulation. These concerns had been voiced by the textile industry as well. President Carter first sided with the Council of Economic Advisors in this debate. However, after an appeal by Secretary of Labor Raymond Donovan, which was augmented by an expression of the affected labor unions’ strong interests, Carter reversed his decision and issued the regulation. What this incident made clear is that even when the leading economic officials present a relatively cogent case concerning the lack of merit of a particular regulation, there are political factors and economic consequences other than simply calculations of benefits and costs that will drive a policy decision.

As a postscript, it is noteworthy that the Reagan administration undertook a review of this cotton dust standard shortly after taking office. Although Reagan administration economists were willing to pursue the possibility of overturning the regulation, at this juncture the same industry leaders who had originally opposed the regulation now embraced it, having already complied with the regulation, and they hoped to force the other, less technologically advanced
firms in the industry to incur these compliance costs as well. The shifting stance by the textile industry reflects the fact that the overall economic costs imposed by the regulation, not the net benefit to society, are often the driving force behind the lobbying efforts involved in the rulemaking process.

The Reagan Administration

Under the Reagan administration there were several pivotal changes in the regulatory oversight mechanism. First, President Reagan moved the oversight function from the Council on Wage and Price Stability to the OMB. Because the OMB is responsible for setting the budgets of all regulatory agencies and has substantial authority over them, this change increases the institutional clout of the oversight mechanism. The second major shift was to increase the stringency of the tests being imposed. Instead of simply imposing a cost-effectiveness requirement, Reagan moved to a full-blown benefit-cost test in his Executive Order No. 12291:

Sec. 2. General Requirements. In promulgating new regulations, reviewing existing regulations, and developing legislative proposals concerning regulation, all agencies, to the extent permitted by law, shall adhere to the following requirements:

a. Administrative decisions shall be based on adequate information concerning the need for and consequences of proposed government action;
b. Regulatory action shall not be undertaken unless the potential benefits to society for the regulation outweigh the potential costs to society;
c. Regulatory objectives shall be chosen to maximize the benefits to society;
d. Among alternative approaches to any given regulatory objective, the alternative involving the least net costs to society shall be chosen; and

e. Agencies shall set regulatory priorities with the aim of maximizing the aggregate net benefits to society, taking into account the condition of the particular industries affected by regulations, the condition of the national economy, and other regulatory actions contemplated for the future.

If, however, the benefit-cost test conflicts with the agency’s legislative mandate—as it does for all risk and environmental regulations—the test is not binding.

The third major change in the executive branch oversight process was the development of a formal regulatory planning process whereby the regulatory agencies would have to clear a regulatory agenda with the OMB. This procedure, which was accomplished through Executive Order No. 12498, was an extension of a concept begun under the Carter administration known as the Regulatory Calendar, which required the agency to list its forthcoming regulatory initiatives. This exercise has served to alert administration officials and the public at large as to the future of regulatory policy, but on a practical basis it has not had as much impact on policy outcomes as has the formal review process, coupled with a benefit-cost test.
The Bush Administration

Under President Bush, the regulatory oversight process remained virtually unchanged. The thrust of the effort was almost identical in character to the oversight procedures that were in place during the second term of the Reagan administration. For example, the same two key executive orders issued by Reagan remained in place under President Bush.

The Clinton Administration

President Clinton continued the regulatory oversight process in a manner that was not starkly changed from the two previous administrations. In his Executive Order No. 12866, President Clinton established principles for regulatory oversight similar to the emphasis on benefits, costs, and benefit-cost analysis of previous administrations. However, the tone of the Clinton executive order was quite different in that it was less adversarial with respect to the relationship with regulatory agencies. Moreover, this executive order correctly emphasized that many consequences of policies are difficult to quantify and that these qualitative concerns should be taken into account as well. The Clinton administration also raised the threshold for reviewing proposed regulations, restricting the focus to the truly major government regulations.

The George W. Bush Administration

The administration of George W. Bush kept Clinton’s Executive Order No. 12866 intact until 2002, when Executive Order No. 13258 introduced some minor structural changes pertaining to the role of the vice president. The two principal advances in the rulemaking process during the recent Bush administration were a fuller articulation of the economic principles to guide regulatory analyses and the introduction of “prompt letters” as a mechanism for urging agencies to initiate regulatory initiatives.

Although the main function of regulatory oversight will continue to be restraining excessive regulations, ideally the OMB will also be able to make assessments of how resources can be allocated more effectively and whether valuable regulatory opportunities are being missed. The OMB prompt letters, which are available to the public, have created pressures that led HHS and the FDA to introduce labeling for trans-fatty acids, strengthen corporate governance of Fannie Mae and Freddie Mac, and consider a proposed EPA rule to reduce pollution from non-road diesel engines.

Regulatory Reform Legislation

Notwithstanding the existence of executive branch oversight, Congress has also sought to bring the cost of regulation under control. There has been increasing recognition that a greater effort must be made to restrict regulatory initiatives to those that are truly worthwhile.
Coupled with this belief is an acknowledgment that executive branch oversight alone cannot ensure sound regulatory outcomes.

The source of the difficulty can be traced to the restrictive legislative mandates of regulatory agencies. In the case of health, safety, and environmental regulations, the legislation drafted by Congress did not require that agencies achieve any balance between benefits and costs. Indeed, in some cases the legislation even precluded that agencies undertake such balancing or consider cost considerations at all. Such an uncompromising approach can be traced in part to ignorance on the part of legislators, who did not understand the potential scope of these regulatory efforts or the fact that absolute safety is unattainable. Society could easily exhaust its entire resources with potential safety-enhancing efforts before achieving a zero risk level.

Typical of such uncompromising mandates is the requirement in the Occupational Safety and Health Act that the agency “assure so far as possible every man and woman in the nation safe and healthful working conditions.” In the 1980 U.S. Supreme Court decision with respect to the proposed OSHA cotton dust standard, the court interpreted this obligation narrowly.5

The court interpreted feasibility as “capable of being done” rather than in terms of benefit-cost balancing. Regulators have used this decision in conjunction with their own restrictive legislative mandates to claim that they are constrained by their legislation to ignore benefit-cost concerns. Agencies consequently seek to bolster their position by claiming that they are constrained by legislation, but these constraints are not necessarily always binding. In a subsequent U.S. Supreme Court decision, the Court ruled that agencies did have the flexibility to interpret their legislative mandate in a reasonable manner.6 In this particular case, the court gave the EPA the flexibility to adopt the “bubble” policy whereby it let firms select the most cost-effective means of reaching an air pollution target rather than requiring that firms meet a specific pollution objective for each emissions source.

To date, regulatory agencies have not attempted to avail themselves of this flexibility, and the OMB has been unsuccessful in urging them to do so. Since 1995 there has been a continuing effort to pass regulatory reform legislation that, in effect, would make the regulatory guidelines issued by the president override the influence of the legislative mandates. The closest such efforts have come to success was in 1995, when both the House and the Senate passed regulatory reform legislation. No consensus legislation emerged, and regulatory reform bills continue to be pending before Congress.

These efforts have failed thus far perhaps because the proposed bills have been overly ambitious. In addition to benefit-cost requirements, proposed legislation would have also

revamped the risk-analysis process by requiring that agencies use mean risk assessments rather than upper-bound values. Many proposed bills also included requirements that went beyond revamping the criteria for regulations, including peer review, judicial review of regulatory analyses, and retrospective assessments of regulatory performance.

The principal components of any such legislation are requirements that agencies assess the benefits and costs of their regulations and demonstrate that the benefits exceed the costs. Other less ambitious possibilities also could be effective, such as permitting agencies to balance benefits and costs but not requiring them to do so. Under this approach, it would be the responsibility of the OMB regulatory oversight group to exert the leverage without the presence of existing legislative constraints. These issues are likely to continue to be on the congressional legislative agenda until some kind of regulatory reform bill resolves the conflict between the national interest in balanced regulatory policies and the agencies’ adherence to restrictive legislative mandates.

**Benefit-Cost Analysis**

From an economic efficiency standpoint, the rationale for a benefit-cost approach seems quite compelling. At a very minimum, it seems reasonable that society should not pursue policies that do not advance our interests. If the benefits of a policy are not in excess of the costs, then clearly it should not be pursued, because such efforts do more harm than good. Ideally we want to maximize the net gain that policies produce. This net gain is the discrepancy between benefits and costs, so our objective should be to maximize the benefit-minus-cost difference.

The underlying economic impetus for the benefit-cost approach is the Hicksian potential compensation principle. The gainers from such policies can potentially compensate the losers, making all parties better off. However, unless potential compensation is actually paid, there is no assurance that everyone’s welfare will be improved. As a practical matter, it is generally impossible to make everyone better off from each individual regulatory policy, but making sound decisions across the entire spectrum of regulatory policies will make almost all of us better off.

The requirement that benefits exceed costs for sound regulatory policies has also given rise to a simple shorthand. The ratio of benefits to costs, or the benefit-cost ratio, must exceed 1.0 for a policy to be potentially attractive. This requirement serves as the minimal test for policy efficacy, as our overall objective should be to maximize the spread between benefits and costs.

To see how one would design a regulatory policy to reap the greatest net benefits, let us consider as a concrete example environment policy choice. The underlying principles are identical in other policy arenas as well. As is indicated in figure 2.2, the cost of providing environmental quality rises as the level of environmental quality improves. Moreover, the cost increases at an increasing rate because improvements in environmental quality become
increasingly costly to achieve. As the most promising policy alternatives are exploited, one must dip into less effective means of enhancing environmental quality, and resorting to these contributes to the rise in costs.

The other curve in the diagram is the total benefits arising from improved environmental quality. The initial gains are the greatest, as they may affect our life and well-being in a fundamental manner. The additional health and welfare effects of environmental quality improvements eventually diminish. Our task of finding the best level of environmental quality to promote through regulation reduces to achieving the largest spread between the total benefit and total cost curves. This maximum is achieved at the environmental quality level $q^*$. At that point, the gap between the cost and benefit curves is the greatest, with the gap giving
the maximum value of the net benefits less costs that are achievable through environmental regulation.

The slope of the total cost and total benefit curves is equal at environmental quality \( q^* \). The slope of the total cost curve is known as the marginal cost, as it represents the incremental increase in cost that arises from a unit increase in environmental quality. Similarly, the slope of the total benefit curve is known as the marginal benefit curve, as it represents the increment in benefits that would be produced by a one-unit increase in environmental quality. An alternative way to assess the optimal policy is to examine the marginal cost and marginal benefit curves, which are illustrated in figure 2.3. Marginal costs are rising because of the decreasing productivity of additional environment-enhancing efforts as we pursue additional improvements in environmental quality. Similarly, the marginal benefits shown in this curve are declining because they experience the greatest incremental benefits from such improvements when the environmental quality is very bad. The optimal policy level is at environmental quality level \( q^* \), at which we equate marginal benefits and marginal costs. Thus the requirement for optimal quality choice can be characterized by the following familiar equation:

\[
\text{Marginal benefits} = \text{Marginal costs.} \tag{2.1}
\]
This result that maximizing net benefits is achieved by equating marginal benefits and marginal costs will be a recurring theme throughout the book. Subsequent chapters will examine decisions by firms that can be recast in this framework. Consider equation 2.1 within the context of a firm choosing how much output to produce. A profit maximizing firm will produce up to the point where the marginal costs of production are equal to the marginal benefit, which equals the additional revenue produced by selling one more unit. In the case of a competitive firm, which is small relative to the entire market, the marginal benefit of selling an extra unit is the product price, so a competitive firm setting marginal benefits equal to marginal costs will produce at the point where price equals the marginal cost of production. A monopolistic firm will be somewhat different in that this firm is so large relative to the market that more sales by the monopoly will affect the market price. The monopolist will set the marginal cost equal to the additional revenue brought in by selling one more unit, which will differ from the price of the last unit sold, since more sales affect the price paid for all units of the good.

Discounting Deferred Effects

If all the effects of regulatory policies were immediate, one could simply sum up these influences, treating effects today the same as one would treat an impact many years from now. Even if one ignores the role of inflation, it is important to take the temporal distribution of benefits and costs into account. If one could earn a riskless real rate of interest \( r \) on one’s own money, then the value of a dollar today is \((1 + r)^{10}\) ten years from now. Thus, resources have an opportunity cost, and one must take this opportunity cost into account when assessing the value of benefit and cost streams over time. This issue is not unique to the social regulation area, but it plays a particularly important role with respect to these regulations because of the long time lags that tend to be involved, particularly when evaluating regulations focusing on cancer and the future of the planet.

Although a substantial literature exists on how one should approach the discount rate issue and estimate the appropriate rate of discount, these approaches can be simplified into two schools of thought.\(^7\) One approach relies on the opportunity cost of capital. In this instance, market-based measures provide the guide as to the appropriate discount rate. A simple but not too unreasonable approximation to this measure is simply the real rate of return on federal bonds. The alternative is the social rate of time preference approach, under which society’s preference for allocating social resources across time may be quite different from the time rate expressed in private markets. How the social rate differs from the private rate and the extent of the difference from private rates of return have remained subjects of considerable debate.

From a practical standpoint, such controversies are not of major consequence in actual regulatory decisions. The U.S. OMB (under OMB circular A-94) now requires that all policy benefits and costs be assessed using a rate of interest of 7 percent and at the agency’s preferred discount rate. Before 1993, the OMB had required a 10 percent rate, which is an extremely high real (that is, inflation-adjusted) rate of return.

**Present Value**

The procedure by which one converts a stream of benefits and costs into a present value is simply to divide any deferred impacts in year $i$ by $(1 + r)^i$. Viewed somewhat differently, if one could earn a rate of interest $r$ on $1$ invested today, the value of this dollar $i$ years from now would be $(1 + r)^i$. Thus the present value calculation simply puts the future payoff into terms that are comparable to payoffs today. More specifically, if one has project benefits $B_i$ and $C_i$ in year $i$, then the formula is given by

$$\text{Present value} = \sum_{i=0}^{n} \frac{B_i - C_i}{(1 + r)^i}. \quad (2.2)$$

To see the implications of the present value calculation, consider a simplified discounting example in table 2.1. Three different sets of results are provided. First, the benefits and costs in which there is no discounting comprise the first part of the table. As can be seen, the benefits exceed the costs by 0.15, and the policy is worth pursuing. If one adopts a discount rate of 5 percent, then the deferred benefits one year from now have a lower present value. Nevertheless, the policy still remains justified on benefit-cost grounds, although the strength of the justification has been weakened. The final example shows the discount rate raised to 10 percent. This higher rate lowers the value of next year’s benefits even further. In this

<table>
<thead>
<tr>
<th>Table 2.1 Discounting Example</th>
<th>Year 0</th>
<th>Year 1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No Discounting</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benefits</td>
<td>1.00</td>
<td>2.15</td>
<td>3.15</td>
</tr>
<tr>
<td>Costs</td>
<td>−3.00</td>
<td>−0.00</td>
<td>−3.00</td>
</tr>
<tr>
<td>Benefits – Costs</td>
<td>−2.00</td>
<td>2.15</td>
<td>+0.15</td>
</tr>
<tr>
<td><strong>Discounting at 5%</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benefits</td>
<td>1.00</td>
<td>2.05</td>
<td>3.05</td>
</tr>
<tr>
<td>Costs</td>
<td>−3.00</td>
<td>−0.00</td>
<td>−3.00</td>
</tr>
<tr>
<td>Benefits – Costs</td>
<td>−2.00</td>
<td>+2.05</td>
<td>0.05</td>
</tr>
<tr>
<td><strong>Discounting at 10%</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benefits</td>
<td>1.00</td>
<td>1.95</td>
<td>2.95</td>
</tr>
<tr>
<td>Costs</td>
<td>3.00</td>
<td>0.00</td>
<td>3.00</td>
</tr>
<tr>
<td>Benefits – Costs</td>
<td>−2.00</td>
<td>1.95</td>
<td>−0.05</td>
</tr>
</tbody>
</table>
instance costs exceed benefits, and the policy is no longer justified. As a rough rule of thumb, since costs are generally imposed early in the life of a regulation and benefits often accrue later, raising the discount rate tends to reduce the overall attractiveness of policies. The exact relationship hinges on the number of sign reversals in the net-benefit-less-cost stream over time. For one sign reversal—net costs in the early periods followed by net benefits—raising the discount rates reduces the attractiveness of a policy. The role of discounting is particularly instrumental in affecting the attractiveness of policies with long-term impacts, such as environmental regulations that address long-run ecological consequences or cancer regulations for which the benefits will not be yielded for two or three decades. Not surprisingly, a major battleground over discounting was asbestos regulation, inasmuch as the deferred nature of the risk made discounting a major policy issue in a debate involving the EPA, OMB, and members of Congress. The EPA advocated a discount rate of zero so that the benefits of the regulation would appear to be large.

Although the practice of reducing the value of deferred benefits may seem to be unduly harsh, it will be muted at least to some extent by increases in the unit benefit value over time. As society continues to become richer, the value we place upon environmental quality and risk reduction will also rise. As a result, there will be some increase in the value benefits over time because of society’s increased affluence, which generally raises the value that people attach to their health or environmental quality.

In general, one will still discount in a manner that reduces the present value of future impacts. If one were in a situation in which one did not discount at all, which is a position that has been frequently advocated by the U.S. EPA and by some congressmen, then any action with permanent adverse effects could never be undertaken. A $1 annual loss that was permanent would swamp in value any finite benefit amount that was for one time only. No policies that would affect a unique natural resource or that would lead to the extinction of a species could ever be pursued. The cost of such efforts would be infinite. Trivial losses that extended forever could never be imposed, irrespective of how great the current benefits are. When confronted with the full implications of not discounting at all, it is likely that there would be few advocates of this practice. We certainly do not follow this practice in our daily lives. Otherwise, we would save all of our resources, earn interest, and spend the money in our last years of life.

In many instances it is necessary to calculate the present value of an infinite stream of payoffs. What, for example, is the value of a taxicab license that generates $V every year? Suppose that the payment is received at the end of each period. It is straightforward to show that the present value of this infinite stream is given by $V/r. For example, with an interest rate of 10 percent, the present value of $5,000 per year would be $5,000/(0.10) = $50,000.

8. Letting $S$ be the present value of this infinite stream,

\[ S = \frac{V}{1+r} + \frac{V}{(1+r)^2} + \frac{V}{(1+r)^3} + \ldots \]
The Criteria Applied in the Oversight Process

Certainly the most dominant criteria that have been used in the oversight process over the last decade have been those pertaining to ensuring the cost-effectiveness of the regulation and, more specifically, ascertaining that the benefits of the regulation exceed the costs. Although the OMB has frequently been unable to enforce the benefit-cost requirements because of conflicts with the agency’s legislative mandate, there have been several notable success stories that illustrate how effective regulation can be if approached in a sound economic manner.

Regulatory Success Stories

One of these success stories is visible every time we ride in an automobile. A prominent regulatory innovation has been the requirement that all cars have center-high mounted stop lamps. When the driver puts on the brakes, the brake lights go on as always, but so does a red light in the bottom center of the rear window. This 1983 regulation was the subject of an extensive analysis whereby the Department of Transportation demonstrated that the benefits of the regulation exceeded the costs. Equally important is that the Department of Transportation also conducted a series of tests with various fleets of automobiles to determine which of several stop lamp designs would be the most effective in reducing rear-end collisions. Thus there was an explicit attempt to evaluate regulatory policy alternatives and to select the most attractive from among these alternatives.

Perhaps the greatest regulatory success story of the 1980s involving the OMB is the phase-down of lead in gasoline. (Telephone deregulation did not involve the OMB but was probably of greater consequence.) Through a series of regulations, EPA requirements have all but eliminated the use of lead in gasoline. This regulation was accompanied by a comprehensive regulatory analysis that clearly established that the benefits of the regulation exceeded the

Multiply $S$ by $[1/(1 + r)]$:

$$
\frac{S}{(1+r)} = \frac{V}{(1+r)^2} + \frac{V}{(1+r)^3} + \frac{V}{(1+r)^4} + \ldots
$$

Subtracting the right-hand side expression and the left-hand side expression in the second equation from the right-hand side expression and the left-hand side expression of the first equation, one gets

$$
\frac{rS}{(1+r)} = \frac{V}{(1+r)}.
$$

Solving this equation for $S$, one finds that

$$
S = \frac{V}{r}.
$$
costs. It is noteworthy that this regulation, one of the few where the EPA clearly established the economic attractiveness of the policy in terms of benefit-cost ratio, is also one that had the greatest demonstrable impact of any pollution regulation instituted in the 1980s. Lead emissions declined dramatically in the 1980s, and the reduction in lead pollution represents the greatest environmental success story of that decade.

**Promotion of Cost-Effective Regulation**

One general way in which the government promotes the most cost-effective regulation is through the encouragement of performance-oriented regulation. Our objective is to promote outcomes that are in the interests of the individuals affected by regulations rather than simply to mandate technological improvements irrespective of their impact. This concern with ends rather than means leads to the promotion of the use of performance-oriented regulations whenever possible.

Rather than mandate nationally uniform standards, it is frequently desirable to give firms some discretion in terms of their means of compliance. The FDA’s tamper-resistant packaging requirements impose effectiveness requirements on the packaging but do not dictate particular types of packaging that must be used. Similarly, the child-resistant cap requirements of the Consumer Product Safety Commission specify safety thresholds that the caps must meet in terms of preventing children from opening the bottles, but they do not prevent firms from adopting particular cap designs that they might believe are most appropriate for the product.

The adoption of performance-oriented alternatives has generally lagged behind economists’ enthusiasm for these policies. Two principal reasons account for this discrepancy. First, the enforcement of some performance-oriented alternatives can be more expensive. If firms were simply given general guidelines to make their workplace safer but were not given any explicit instructions for doing so, then government inspectors would have a more difficult task in determining whether the firm had met the minimal safety requirements.10

Another major barrier to performance-oriented regulation has been political. In the case of air pollution requirements, congressmen from soft-coal-producing states lobbied for legislation that required firms to develop technological solutions to air pollution (that is, use of scrubbers) as opposed to changing the type of fuel they used to a less polluting form of coal. This emphasis was dictated by regional economic self-interests, not by national efficiency concerns.


10. The government could utilize an outcomes-based performance measure, such as total worker deaths and injuries. However, such a measure would be more effective for large firms than for smaller firms, which have a sufficiently small sample of workers that precise inferences cannot be drawn regarding the firms’ safety performance.
Distortion of Benefit and Cost Estimates

Another principle that has been promoted through the oversight process is the utilization of unbiased estimates of the benefits and costs. The need for lack of bias may appear to be both obvious and uncontroversial, but in fact it represents an ongoing problem with respect to risk regulations.

The scientific analyses underlying risk regulations typically include a variety of assumptions for the purpose of “conservatism,” but which in effect distort the assessment of the merits of the regulation. For example, projections of the cancer-causing implications of some chemical may be made by relying on the most sensitive animal species, as opposed to the animal species most relevant to extrapolation to humans. In addition, scientific analysts frequently focus on the upper end of the 95 percent confidence interval, thus placing great emphasis on how high the risk potentially could be as opposed to their best estimate of how high the risk actually is.

Focusing on the upper limit of the potential risk distorts the policy mix in a number of ways. Most important is that it shifts our attention to those hazards about which the least is known, as opposed to those hazards that pose the greatest threat and will endanger the greatest number of lives. Because we often know the least about the very-low-probability events because we have little experience to guide us, the effect has often been to tilt policies in the direction of the inconsequential low-probability events that we dimly understand, whereas the major sources of accidents and illness that are precisely understood receive less attention.

In some cases, there are additional conservatism factors incorporated arbitrarily within the risk analysis process. For example, risk analysts assessing the reproductive toxicity of different chemicals may simply multiply these risk levels by a factor of 1,000 for the purposes of “conservatism,” but there is no justification for multiplying by any factor.

The problem that these conservatism adjustments pose from the standpoint of government policy is that when we address different regulations and are comparing their efficacy, we do not know the extent to which the benefits have been distorted. Various conservatism factors are used by different agencies in different contexts. These adjustments are seldom detailed in the regulatory analysis and are often compounded in the successive stages of analysis. Conservatism multipliers are often added in each round of the calculations. Such distortions prevent the regulatory policymakers from having the accurate information they need to choose among policies. The overall judgment as to how conservative society wishes to be in bearing risk or in incurring other outcomes is a social policy decision that should be made at the policymaking level of the regulatory agencies and the executive branch. Arbitrary conservatism factors incorporated in the risk analysis in effect involve little more than stealth policymaking that is masquerading as a scientific exercise.
The Regulatory Role of Price and Quality

A general principle that has guided the development of regulation and in particular the deregulation effort is that “regulation of prices and production in competitive markets should be avoided.” The price system has a legitimate role to play, as is evidenced in the discussion of markets in all elementary economics textbooks. Recognition of the role of the price mechanism has provided the impetus for the deregulation of the rate entry regulations that were formerly present in industries like airlines, trucking, and communications. Some regulations, such as minimum wage requirements, explicitly interfere with these prices. The purported benefits of these regulations is that they will raise workers’ income level to a fairer wage amount needed for subsistence, although most labor economists believe that the long-run effect of minimum wage regulations is to displace workers from jobs. It appears in this regard that teenagers, particularly minority teenagers, have been most hard-hit by the adverse employment effects of higher minimum wage levels.

Just as we do not want to standardize product prices, we also do not wish to standardize quality except when there are legitimate reasons for doing so, as in the case of provision of minimal safety levels for cars. Antilock brakes and passenger side curtain airbags are beneficial safety features, but they are also quite expensive. We would like to give consumers the option to purchase such equipment; the more expensive cars typically offer these features. However, we do not require that all cars have them, for those features would comprise a substantial part of the product price for the low end of the market. Instead of mandating all available safety devices for all cars, we have required that certain minimal safety features be universal, and we permit other safety features to be optional. Consumers who place substantial value on safety can purchase the cars offering these additional features, and we can continually revise the nationally mandated safety standards to reflect the safety floor that is most sensible from the standpoint of being imposed on a universal basis.

The Impact of the Oversight Process

The objective of regulatory oversight is to foster better regulations, not necessarily less regulation. However, one consequence of improving regulation is that we will eliminate those regulations that are unattractive from the standpoint of advancing the national interest. Moreover, much of the impetus for regulatory oversight has been a concern with the excessive costs imposed by unattractive regulations, so that there has been considerable attention devoted to these costs.

The Cost of Regulation

The stakes involved are enormous. In 1990 President Bush noted the staggering levels of costs involved:

Federal regulations impose estimated direct costs on the economy as high as $175 billion—more than $1,700 for every taxpayer in the United States. These costs are in effect indirect “taxes” on the American public—taxes that should only be levied when the benefits clearly exceed the costs.12

Roughly half of these costs are attributable to EPA regulations, as earlier estimates of the costs imposed by EPA policies indicated that these regulatory costs alone were in the range of $70–$80 billion per year.13

In the absence of regulatory reform efforts, these costs would be substantially higher. The Council of Economic Advisors estimates that airline deregulation led to $15 billion worth of gains to airline travelers and airline companies.14 Similarly, estimates suggest that savings resulting from trucking deregulation have been in excess of $30 billion annually.15 The annual benefits from railroad deregulation have also been substantial—on the order of $15 billion annually.16 The total savings from these deregulation efforts in the transportation field are on the order of $60 billion per year, a substantial payoff indeed for a return to greater reliance on market forces.

Other Measures of the Size of Regulation

The most pertinent estimate of regulatory activity is the level of the costs that are generated by the regulation. Professor Thomas Hopkins, once a prominent regulatory oversight official, has compiled a comprehensive assessment of the costs of different federal regulatory programs. This tally appears in table 2.2, where the primary inputs to these calculations are the regulatory analyses prepared by government regulations on a prospective basis for new regulations.17 Actual costs of regulations may of course differ from those that are estimated at the time of the regulation’s promulgation. However, these cost measures are likely to be much more indicative of the scale of regulatory activity than are Federal Register counts.

---


As the information in table 2.2 indicates, the cost of these regulations is substantial. The total cost level in 1995 was $668 billion, which includes regulations that were simply transfers, such as the minimum wage. Transfers accounted for $147 billion of the costs. The minimum wage leads to higher wage payments for low-income workers. From an economic standpoint this is not an efficiency loss, but simply an effort that passes money around in society. The gains to workers offset the losses to firms. However, from the standpoint of the potential costs to the rest of society, the appropriate amount to be recognized is the total regulatory cost, since it is this regulatory cost amount that firms (or consumers and workers) must pay. In practice, however, the shifting of this and other costs among consumers, shareholders, workers, and other parties is a very complex matter.

In 1995 the total gross domestic product was $7.3 trillion, so the regulatory cost share of the gross domestic product was 9.2 percent. Another useful measure of regulatory costs is the regulatory cost per household. In 1995 these costs were estimated to be $6,809 per household. Regulatory costs consequently are not a trivial component of the gross domestic product, but it should also be taken into account that benefits are derived from these efforts as well. It is quite striking that for the 1995 federal regulatory costs, the largest component was for process regulation, or $218 billion in annual expenditures related to government paperwork requirements. Environmental regulation, such as that administered by the U.S. EPA, was next greatest in importance at $168 billion, followed by economic regulation at $80 billion. The role of deregulation in the economic regulation context is apparent, as economic regulations decreased substantially in cost from 1977 to 1995. Moreover, there has been a remarkable change in the mix of regulations, as environmental regulation has assumed increasing importance during the same period in which economic regulation has diminished in terms of the efficiency costs. Estimates for the year 2000 indicated additional regulatory cost growth due largely to environmental regulation and process regulations.

One of the most striking aspects of the regulatory cost mix is the substantial process regulation component of $218 billion in 1995 federal paperwork costs. A concern with

<table>
<thead>
<tr>
<th>Table 2.2</th>
<th>Annual Costs of Federal Regulation (Billions of 1995 Dollars)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Environmental and risk regulation</td>
<td>99</td>
</tr>
<tr>
<td>Price and entry controls</td>
<td>364</td>
</tr>
<tr>
<td>Paperwork</td>
<td>143</td>
</tr>
<tr>
<td>Total regulatory costs</td>
<td>606</td>
</tr>
</tbody>
</table>


18. Ibid.
paperwork required by federal activities has long been widespread. Moreover, unlike the regulatory efforts themselves, paperwork often lacks the clear-cut link to perceived societal benefits, such as improved environmental quality. Although politicians frequently voice commitments to reduce paperwork, this burden continues to grow. One difficulty is that gathering information generally appears to be attractive, inasmuch as more knowledge is better than less, but the benefits derived from the information are not always valued to determine whether the associated paperwork burden is justified. One frequently proposed policy that might address this issue is to establish a federal paperwork budget to limit the annual dollar value of paperwork costs.

Whereas the previous estimates referred to the total costs of all existing regulations, the annual change in these costs due to new regulations is instructive as well. The costs of new regulations are indicative of the annual pace of regulation, which varies substantially over time, as the annual totals from 1987–2003 reported in table 2.3 indicate. This table focuses only on the costs of new major rules reviewed by the OMB.

The differences across presidential administrations are often quite stark. The first three years of the George W. Bush administration accounted for a total of only $4.4 billion. This amount is dwarfed by the pace in the Clinton administration, which issued $13.1 billion in new regulations in year 2000 alone. The number of major new rules with a cost in

Table 2.3
The Economic Costs of New Major Rules by Year, 1987–2003

<table>
<thead>
<tr>
<th>Year</th>
<th>Cost (in Billions of Dollars)</th>
<th>Number of Rules over $1 Billion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1987</td>
<td>3.6</td>
<td>N/A</td>
</tr>
<tr>
<td>1988</td>
<td>12.5</td>
<td>N/A</td>
</tr>
<tr>
<td>1989</td>
<td>4.1</td>
<td>N/A</td>
</tr>
<tr>
<td>1990</td>
<td>3.8</td>
<td>0</td>
</tr>
<tr>
<td>1991</td>
<td>9.7</td>
<td>2</td>
</tr>
<tr>
<td>1992</td>
<td>16.3</td>
<td>7</td>
</tr>
<tr>
<td>1993</td>
<td>5.1</td>
<td>2</td>
</tr>
<tr>
<td>1994</td>
<td>8.7</td>
<td>2</td>
</tr>
<tr>
<td>1995</td>
<td>3.5</td>
<td>0</td>
</tr>
<tr>
<td>1996</td>
<td>2.6</td>
<td>1</td>
</tr>
<tr>
<td>1997</td>
<td>2.4</td>
<td>0</td>
</tr>
<tr>
<td>1998</td>
<td>5.4</td>
<td>1</td>
</tr>
<tr>
<td>1999</td>
<td>8.4</td>
<td>3</td>
</tr>
<tr>
<td>2000</td>
<td>13.1</td>
<td>4</td>
</tr>
<tr>
<td>2001</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2002</td>
<td>1.9</td>
<td>0</td>
</tr>
<tr>
<td>2003</td>
<td>2.5</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>103.6</td>
<td>23</td>
</tr>
</tbody>
</table>

excess of $1 billion also plummeted from the Clinton administration to the George W. Bush administration.

Is the decline in regulatory activity necessarily a favorable development? Regulatory costs in and of themselves are an undesirable economic burden. However, if the regulations generate benefits that exceed the costs, then it is desirable to promote such regulation rather than discourage it. Thus, the ultimate test of regulatory efficacy requires consideration of both the benefit and cost consequences of regulation.

A less precise tally of trends in regulatory burdens is provided by the index of the number of pages published in the Federal Register. One would expect there to be a correlation between the number of pages devoted to government rules and regulations and the cost these regulations impose. This need not be the case if, for example, agencies become adept at editing their regulatory documents to make them shorter but no less burdensome. Moreover, some Federal Register entries modify regulations and decrease costs rather than increase them. However, it is generally believed that there is a positive, albeit highly imperfect, correlation between the amount of federal regulation published in the Federal Register and the regulatory costs imposed.

Figure 2.4 indicates the trends in these costs for the past half-century. In 1936 the number of pages in the Federal Register was relatively modest—2,599. The pace of regulation

![Figure 2.4](image-url)

**Figure 2.4**
Trends in Federal Register Analysis Pages, 1936–2003

*Source: Office of the Federal Register*
increased steadily but slowly until 1970. It is apparent from figure 2.4 that there was a rapid escalation in regulation beginning in that decade. The 1970s marked the establishment of the new wave of health, safety, and environmental regulation, which greatly expanded the role of the government and its regulatory activities. By 1980 the number of pages in the Federal Register had reached 87,012. The first half of the 1980s marked a decrease in the dissemination of new regulation, which was consistent with the Reagan administration’s efforts to deregulate and roll back regulations. However, by the second term of the Reagan administration there was renewed regulatory activity, which is also reflected in the subsequent increase in the number of pages of regulations published in the Federal Register.

The more recent upward trend in the total number of pages published in the Federal Register is more reflective of the increased volume of regulatory initiatives under the Clinton administration and the George W. Bush administration. Whereas there were about 50,000 pages published during many of the years in the 1980s, since 1993 the total Federal Register page count has ranged from 67,518 to 83,294. How much meaning one should attach to such statistics is unclear. For example, the number of final rules documents appearing in 1997 was 4,615, as compared to a slightly lower figure of 4,581 a decade earlier. Moreover, some years of peak regulatory activity, such as 1980, include statistics that are quite misleading as a measure of regulatory burden. That year featured a flurry of regulatory initiatives at the end of the Carter administration in January 1980, which was subsequently followed by a rescinding of regulations and a major deregulation effort on the part of the Reagan administration later that year. With this principal exception, however, the overall implication of figure 2.4, that regulation has become an increasingly important part of our lives, is certainly valid.

Other measures of regulatory activity have similar implications. The Code of Federal Regulations summarizes the stock of existing regulations, whereas the Federal Register page count provides a measure of the flow of annual regulations. The total number of pages of regulation in the Code of Federal Regulations was under 10,000 in 1950, but had grown to more than 100,000 by 1980. By the end of that decade, the number of pages in the Code of Federal Regulations was just over 50,000, which has been consistent with the effort to scale back the role of regulation, particularly in the transportation area.

The trends in regulatory agency spending shown in table 2.4 show a similar dramatic upward trend. In inflation-adjusted 2000 dollars, total spending by regulatory agencies rose from $2.5 billion in 1960 to $36 billion in 2004–2005. The composition of the spending has also shifted as well. Whereas social regulations comprised 66 percent of all agency spending in 1960, by 2004–2005 this amount had grown to 85 percent. In addition to the jump in spending on social regulation between 1970 and 1980, the decade that marked the establishment of agencies such as the EPA and OSHA, there has also been a tremendous increase in spending on homeland security in response to the September 11, 2001, attack on the World Trade Center. The appendix to this chapter includes trends in agency staffing and detailed
breakdowns of spending patterns that also document the increased prominence of social regulation.

The Character of Regulatory Oversight Actions

It is also instructive to consider the mix of actions undertaken through the regulatory oversight process to obtain an assessment of the nature of the oversight activity that has led to many of these changes. Table 2.5 summarizes the oversight actions undertaken since 1981. When the oversight process began, the OMB approved almost 90 percent of regulations without change. At the present time, the overall approval rate is just 30 percent.

One should be cautious in attributing any change in character of the regulatory oversight process to the trends exhibited by the statistics in table 2.5. A higher percentage of regulations are changed as a result of the current review process, in large part because of the increased selectivity of the regulations that are earmarked for review. The number of executive order reviews plummeted from 2,800 in 1981 to 509 in 1997. The OMB’s review efforts are consequently much more targeted than before, so that one would expect a higher

Table 2.4
Spending Summary for Federal Agencies, Selected Years (Millions of 2000 Dollars)*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Social Regulation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consumer safety and health</td>
<td>485</td>
<td>792</td>
<td>2,136</td>
<td>2,250</td>
<td>3,456</td>
<td>4,640</td>
<td>4,697</td>
<td>4,874</td>
<td>1.2</td>
<td>3.8</td>
</tr>
<tr>
<td>Homeland security</td>
<td>689</td>
<td>1,218</td>
<td>2,938</td>
<td>4,116</td>
<td>7,874</td>
<td>17,883</td>
<td>15,403</td>
<td>15,556</td>
<td>-13.9</td>
<td>1.0</td>
</tr>
<tr>
<td>Transportation</td>
<td>198</td>
<td>643</td>
<td>1,017</td>
<td>981</td>
<td>1,439</td>
<td>1,816</td>
<td>2,082</td>
<td>1,975</td>
<td>14.6</td>
<td>-5.1</td>
</tr>
<tr>
<td>Workplace</td>
<td>171</td>
<td>418</td>
<td>1,384</td>
<td>1,240</td>
<td>1,421</td>
<td>1,583</td>
<td>1,603</td>
<td>1,628</td>
<td>1.2</td>
<td>1.6</td>
</tr>
<tr>
<td>Environment</td>
<td>78</td>
<td>664</td>
<td>2,740</td>
<td>4,284</td>
<td>5,586</td>
<td>5,761</td>
<td>5,341</td>
<td>5,957</td>
<td>-7.3</td>
<td>11.5</td>
</tr>
<tr>
<td>Energy</td>
<td>57</td>
<td>236</td>
<td>808</td>
<td>543</td>
<td>607</td>
<td>659</td>
<td>652</td>
<td>685</td>
<td>-1.0</td>
<td>5.0</td>
</tr>
<tr>
<td><strong>Total Social Regulation</strong></td>
<td>1,678</td>
<td>3,970</td>
<td>11,204</td>
<td>13,413</td>
<td>20,383</td>
<td>32,342</td>
<td>29,778</td>
<td>30,676</td>
<td>-7.9</td>
<td>3.0</td>
</tr>
<tr>
<td><strong>Economic Regulation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finance and banking</td>
<td>195</td>
<td>356</td>
<td>725</td>
<td>1,598</td>
<td>1,856</td>
<td>1,797</td>
<td>1,892</td>
<td>1,908</td>
<td>5.3</td>
<td>0.8</td>
</tr>
<tr>
<td>Industry-specific regulation</td>
<td>432</td>
<td>1,002</td>
<td>869</td>
<td>603</td>
<td>712</td>
<td>791</td>
<td>863</td>
<td>890</td>
<td>9.2</td>
<td>3.1</td>
</tr>
<tr>
<td>General business</td>
<td>228</td>
<td>410</td>
<td>659</td>
<td>888</td>
<td>1,673</td>
<td>2,034</td>
<td>2,235</td>
<td>2,763</td>
<td>9.9</td>
<td>23.6</td>
</tr>
<tr>
<td><strong>Total Economic Regulation</strong></td>
<td>855</td>
<td>1,768</td>
<td>2,253</td>
<td>3,089</td>
<td>4,241</td>
<td>4,622</td>
<td>4,991</td>
<td>5,561</td>
<td>8.0</td>
<td>11.4</td>
</tr>
<tr>
<td><strong>GRAND TOTAL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2,533</td>
<td>5,739</td>
<td>13,457</td>
<td>16,502</td>
<td>24,624</td>
<td>36,964</td>
<td>34,769</td>
<td>36,236</td>
<td>-5.9</td>
<td>4.2</td>
</tr>
</tbody>
</table>

* Numbers may not add to totals due to rounding. Numbers are in millions of dollars in outlays, in constant (real) 2,000 dollars. All years are fiscal years.

Source: Weidenbaum Center, Washington University, and Mercatus Center at George Mason University. Derived from the *Budget of the United States Government* and related documents, various fiscal years.
Table 2.5  
Types of Action Taken by the OMB Regulatory Oversight Process on Agency Rules, 1981–2003 (Percent)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Consistent without change</td>
<td>87.2</td>
<td>70.8</td>
<td>71.8</td>
<td>63.1</td>
<td>64.5</td>
<td>67.3</td>
<td>53.4</td>
<td>53.1</td>
<td>41.4</td>
<td>37.4</td>
<td>36.1</td>
<td>31.5</td>
<td>34.3</td>
<td>28.1</td>
<td>31.7</td>
<td>30.1</td>
</tr>
<tr>
<td>Consistent with change</td>
<td>4.9</td>
<td>23.1</td>
<td>19.3</td>
<td>27.2</td>
<td>25.9</td>
<td>23.5</td>
<td>37.3</td>
<td>39.0</td>
<td>51.5</td>
<td>56.0</td>
<td>59.3</td>
<td>62.2</td>
<td>60.4</td>
<td>45.6</td>
<td>54.3</td>
<td>60.5</td>
</tr>
<tr>
<td>Withdrawn by agency</td>
<td>1.8</td>
<td>3.1</td>
<td>2.5</td>
<td>2.8</td>
<td>4.6</td>
<td>6.0</td>
<td>4.3</td>
<td>5.2</td>
<td>5.1</td>
<td>5.1</td>
<td>3.1</td>
<td>3.9</td>
<td>22.0</td>
<td>7.6</td>
<td>6.9</td>
<td></td>
</tr>
<tr>
<td>Returned for reconsideration</td>
<td>1.6</td>
<td>1.5</td>
<td>1.0</td>
<td>1.1</td>
<td>0.4</td>
<td>0.4</td>
<td>0.0</td>
<td>0.5</td>
<td>0.0</td>
<td>0.8</td>
<td>0.0</td>
<td>0.0</td>
<td>2.6</td>
<td>0.7</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Sent improperly or exempt</td>
<td>3.1</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
<td>0.6</td>
<td>0.5</td>
<td>1.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>1.0</td>
<td>0.2</td>
<td>0.3</td>
<td>0.1</td>
<td>0.8</td>
</tr>
<tr>
<td>Emergency, statutory, or judicial deadline</td>
<td>1.4</td>
<td>1.2</td>
<td>2.5</td>
<td>3.0</td>
<td>1.8</td>
<td>1.8</td>
<td>3.7</td>
<td>2.1</td>
<td>1.8</td>
<td>0.4</td>
<td>0.8</td>
<td>2.2</td>
<td>1.2</td>
<td>1.4</td>
<td>5.5</td>
<td>1.4</td>
</tr>
<tr>
<td>Total</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

*Note:* All data are for calendar years. Some percentages may not sum to 100.0 percent because of rounding. In addition, 1990–1993 include up to 2.7 percent of rules that were suspended.

percentage of the regulations to be revised in response to the review efforts. These expectations are in fact borne out by the data in table 2.5, which indicate that more than 60 percent of all regulations are now altered before being issued by the agency.

Some of these changes have been quite consequential. For example, at OMB’s insistence, OSHA offered firms a variety of alternative means of compliance to reduce the explosion hazards arising from the dust levels in grain mills. This expanded flexibility did not impede the safety effects of the regulation, but it did lower the regulatory costs. Over 90 percent of the regulations are consistent with OMB principles after such changes are made or without change. This high percentage indicates that the dominant emphasis of the OMB process is to promote negotiated solutions to enhance regulatory policy as opposed to simply serving in an obstructionist role. The OMB oversight process has limited political resources, so that it cannot afford to do battle in every regulatory arena, even though few would claim that 90 percent of the regulations proposed will in fact maximize the net benefits to society.

The percentage of instances in which the OMB blocks regulations is quite small. In 2003, for example, 6.9 percent of the regulations reviewed were withdrawn by the regulatory agency and 0.3 percent were returned for consideration. Many of these regulations are among the most burdensome.

Perhaps the most interesting trend exhibited in table 2.5 pertains to the first two rows of the table. The percentage of regulations that are consistent with OMB guidelines without any change dropped by 57 percent from 1981 to 2003, and the percentage of regulations that are consistent with change rose by a comparable amount over that period. The dominant emphasis of OMB actions has been either to approve regulations or to promote moderate modifications of them, and over time there has been an increased attempt to alter regulations in an incremental fashion rather than simply to approve them without any change whatsoever.

Such incremental modifications in regulation are where we would expect the regulatory oversight process to have its greatest influence because major conflicts, such as those over the entire thrust of a regulatory policy, would be escalated to higher political levels. If all regulatory policy decisions were escalated in this manner, the president would have little opportunity to devote time to other national problems. In any year, there are hundreds of major regulations and an even greater number of minor regulations that agencies will issue. In 1997, for example, the OMB reviewed 92 major regulations from the U.S. Department of Health and Human Services and 52 major regulations from the U.S. EPA. Given the substantial volume of regulatory activity, the only feasible way to address these issues is to remain within the interagency negotiations between the regulatory agency and the OMB, saving appeals to a higher level for the small percentage of regulatory issues that involve controversial issues of national policy. In the Reagan administration, one such policy merit-

In this preview copy of Economics of Regulation and Antitrust is provided for course evaluation purposes only. It is not to be reproduced or redistributed. For information on adopting this book for class use, please contact: Textbook Manager, The MIT Press, textbook@mit.edu.
the Clinton administration there was substantial high-level involvement in the rewriting of the Superfund law, which governs the treatment of hazardous wastes. More routine regulations, such as standards for the combustion of municipal waste, are handled without a national debate.

**What Do Regulators Maximize?**

In theory, regulatory agencies serve to maximize the national interest subject to their legislative mandates. Similarly, the OMB is presumably motivated to maximize the net benefits minus costs to society. Such a characterization of regulatory objectives is, unfortunately, excessively naive. There are a number of diverse factors that influence policy decisions, many of which have very little to do with these formal statements of purpose.

What is clear at this stage is that there are certainly influences at work other than those that are formally specified. However, economists have yet to reach a consensus regarding the specific formulation that best captures the political mechanisms at work. A brief review of some of these theories can, however, highlight the range and the types of approaches that have been taken.

**The Capture Theory**

Under the capture theory of regulation, such as that espoused by George Stigler, the regulatory agency is captured by the economic interests that it serves. Stigler has been most successful in testing this model with respect to the economic regulation agencies, such as the Interstate Commerce Commission. Examples of how government regulation can foster industry interests abound. Regulation of airline fares can, for example, provide a floor on airline rates that enables firms to make greater profits than if there were price competition. Similarly, minimum quality standards for products can promote the interests of the more established and advanced firms in the industry, which will use these mandated quality standards to squeeze the producers with less advanced technological capabilities.

Most models based on the capture theory recognize the competing demands on regulatory agencies. Private interests as well as public interests may affect the political survival of the regulatory officials as well as the agency’s budget. Although the most direct descendant of Stigler’s work is that of Peltzman, a number of authors have developed similar models reflecting the diversity of political influences at work. Roger Noll has developed an external signaling theory of regulation whereby regulatory agencies attempt to minimize the conflict-

ing criticism that appears through signals from the economic and social environment in which the regulatory agency operates. 21 Noll proposes that agencies construct an administrative apparatus for the development and enforcement of their regulations to promote the ability of groups that approve their actions and to limit the ability of political forces that disapprove their actions.

Other Theories of Influence Patterns

Other researchers have also formulated models reflecting diverse patterns of influence, but have concluded that there are particular sets of influences that are most influential. For example, Wilson and Stewart suggest that regulatory agencies have substantial discretion with respect to the regulatory actions they take, so that it is the regulatory agency that plays the dominant role. 22 Other authors have advocated a quite different view in which Congress has the dominant role, not the regulatory agency. 23 The leverage of Congress stems from the fact that the congressional committees are responsible for setting the budgets of the regulatory agencies and for confirming the leading administrators in these agencies.

Comprehensive Models of Regulatory Objectives

In all likelihood, the actual outcomes are influenced by a multiplicity of factors that cannot be characterized by any simple, single model. The regulatory agency does not have sole control, nor does the OMB. Moreover, Congress and the judiciary play a restraining role, and lobbyists for and against the regulation can affect the political payoffs to the regulatory agency as well. The actual strength of the influences undoubtedly varies depending on the particular context.

An interesting case study of the extent to which there are multiple influences at work is provided through detailed analysis of the rulemaking process for the EPA regulations that implemented the industrial effluent standards that are used to control water pollution. The study by Magat, Krupnick, and Harrington highlights the types of outcomes that will ultimately be explained through an analysis of the competing interests affecting regulatory outcomes:


The factors determining the outcomes of EPA's effluent standard-setting process are by no means self-evident. For instance, on December 7, 1973, EPA proposed effluent discharge standards for water pollution from the leather tanning industry. These standards required that by 1977 discharges of biological oxygen demand (BOD) not exceed 40 milligrams per liter (mg/l) of waste water. Four months and two days later, EPA promulgated the final BOD standard for the industry of 102 mg/l. Why was the stringency of the standard weakened by 155 percent between its initial proposal and final promulgation? Why did EPA issue a tighter final standard for the meat packing industry, which produces wastes with similar characteristics to leather tanning, of only 24 mg/l BOD? And why did smaller firms receive weaker regulations?24

The heterogeneity of the regulation in different industries and for firms of different sizes clearly suggests that there is no simple or naive regulatory objective guiding behavior. Through detailed statistical analysis of a series of decisions made by EPA as part of this rule-making process, Magat et al. have identified a variety of factors that were influential in the setting of these water pollution standards.

One such influence was efficiency concerns. EPA did adjust the stringency of regulations in different industries to reflect the differences in compliance costs across firms. This is the kind of heterogeneity one would want to promote, in that standards should not be as stringent for industries that must bear greater burdens to reduce pollution. In those contexts, the costs of compliance will be greater, so that to maximize the net benefits of the standard one would want to reflect these cost differences in the standard level.

Second, the quality of the economic analysis supporting the standard also was influential. Standards supported by high-quality economic analyses were more likely to lead to more stringent effluent guidelines than those lacking substantive support. This result as well suggests that there is a sense of economic rationality to the process whereby the strength of the analysis does affect the policy outcome. It should be noted, however, that the particular price and cost effects of the regulation did not appear to be as influential as the overall quality of the economic analysis.

Other players have an impact as well. The economic resources of the trade association for the particular industry affect the stringency of the standards in the expected manner. In particular, industries with large budgets for their trade association are able to obtain weaker standards, after taking into account other factors that should determine the stringency of the regulation. The total financial resources appear to be much more influential than the volume of industry comments provided, in that these resources presumably reflect the political clout of the agency to a greater degree than does the number of pages of comments submitted.

Conclusion

In later chapters we will develop a series of models of the regulatory process. All such models should be viewed as a simplification of the actual objectives guiding the regulatory agencies. Economists have made substantial progress in recent decades in developing approaches to indicate how regulators make decisions, which is often quite different than one would predict based on their legislative mandates or their stated agency objectives. A variety of political factors also are at work and will affect the policy outcomes that result.

Despite the multiplicity of these influences, one should not understate the pivotal role that legislative mandates have. These mandates, which are written by Congress, in many circumstances define the terms of the regulatory debate and impose stringent limits on the scope of discretion of the regulatory officials. It is through these mandates that Congress has a long-run influence on regulatory policy, even though most short-run regulatory decisions appear to be governed by actions of the regulatory agency, the influence of the regulatory oversight process, and recognition of the political factors at stake in the regulatory policy decision.

Questions and Problems

1. A frequent proposal has been to replace the oversight process through a system known as a “regulatory budget.” Each agency would be assigned a total cost that it could impose on the American economy, and its task would be to select the regulations that best foster the national interest subject to this cost. Can you identify any problems with the regulatory budget approach? How feasible do you believe it would be to calculate the costs of all the regulations of a particular agency? What, for example, are the costs associated with affirmative action? Are they positive or negative?

2. Inadequacies in government action are frequently called “government failure.” In some cases, government failures reinforce market failures. In particular, the government may promote inefficient outcomes in a way that exacerbates the shortcomings of the market rather than alleviates these shortcomings. Can you think of any examples where such mutually reinforcing failures might occur and the reasons why they might occur?

3. One justification often given for the utilization of a variety of conservatism factors in risk analyses is that society is risk-averse, so that we should be conservative. Can you identify any flaws in this reasoning?

4. Regulatory agencies are not permitted to publicly release the details of their regulatory proposals until after the appropriate review by the OMB, as outlined in figure 2.2. How do you believe the process would change if the agency first issued the proposal publicly and then began its discussions with the OMB? Do you believe this change would improve the regulatory decision-making process? What new factors would be brought to bear?

5. What are the problems in using measures such as Federal Register page counts to assess the costs imposed by regulation? In the chapter as well as in the appendix, the measures of
regulatory trends include Federal Register page counts, page counts from the Code of Federal Regulations, agency budget trends, and agency staffing trends. Which of these sets of information do you believe is most informative with respect to the regulatory costs imposed on society? What other measures do you believe would be useful in assessing the changing regulatory burden?

6. In your view, what is the appropriate rate of discount for regulatory policies? Suppose that the measure is the real rate of return to capital. How would you measure this? If a group of economists were given the task, do you believe they would all arrive at the same answer? Why might there be differences in the discount rate estimate?

Appendix: Trends in Regulatory Agency Budgets and Staff

An instructive measure of the changing role of government regulation is provided by the magnitude of government expenditures in this area. Although the principal costs of regulations are those borne by business and the public at large, the levels of the budgets of the regulatory agencies do provide some index of the degree of regulatory activity.

The Weidenbaum Center at Washington University, which was directed by Murray Weidenbaum (chairman of President Reagan’s Council of Economic Advisors), regularly compiles a series of tables summarizing these budgetary and staffing trends. This compilation is now done in conjunction with the Mercatus Center at George Mason University, which functions as a university-based counterpart of OIRA. The Mercatus Center also provides detailed comments on regulatory proposals. Tables A.1 and A.2 summarize the key data. Table A.1 reviews the staffing trends, and table A.2 provides a very detailed breakdown of agency budgetary trends. These patterns are generally consistent with those displayed by the Federal Register page counts. Regulation accelerated dramatically in the 1970s, as there was a substantial growth in the health, safety, and environmental regulation agencies. The deregulation in the transportation fields in the 1980s, coupled with the moderation in the health, safety, and environmental regulation area, led to some reduction in the regulatory effort in the early 1980s. However, there is some evidence of a resurgence in regulation in the latter 1980s and early 1990s.
### Table A.1
Staffing Summary for Federal Regulatory Agencies, Selected Years*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Social Regulation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consumer safety and health</td>
<td>13,912</td>
<td>33,242</td>
<td>28,730</td>
<td>31,150</td>
<td>34,438</td>
<td>34,314</td>
<td>35,048</td>
</tr>
<tr>
<td>Homeland security</td>
<td>22,510</td>
<td>37,316</td>
<td>46,512</td>
<td>60,347</td>
<td>123,504</td>
<td>121,912</td>
<td>123,499</td>
</tr>
<tr>
<td>Transportation</td>
<td>7,614</td>
<td>8,401</td>
<td>7,497</td>
<td>8,944</td>
<td>8,889</td>
<td>8,779</td>
<td>8,670</td>
</tr>
<tr>
<td>Workplace</td>
<td>6,486</td>
<td>17,894</td>
<td>13,610</td>
<td>12,141</td>
<td>12,052</td>
<td>12,192</td>
<td>12,312</td>
</tr>
<tr>
<td>Environment</td>
<td>4,525</td>
<td>16,993</td>
<td>20,057</td>
<td>24,555</td>
<td>25,537</td>
<td>25,757</td>
<td>25,751</td>
</tr>
<tr>
<td>Energy</td>
<td>219</td>
<td>3,225</td>
<td>3,293</td>
<td>2,923</td>
<td>3,048</td>
<td>3,149</td>
<td>3,199</td>
</tr>
<tr>
<td><strong>Total Social Regulation</strong></td>
<td>55,266</td>
<td>117,071</td>
<td>119,699</td>
<td>140,060</td>
<td>207,468</td>
<td>206,103</td>
<td>208,479</td>
</tr>
<tr>
<td><strong>Economic Regulation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finance and banking</td>
<td>4,969</td>
<td>9,524</td>
<td>16,353</td>
<td>14,188</td>
<td>12,375</td>
<td>12,604</td>
<td>12,562</td>
</tr>
<tr>
<td>Industry-specific regulation</td>
<td>18,548</td>
<td>11,885</td>
<td>7,977</td>
<td>6,438</td>
<td>6,504</td>
<td>6,570</td>
<td>6,667</td>
</tr>
<tr>
<td>General business</td>
<td>6,609</td>
<td>9,251</td>
<td>9,611</td>
<td>12,509</td>
<td>13,324</td>
<td>14,347</td>
<td>14,765</td>
</tr>
<tr>
<td><strong>Total Economic Regulation</strong></td>
<td>30,126</td>
<td>30,660</td>
<td>33,941</td>
<td>33,135</td>
<td>32,203</td>
<td>33,521</td>
<td>33,994</td>
</tr>
<tr>
<td><strong>GRAND TOTAL</strong></td>
<td>85,392</td>
<td>147,731</td>
<td>153,640</td>
<td>173,195</td>
<td>239,671</td>
<td>239,624</td>
<td>242,473</td>
</tr>
</tbody>
</table>

**Annualized percentage change**

|               | 5.6   | 0.4   | 1.2   | 11.4  | 0.0   | 1.2 |

* Employment described is full-time equivalent employment; years are fiscal years.

**Source:** Weidenbaum Center, Washington University, and Mercatus Center at George Mason University. Derived from the *Budget of the United States Government* and related documents, various fiscal years.
Table A.2
Agency Detail of Spending on Federal Regulatory Activity: Current Dollars (Fiscal Years, Millions of Dollars in “Outlays”)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Social Regulation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Consumer Safety and Health</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consumer Product Safety Commission</td>
<td>n/o</td>
<td>n/o</td>
<td>44</td>
<td>35</td>
<td>51</td>
<td>62</td>
<td>63</td>
<td>66</td>
<td>1.6</td>
<td>4.8</td>
</tr>
<tr>
<td><strong>Department of Agriculture:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Animal and Plant Health Inspection Service</td>
<td>59</td>
<td>96</td>
<td>257</td>
<td>406</td>
<td>735</td>
<td>1,180</td>
<td>1,310</td>
<td>1,154</td>
<td>11.0</td>
<td>−11.9</td>
</tr>
<tr>
<td>Food Safety and Inspection Service</td>
<td>n/o</td>
<td>n/o</td>
<td>393</td>
<td>475</td>
<td>743</td>
<td>841</td>
<td>887</td>
<td>950</td>
<td>5.5</td>
<td>7.1</td>
</tr>
<tr>
<td>Grain Inspection, Packers and Stockyards</td>
<td>n/o</td>
<td>3</td>
<td>66</td>
<td>50</td>
<td>60</td>
<td>66</td>
<td>82</td>
<td>85</td>
<td>24.2</td>
<td>3.7</td>
</tr>
<tr>
<td><strong>Subtotal</strong></td>
<td>59</td>
<td>99</td>
<td>716</td>
<td>931</td>
<td>1,538</td>
<td>2,087</td>
<td>2,279</td>
<td>2,189</td>
<td>9.2</td>
<td>−3.9</td>
</tr>
<tr>
<td><strong>Department of Health and Human Services:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Food and Drug Administration</td>
<td>16</td>
<td>68</td>
<td>326</td>
<td>561</td>
<td>1,209</td>
<td>1,736</td>
<td>1,662</td>
<td>1,844</td>
<td>−4.3</td>
<td>11.0</td>
</tr>
<tr>
<td><strong>Department of Housing and Urban Development:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consumer Protection Programs</td>
<td>n/o</td>
<td>n/o</td>
<td>2</td>
<td>5</td>
<td>14</td>
<td>7</td>
<td>13</td>
<td>13</td>
<td>85.7</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Department of Justice:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drug Enforcement Administration</td>
<td>n/o</td>
<td>2</td>
<td>13</td>
<td>27</td>
<td>74</td>
<td>81</td>
<td>102</td>
<td>137</td>
<td>25.9</td>
<td>34.3</td>
</tr>
<tr>
<td>Bureau of Alcohol. Tobacco, Firearms, Explosives</td>
<td>27</td>
<td>49</td>
<td>147</td>
<td>273</td>
<td>555</td>
<td>828</td>
<td>795</td>
<td>914</td>
<td>−4.0</td>
<td>15.0</td>
</tr>
<tr>
<td><strong>Subtotal</strong></td>
<td>27</td>
<td>51</td>
<td>160</td>
<td>300</td>
<td>629</td>
<td>909</td>
<td>897</td>
<td>1,051</td>
<td>−1.3</td>
<td>17.2</td>
</tr>
<tr>
<td><strong>Department of Treasury:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alcohol and Tobacco Tax and Trade Bureau</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>79</td>
<td>80</td>
<td>84</td>
<td>1.3</td>
<td>5.0</td>
</tr>
<tr>
<td>Chemical Safety and Hazard Investigation Board</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Federal Mine Safety and Health Review Commission</td>
<td>n/o</td>
<td>n/o</td>
<td>4</td>
<td>4</td>
<td>7</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>33.3</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>TOTAL—Consumer Safety and Health</strong></td>
<td>102</td>
<td>218</td>
<td>1,252</td>
<td>1,836</td>
<td>3,456</td>
<td>4,895</td>
<td>5,011</td>
<td>5,264</td>
<td>2.4</td>
<td>5.0</td>
</tr>
<tr>
<td><strong>Homeland Security</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Department of Homeland Security:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Area Maritime Security</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>52.0</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Customs and Border Protection</td>
<td>62</td>
<td>175</td>
<td>837</td>
<td>1,664</td>
<td>2,802</td>
<td>5,294</td>
<td>7,008</td>
<td>6,487</td>
<td>32.4</td>
<td>−7.4</td>
</tr>
<tr>
<td>Immigration and Customs Enforcement</td>
<td>38</td>
<td>66</td>
<td>254</td>
<td>786</td>
<td>3,355</td>
<td>2,218</td>
<td>2,334</td>
<td>2,818</td>
<td>5.2</td>
<td>20.7</td>
</tr>
<tr>
<td>Coast Guard</td>
<td>45</td>
<td>94</td>
<td>498</td>
<td>909</td>
<td>1,717</td>
<td>1,680</td>
<td>1,795</td>
<td>2,092</td>
<td>6.8</td>
<td>16.5</td>
</tr>
<tr>
<td>Category</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>7</td>
<td>109</td>
<td>87</td>
<td>1457.1</td>
<td>−20.2</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>----</td>
<td>-----</td>
<td>-----</td>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>Science and Technology</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>7</td>
<td>109</td>
<td>87</td>
<td>1457.1</td>
<td>−20.2</td>
</tr>
<tr>
<td>Transportation Security Administration</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>7</td>
<td>109</td>
<td>87</td>
<td>1457.1</td>
<td>−20.2</td>
</tr>
<tr>
<td>TOTAL—Homeland Security</td>
<td>145</td>
<td>335</td>
<td>1,589</td>
<td>3,359</td>
<td>7,874</td>
<td>18,865</td>
<td>16,434</td>
<td>16,799</td>
<td>−12.9</td>
<td>2.2</td>
</tr>
<tr>
<td><strong>Transportation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Department of Transportation:</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Federal Aviation Administration</td>
<td>42</td>
<td>124</td>
<td>283</td>
<td>477</td>
<td>881</td>
<td>1,144</td>
<td>1,175</td>
<td>1,187</td>
<td>2.7</td>
<td>1.0</td>
</tr>
<tr>
<td>Federal Highway Administration</td>
<td>n/o</td>
<td>6</td>
<td>20</td>
<td>98</td>
<td>9</td>
<td>17</td>
<td>17</td>
<td>9</td>
<td>*</td>
<td>0.0</td>
</tr>
<tr>
<td>Federal Motor Carrier Safety Administration</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>163</td>
<td>274</td>
<td>513</td>
<td>418</td>
<td>87.2</td>
</tr>
<tr>
<td>Federal Railroad Administration</td>
<td>n/o</td>
<td>16</td>
<td>92</td>
<td>52</td>
<td>119</td>
<td>154</td>
<td>199</td>
<td>181</td>
<td>29.2</td>
<td>−9.0</td>
</tr>
<tr>
<td>National Highway Traffic Safety Administration</td>
<td>n/o</td>
<td>26</td>
<td>137</td>
<td>148</td>
<td>173</td>
<td>233</td>
<td>223</td>
<td>253</td>
<td>−4.3</td>
<td>13.5</td>
</tr>
<tr>
<td>Surface Transportation Board</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>17</td>
<td>23</td>
<td>21</td>
<td>20</td>
<td>−8.7</td>
</tr>
<tr>
<td><strong>Subtotal</strong></td>
<td>42</td>
<td>172</td>
<td>532</td>
<td>775</td>
<td>1,362</td>
<td>1,845</td>
<td>2,148</td>
<td>2,059</td>
<td>16.4</td>
<td>−4.1</td>
</tr>
<tr>
<td>National Transportation Safety Board</td>
<td>n/o</td>
<td>5</td>
<td>18</td>
<td>26</td>
<td>77</td>
<td>71</td>
<td>73</td>
<td>74</td>
<td>2.8</td>
<td>1.4</td>
</tr>
<tr>
<td><strong>TOTAL—Transportation</strong></td>
<td>42</td>
<td>177</td>
<td>550</td>
<td>801</td>
<td>1,439</td>
<td>1,916</td>
<td>2,221</td>
<td>2,133</td>
<td>15.9</td>
<td>−4.0</td>
</tr>
<tr>
<td><strong>Job Safety and Other Working Conditions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Department of Labor:</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employment Standards Administration</td>
<td>14</td>
<td>36</td>
<td>123</td>
<td>156</td>
<td>227</td>
<td>248</td>
<td>256</td>
<td>258</td>
<td>3.2</td>
<td>0.8</td>
</tr>
<tr>
<td>Office of the American Workplace</td>
<td>n/o</td>
<td>12</td>
<td>55</td>
<td>79</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>Employee Benefits Security Administration</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>99</td>
<td>130</td>
<td>145</td>
<td>148</td>
<td>11.5</td>
</tr>
<tr>
<td>Mine Safety and Health Administration</td>
<td>7</td>
<td>17</td>
<td>142</td>
<td>171</td>
<td>225</td>
<td>266</td>
<td>266</td>
<td>274</td>
<td>0.0</td>
<td>3.0</td>
</tr>
<tr>
<td>Occupational Safety and Health Admin.</td>
<td>n/o</td>
<td>n/o</td>
<td>180</td>
<td>275</td>
<td>370</td>
<td>465</td>
<td>460</td>
<td>465</td>
<td>−1.1</td>
<td>1.1</td>
</tr>
<tr>
<td><strong>Subtotal</strong></td>
<td>21</td>
<td>65</td>
<td>500</td>
<td>681</td>
<td>921</td>
<td>1,109</td>
<td>1,127</td>
<td>1,145</td>
<td>1.6</td>
<td>1.6</td>
</tr>
<tr>
<td>Architectural and Transportation Barriers</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Compliance Board</td>
<td>n/o</td>
<td>12</td>
<td>131</td>
<td>181</td>
<td>290</td>
<td>315</td>
<td>325</td>
<td>347</td>
<td>3.2</td>
<td>6.8</td>
</tr>
<tr>
<td>Equal Employment Opportunity Commission</td>
<td>15</td>
<td>38</td>
<td>109</td>
<td>142</td>
<td>198</td>
<td>231</td>
<td>242</td>
<td>249</td>
<td>4.8</td>
<td>2.9</td>
</tr>
<tr>
<td>National Labor Relations Board</td>
<td>n/o</td>
<td>n/o</td>
<td>8</td>
<td>6</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>11.1</td>
<td>10.0</td>
</tr>
<tr>
<td><strong>TOTAL—Job Safety and Other Working Conditions</strong></td>
<td>36</td>
<td>115</td>
<td>748</td>
<td>1,012</td>
<td>1,421</td>
<td>1,670</td>
<td>1,710</td>
<td>1,758</td>
<td>2.4</td>
<td>2.8</td>
</tr>
<tr>
<td>-------------------------------------------------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------------</td>
<td>----------------</td>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td><strong>Environment</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Council on Environmental Quality</td>
<td>n/o</td>
<td>n/o</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td><strong>Department of Agriculture:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forest and Rangeland Research</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>232</td>
<td>305</td>
<td>324</td>
<td>328</td>
<td>6.2</td>
<td></td>
</tr>
<tr>
<td><strong>Department of Defense:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Army Corps of Engineers</td>
<td>1</td>
<td>2</td>
<td>41</td>
<td>66</td>
<td>111</td>
<td>141</td>
<td>145</td>
<td>151</td>
<td>2.8</td>
<td></td>
</tr>
<tr>
<td><strong>Department of Interior:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fish and Wildlife Service</td>
<td>3</td>
<td>7</td>
<td>71</td>
<td>152</td>
<td>236</td>
<td>288</td>
<td>293</td>
<td>299</td>
<td>1.7</td>
<td></td>
</tr>
<tr>
<td>Office of Surface Mining Reclamation and</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enforcement</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>85</td>
<td>327</td>
<td>392</td>
<td>384</td>
<td>318</td>
<td>-17.2</td>
<td></td>
</tr>
<tr>
<td>U.S. Geological Survey</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>134</td>
<td>171</td>
<td>161</td>
<td>170</td>
<td>-5.8</td>
<td></td>
</tr>
<tr>
<td><strong>Subtotal</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Environmental Protection Agency</td>
<td>13</td>
<td>174</td>
<td>1,282</td>
<td>2,950</td>
<td>4,478</td>
<td>4,785</td>
<td>4,454</td>
<td>5,121</td>
<td>-6.9</td>
<td></td>
</tr>
<tr>
<td><strong>TOTAL—Environment</strong></td>
<td>17</td>
<td>183</td>
<td>1,482</td>
<td>3,496</td>
<td>5,586</td>
<td>6,077</td>
<td>5,698</td>
<td>6,433</td>
<td>-6.2</td>
<td></td>
</tr>
<tr>
<td><strong>Energy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Department of Energy:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Petroleum Regulation</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>14</td>
<td>24</td>
<td>23</td>
<td>22</td>
<td>23</td>
<td>-4.3</td>
<td></td>
</tr>
<tr>
<td>Federal Inspector for the Alaska Natural Gas</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>5</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>Pipeline</td>
<td>n/o</td>
<td>n/o</td>
<td>54</td>
<td>38</td>
<td>103</td>
<td>97</td>
<td>59</td>
<td>58</td>
<td>-39.2</td>
<td></td>
</tr>
<tr>
<td>Energy Conservation</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>59</td>
<td>52</td>
<td>127</td>
<td>120</td>
<td>81</td>
<td>-32.5</td>
<td></td>
</tr>
<tr>
<td><strong>Subtotal</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nuclear Regulatory Commission</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>65</td>
<td>378</td>
<td>391</td>
<td>480</td>
<td>575</td>
<td>615</td>
<td>659</td>
<td>7.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TOTAL—Energy</strong></td>
<td>12</td>
<td>65</td>
<td>437</td>
<td>443</td>
<td>607</td>
<td>695</td>
<td>696</td>
<td>740</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td><strong>TOTAL SOCIAL REGULATION</strong></td>
<td>353</td>
<td>1,093</td>
<td>6,057</td>
<td>10,947</td>
<td>20,383</td>
<td>34,118</td>
<td>31,770</td>
<td>33,127</td>
<td>-6.9</td>
<td></td>
</tr>
</tbody>
</table>
### Economic Regulation

#### Finance and Banking

*Department of the Treasury:*
- Comptroller of the Currency
- Office of Thrift Supervision

<table>
<thead>
<tr>
<th></th>
<th>429</th>
<th>460</th>
<th>471</th>
<th>7.2</th>
<th>2.4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>11</td>
<td>32</td>
<td>113</td>
<td>256</td>
<td>382</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>21</td>
<td>20</td>
<td>249</td>
<td>159</td>
</tr>
</tbody>
</table>

**Subtotal**

<table>
<thead>
<tr>
<th></th>
<th>20</th>
<th>53</th>
<th>133</th>
<th>505</th>
<th>541</th>
</tr>
</thead>
<tbody>
<tr>
<td>Farm Credit Administration</td>
<td>2</td>
<td>4</td>
<td>12</td>
<td>36</td>
<td>32</td>
</tr>
<tr>
<td>Federal Deposit Insurance Corporation</td>
<td>13</td>
<td>30</td>
<td>121</td>
<td>476</td>
<td>580</td>
</tr>
<tr>
<td>Federal Housing Finance Board</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>1</td>
<td>18</td>
</tr>
</tbody>
</table>

**Federal Reserve System**
- Federal Reserve Banks
- Federal Reserve System Board of Governors

<table>
<thead>
<tr>
<th></th>
<th>3</th>
<th>5</th>
<th>19</th>
<th>30</th>
<th>79</th>
<th>110</th>
<th>116</th>
<th>116</th>
</tr>
</thead>
<tbody>
<tr>
<td>National Credit Union Administration</td>
<td>3</td>
<td>5</td>
<td>105</td>
<td>242</td>
<td>616</td>
<td>613</td>
<td>636</td>
<td>636</td>
</tr>
</tbody>
</table>

**TOTAL—Finance and Banking**

<table>
<thead>
<tr>
<th></th>
<th>41</th>
<th>98</th>
<th>392</th>
<th>1,304</th>
<th>1,856</th>
<th>1,896</th>
<th>2,019</th>
<th>2,060</th>
</tr>
</thead>
<tbody>
<tr>
<td>Industry-Specific Regulation</td>
<td>42</td>
<td>186</td>
<td>60</td>
<td>161</td>
<td>198</td>
<td>184</td>
<td>247</td>
<td>258</td>
</tr>
</tbody>
</table>

*Department of Agriculture:*
- Agriculture Marketing Service

<table>
<thead>
<tr>
<th></th>
<th>132</th>
<th>16</th>
<th>2</th>
<th>1</th>
<th>1</th>
<th>n/o</th>
<th>0.0</th>
<th>—</th>
</tr>
</thead>
<tbody>
<tr>
<td>Economic Regulatory Administration</td>
<td>n/o</td>
<td>n/o</td>
<td>16</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>n/o</td>
<td>0.0</td>
</tr>
<tr>
<td>Civil Aeronautics Board</td>
<td>7</td>
<td>11</td>
<td>28</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>—</td>
</tr>
<tr>
<td>Commodity Futures Trading Commission</td>
<td>1</td>
<td>2</td>
<td>16</td>
<td>36</td>
<td>62</td>
<td>83</td>
<td>91</td>
<td>94</td>
</tr>
<tr>
<td>Federal Communications Commission</td>
<td>11</td>
<td>24</td>
<td>76</td>
<td>108</td>
<td>269</td>
<td>351</td>
<td>361</td>
<td>381</td>
</tr>
<tr>
<td>Federal Energy Regulatory Commission</td>
<td>7</td>
<td>18</td>
<td>67</td>
<td>113</td>
<td>167</td>
<td>198</td>
<td>203</td>
<td>209</td>
</tr>
<tr>
<td>Federal Maritime Commission</td>
<td>n/o</td>
<td>4</td>
<td>11</td>
<td>15</td>
<td>14</td>
<td>17</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>Interstate Commerce Commission</td>
<td>20</td>
<td>27</td>
<td>80</td>
<td>43</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>—</td>
</tr>
<tr>
<td>Renegotiation Board</td>
<td>3</td>
<td>4</td>
<td>L</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>—</td>
</tr>
</tbody>
</table>

**TOTAL—Industry-Specific Regulation**

<table>
<thead>
<tr>
<th></th>
<th>91</th>
<th>276</th>
<th>470</th>
<th>492</th>
<th>712</th>
<th>834</th>
<th>921</th>
<th>961</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>34.2</td>
<td>4.5</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>
Table A.2 (continued)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>General Business</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost Accounting Standards Board</td>
<td>n/o</td>
<td>n/o</td>
<td>1</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Council on Wage and Price Stability</td>
<td>n/o</td>
<td>n/o</td>
<td>9</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Department of Commerce:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>International Trade Administration</td>
<td>3</td>
<td>6</td>
<td>13</td>
<td>21</td>
<td>35</td>
<td>43</td>
<td>62</td>
<td>67</td>
<td>44.2</td>
<td>8.1</td>
</tr>
<tr>
<td>Bureau of Industry and Security</td>
<td>n/o</td>
<td>n/o</td>
<td>n/o</td>
<td>41</td>
<td>61</td>
<td>68</td>
<td>80</td>
<td>85</td>
<td>17.6</td>
<td>6.3</td>
</tr>
<tr>
<td>Patent and Trademark Office</td>
<td>22</td>
<td>49</td>
<td>103</td>
<td>320</td>
<td>872</td>
<td>1,145</td>
<td>1,096</td>
<td>1,339</td>
<td>44.2</td>
<td>8.1</td>
</tr>
<tr>
<td>Subtotal</td>
<td>25</td>
<td>55</td>
<td>116</td>
<td>382</td>
<td>968</td>
<td>1,256</td>
<td>1,238</td>
<td>1,491</td>
<td>-1.4</td>
<td>20.4</td>
</tr>
<tr>
<td>Department of Justice:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antitrust Division</td>
<td>5</td>
<td>9</td>
<td>49</td>
<td>47</td>
<td>102</td>
<td>117</td>
<td>114</td>
<td>136</td>
<td>-2.6</td>
<td>19.3</td>
</tr>
<tr>
<td>Federal Election Commission</td>
<td>n/o</td>
<td>n/o</td>
<td>10</td>
<td>15</td>
<td>40</td>
<td>46</td>
<td>53</td>
<td>55</td>
<td>15.2</td>
<td>3.8</td>
</tr>
<tr>
<td>Federal Trade Commission</td>
<td>7</td>
<td>20</td>
<td>69</td>
<td>71</td>
<td>125</td>
<td>166</td>
<td>200</td>
<td>204</td>
<td>20.5</td>
<td>2.0</td>
</tr>
<tr>
<td>International Trade Commission</td>
<td>2</td>
<td>4</td>
<td>14</td>
<td>37</td>
<td>47</td>
<td>52</td>
<td>60</td>
<td>61</td>
<td>15.4</td>
<td>1.7</td>
</tr>
<tr>
<td>Library of Congress:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Copyright Office</td>
<td>1</td>
<td>3</td>
<td>14</td>
<td>19</td>
<td>34</td>
<td>40</td>
<td>46</td>
<td>51</td>
<td>15.0</td>
<td>10.9</td>
</tr>
<tr>
<td>Securities and Exchange Commission</td>
<td>8</td>
<td>22</td>
<td>74</td>
<td>154</td>
<td>357</td>
<td>469</td>
<td>674</td>
<td>986</td>
<td>43.7</td>
<td>46.3</td>
</tr>
<tr>
<td>TOTAL—General Business</td>
<td>48</td>
<td>113</td>
<td>356</td>
<td>725</td>
<td>1,673</td>
<td>2,146</td>
<td>2,385</td>
<td>2,984</td>
<td>11.1</td>
<td>25.1</td>
</tr>
<tr>
<td>TOTAL ECONOMIC REGULATION</td>
<td>180</td>
<td>487</td>
<td>1,218</td>
<td>2,521</td>
<td>4,241</td>
<td>4,876</td>
<td>5,325</td>
<td>6,005</td>
<td>9.2</td>
<td>12.8</td>
</tr>
<tr>
<td>GRAND TOTAL</td>
<td>533</td>
<td>1,580</td>
<td>7,275</td>
<td>13,468</td>
<td>24,624</td>
<td>38,994</td>
<td>37,095</td>
<td>39,132</td>
<td>-4.9</td>
<td>5.5</td>
</tr>
</tbody>
</table>

Notes:  
L = less than $500,000; n/o = agency not operational;  
* Agency no longer has regulatory functions.

Source: Weidenbaum Center, Washington University, and Mercatus Center at George Mason University. Derived from the *Budget of the United States Government* and related documents, various fiscal years.
3

Introduction to Antitrust

There is no such thing as “unconstrained competition” in the modern economy. A firm is not allowed to blow up a rival’s factory or renege on agreements with customers. The government is relied upon to enforce property rights and contractual arrangements. The issue, then, is not whether the government has a role in the economy but rather the extent of its role. As a general rule, as long as property rights and contracts are respected, most economists think firms and consumers should be left unconstrained. But there are always exceptions to any general rule. Parts II and III explore some of those that provide a rationale for governments to regulate price, profit, product standards, worker safety, pollution, entry, exit, and the like.

Here in part I, we will focus on the unregulated sector of the economy, where we rely on competition to be the primary mechanism to produce good economic results. What we will find, however, is that even here some constraints need to be placed on what firms do. This is the role of antitrust law and policy. It is to ensure that competition does not evaporate because firms decide to merge into a monopoly, or because a firm that legitimately achieved monopoly power, tries to illegitimately perpetuate or extend that monopoly power, or because firms decide to cooperate rather than compete (as in the lysine cartel, where the cartel’s slogan was “The competitor is our friend, and the customer is our enemy”).

To understand the implications of existing antitrust policy and to design better policies, we need an understanding of the circumstances under which anticompetitive behavior might emerge. What conditions are ripe for it? What should we look for? And how can we correct it? This is the role played by the field in economics known as industrial organization (also called industrial economics). By developing theoretical models of an industry and empirically analyzing actual industries, industrial organization economists seek to answer questions such as, What determines the extent of competition? What is the effect of the number of firms on prices, investment, product variety, and other important variables? What industry conditions are conducive to cartel formation? What determines the number of firms in an industry? How can incumbent firms deter entry and promote exit? Under what conditions should we observe an industry being dominated by a firm? If we observe such dominance, is it bad for society? And, more broadly, when does an industry fall well short of a social welfare optimum?

The objective of this chapter is to provide a brief overview of some of the important concepts that have been developed by industrial organization specialists. We will also provide background information on the antitrust laws and their enforcement. The remainder of part I will develop the necessary knowledge of industrial organization and then use it to explore antitrust law and policy.
In microeconomic theory, one begins by analyzing the market structures of monopoly and perfect competition, but all of the action in real economies actually lies between those two extremes. Most industries are characterized as having multiple firms, often of drastically varying sizes, with some or all having market power, that is, the ability to raise price above their competitors’ prices and still have positive demand. Such a situation is known as an oligopoly or imperfect competition, and modeling and understanding those industries is the primary task of industrial organization.

The field of industrial organization began with research by economists at Harvard University in the 1930s and 1940s. They developed a general approach to the economic analysis of markets that is based on three key concepts: (1) structure, (2) conduct (or behavior), and (3) performance. They hypothesized a causal relationship between these three concepts: structure (number of sellers, ease of entry, etc.) determines firm conduct (pricing, advertising, etc.), which then determines market performance (efficiency, technical progress). Known as the structure-conduct-performance paradigm (SCPP), it is depicted in figure 3.1.

During the 1950 and 1960s, empirical work based on this framework sought to identify general relationships that would hold for all industries, such as a general coefficient that would indicate how adding one more firm would affect price. Time has shown that such a research program was misguided. Industrial organization economists now recognize that each industry is too idiosyncratic for us to think there is such a general stable relationship that would be applicable to many industries.

It was also found that the causal story told above is too simplistic; there are more causal relationships running around then originally described. In figure 3.1, the dashed arrow between the conduct and structure blocks indicates that conduct can sometimes “feed back” to change structure. There are a number of ways in which the behavior of existing firms in a market can affect future market structure. For example, through investing in research and development, a firm can lower its cost to a point where it can profitably price its competitors out of the market. Alternatively, firms can influence market structure by affecting the decisions of potential entrants to enter through the strategic manipulation of price or capital. Perhaps the bluntest way in which conduct affects structure is through merger. Because causality runs in many directions, the SCPP is of limited use in many predictions.

Although the SCPP is no longer the foundation for theory and empirical work in industrial organization, the categories of structure, conduct, and performance remain useful in organizing knowledge about an industry. These three elements will be examined in detail in later chapters; a short description of them is given here.
Theories of oligopoly typically assume sellers of equal size, and thereby specify only the number of sellers. Actual industries, however, contain sellers of unequal size, often vastly unequal size. The concept of concentration is intended to arrive at a single number that takes account not only of the number of firms but also of how sales are distributed among those firms.

By way of example, table 3.1 lists the number of companies that brewed beer in the United States over 1947 to 1998, as well as the number of plants. If one simply counts the number of firms, the industry looks very competitive. Even after the exit of a large number of firms (many of them were acquired by or merged with other brewing companies), the number of firms has always exceeded 25. Also reported, however, is the sum of the market shares of the five largest firms, known as the five-firm concentration ratio. That measure tells quite a different story. Although 89 firms were active in 1998, the largest five firms controlled 87 percent of the market. By itself, Anheuser-Busch’s market share was almost 47 percent.

Perhaps, then, the industry is not as competitive as we originally thought. But just because an industry is concentrated does not mean it is not competitive. If existing firms set price too high, new firms can come in with a lower price and take away much of their demand—or can they? This brings us to two other elements of structure, entry conditions and product differentiation.

Entry conditions describe the ease with which a new firm can enter an industry. Ease depends on the cost of entry, but also on the extent to which incumbent firms have an advantage, not because their product is better or their cost is lower but because they were there first. If entry is difficult, then a high price by existing firms may not be driven down by the arrival of new firms. An important related concept is that of an entry barrier, which, for the present discussion, can be thought of as something that makes entry more costly or more
difficult. The significance of entry barriers is that they may permit existing firms to charge prices above the competitive level without attracting entry. A clear example is a patent on a product that has no close substitutes. The patent holder on a drug for which there are no available substitutes can charge a monopoly price for the legal life of the patent (or at least until some other firm develops a better drug). Strong brand loyalties created through intensive advertising have been cited as an entry barrier to new firms. As we will find in chapter 6, the concept of entry barriers is controversial, but it persistently arises in many antitrust cases.

In relation to entry conditions for the beer industry, we have recently witnessed entry by craft breweries that make small quantities of high-quality beer and sell it at high prices. Given the extent of their profitable entry, it does not appear difficult to enter that segment of the market. Entry into the broader market is another matter. Much of the attrition of brewing companies from the 1940s through the 1970s is thought to have resulted from the economies scale realized by the large companies, which could produce more beer at lower cost. The construction of a modern brewery with a capacity comparable to that used by the largest firm is expensive, on the order of $250 million. There is also the large marketing cost that must be incurred to inform consumers of a new beer. These capital cost requirements can make entry into the broadest segment of the beer market difficult. It is noteworthy that all of the top five firms entered many decades ago.

---


---

### Table 3.1
U.S. Brewing Companies, 1947–1998

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of Companies</th>
<th>Number of Plants</th>
<th>Five-Firm Concentration Ratio (Percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1947</td>
<td>404</td>
<td>465</td>
<td>19.0</td>
</tr>
<tr>
<td>1954</td>
<td>262</td>
<td>310</td>
<td>24.9</td>
</tr>
<tr>
<td>1958</td>
<td>211</td>
<td>252</td>
<td>28.5</td>
</tr>
<tr>
<td>1963–64</td>
<td>150</td>
<td>211</td>
<td>39.0</td>
</tr>
<tr>
<td>1967–68</td>
<td>124</td>
<td>153</td>
<td>47.6</td>
</tr>
<tr>
<td>1974</td>
<td>57</td>
<td>107</td>
<td>64.0</td>
</tr>
<tr>
<td>1978</td>
<td>44</td>
<td>196</td>
<td>74.3</td>
</tr>
<tr>
<td>1983–84</td>
<td>35</td>
<td>73</td>
<td>83.9</td>
</tr>
<tr>
<td>1986–87</td>
<td>29</td>
<td>67</td>
<td>88.1</td>
</tr>
<tr>
<td>1989–90</td>
<td>29</td>
<td>61</td>
<td>88.6</td>
</tr>
<tr>
<td>1992</td>
<td>29</td>
<td>58</td>
<td>87.4</td>
</tr>
<tr>
<td>1995</td>
<td>42</td>
<td>77</td>
<td>89.9</td>
</tr>
<tr>
<td>1997</td>
<td>58</td>
<td>89</td>
<td>88.1</td>
</tr>
<tr>
<td>1998</td>
<td>89</td>
<td>115</td>
<td>87.2</td>
</tr>
</tbody>
</table>

Another source of market power is product differentiation. If a firm has a unique product, consumers may be willing to buy it even if the price well exceeds the prices of competitors. To give a flavor of product differentiation, let us return to Homer Simpson’s favorite beverage. Two identified dimensions of beer are its bitterness and how malty it is. Figure 3.2 depicts how various classes of products fall in this product space. Different consumers weigh these dimensions differently; it is not just price that explains why some buy Sam Adams and others buy Coors. Of course, there are other product dimensions as well, including calories, carbohydrates, and the more ephemeral properties that marketers try to influence.

An important part of structure is not only the existing degree to which products are differentiated but also the possibilities for further differentiation. It is not difficult for firms to distinguish automobiles, breakfast cereals, clothes, and the like (though making a product that is highly desired by many consumers is a challenge). In contrast, products like sugar,

Figure 3.2
Product Differentiation in the Beer Market
vitamins, and natural gas are by nature rather homogeneous. The technological opportunities to engage in product differentiation are then also relevant.

Conduct

Conduct refers to the decisions made by firms as regards price, quantity, advertising, research and development, capacity, and other variables of importance. In that product differentiation is partly influenced by firm choice, it appears as an element of conduct as well as structure. As will become clear in chapter 5, economists think of two general states of conduct: competition and collusion. Collusion refers to forms of coordination among firms; specifically, firms are able to coordinate in jointly raising price, which serves to deliver higher profit, though at a cost to consumers and social welfare. Within collusion, there are two types. Explicit collusion, which occurs when firms form a cartel, entails overt communication among firms. With tacit collusion (also called conscious parallelism), firms are able to achieve some mutual understanding without such communication. Even when firms are not colluding, industries can differ tremendously in the extent to which price exceeds cost. This is due to the elements of structure, especially the number of firms. Structure also partially influences whether firms are colluding or competing.

Industries differ not only in the intensity of competition but also in the instruments of competition. Do firms largely compete on price or product design or service or some other variable? For example, historically, the tobacco industry was described as lacking competition in price. Firms charged the same price for cigarettes, regardless of whether one brand was much more popular than other ones, and price changes were done in unison. Though price competition largely appeared absent, competition was intense on two other dimensions—advertising and brands. The tobacco companies were regularly introducing new brands and heavily advertising existing ones. Figure 3.3 shows the considerable advertising and promotional expense incurred by the cigarette industry. (Because the promotional expense includes coupons, there is at least some price competition.) The tobacco companies stopped advertising on television and radio in the late 1960s because government regulation prohibited it.

Performance

The performance block contains two elements, efficiency and technical progress. Efficiency concerns the allocation of resources with a given state of technology. For example, the monopolist that sets price above marginal cost causes a loss in economic surplus. Of course, one could list other desirable attributes of economic performance. For example, most would agree that industrial performance should facilitate full employment of labor and other resources. It can also be argued that the operations of industries should produce an equitable distribution of income. While these additional elements of performance are important, they are heavily influenced by various macroeconomic policies, such as tax policy, and
only marginally by antitrust. Hence in this book we shall focus on efficiency and put equity aside.

An example of performance measures is provided in chapter 17 for the airline industry. Figure 17.4 reports the change in air fare since deregulation, depending on the distance of the route. To make this a more meaningful measure, one would want to use a demand curve for airline services to calculate how much higher (lower) consumer surplus is from lower (higher) airfares. Indeed, this is reported in table 17.9. But there can be other measures of performance, such as product quality. An example is airline safety, which is reported in figure 17.7.

*Technical progress* is the term used in the economics literature for what might better be called *dynamic efficiency*. It is the efficiency with which an industry develops new and better production methods and products. Returning to the airline industry, table 17.10 reports the average annual percentage change in unit cost, which is a reflection of investment decisions by the airlines and advances in technology.

One of the more striking examples of technical progress is the performance of microprocessors in the last several decades. As shown in figure 3.4, a microprocessor could perform hundreds of thousands of instructions per second in the early 1970s, but the Pentium chip in the mid-1990s could perform hundreds of millions.
Measuring performance is essential to assessing whether markets are working well and, when there is some government intervention, whether it is enhancing or harming social welfare. We will then devote the next chapter to defining and measuring efficiency.

**Government**

Figure 3.1 also shows a government policy block that contains the two major categories of policy to be examined in this book: antitrust and regulation. The arrows show that antitrust and regulation can be viewed as influencing the structure and conduct of an industry in order to improve the industry’s economic performance.

An antitrust decision might lead to the dissolution of a monopoly into a number of independent sellers. This would directly affect the concentration, or industry structure. A 1911 antitrust case resulted in the creation of 33 companies by splitting up John D. Rockefeller’s
famous monopoly (or trust) of the oil-refining industry. Alternatively, antitrust laws against price fixing influence the conduct block (rather than the structure block).

The dashed arrow indicates a “feedback” relationship from the conduct block to government policy. Business firms often maintain public affairs departments or lobbyists whose purpose is to try to change government policy to favor the firm. The Robinson-Patman Act of 1936 is generally viewed as an economically harmful law enacted by Congress under strong political pressure from hundreds of small businessmen. The act was designed to protect these small businesses from the operations of large discount chains that emerged during the 1930s. Among other things, the act made it illegal for large food chains to pay lower prices than smaller independent food stores for their produce (even though the large chains performed their own brokerage function). How economic agents influence the type of regulation in place will be examined in chapter 10.

Antitrust

[Nobel laureate] Ronald [Coase] said he had gotten tired of antitrust because when the prices went up the judges said it was monopoly, when the prices went down they said it was predatory pricing, and when they stayed the same they said it was tacit collusion. 2

While Professor Coase was presumably speaking in satiric hyperbole, there is an ounce of truth in what he says. The challenge of antitrust is to distinguish anticompetitive behavior, such as collusion and monopolization, from competition and the exercise of fairly attained monopoly power (for example, due to better products).

Federal Antitrust Laws

The major federal antitrust statute in the United States, the Sherman Act of 1890, was the political reaction to the widespread growth of large-scale business combinations, or trusts, formed in the 1880s. Severe business depression had brought about pricing practices that were disastrous to firms in certain industries. To avoid this cutthroat competition, trusts were formed in many industries, including petroleum, meatpacking, sugar, lead, coal, tobacco, and gunpowder. Farmers’ organizations, labor unions, and small businessmen united in urging passage of a law to protect themselves from the economic power of these new trusts.

There are two main sections of the Sherman Act. Section 1 prohibits contracts, combinations, and conspiracies in restraint of trade. Penalties for violators can be imprisonment and/or a fine. Section 2 prohibits monopolization, attempts to monopolize, and combinations or conspiracies to monopolize “any part of the trade or commerce among the several states, or with

foreign nations.” Penalties are similar to those for Section 1. The classic target under Section 1 is price-fixing arrangements, while Section 2 is applied to market dominance. We shall examine price fixing in chapter 5 and monopolization in chapters 8 and 9.

As a result of dissatisfaction with the Sherman Act during the first few decades, two additional statutes were enacted in 1914. The Clayton Act was designed to define anticompetitive acts more clearly. It outlawed price discrimination, various vertical restraints such as tying clauses and exclusive dealing agreements, interlocking directorates, and mergers between competitors. However, these practices were illegal only where they would “substantially lessen competition or tend to create a monopoly.” Section 7, which dealt with mergers, was largely ineffective because of a legal loophole. This problem was remedied by the Celler-Kefauver Act of 1950, which amended Section 7. The law concerning mergers and vertical restraints will be discussed in detail in chapters 7 and 8. Also, Section 2, having to do with price discrimination, was heavily amended in 1936 by the Robinson-Patman Act. This will be briefly covered in chapter 9.

The second statute passed in 1914 was the Federal Trade Commission (FTC) Act. The objective of this legislation was to create a special agency that could perform both investigatory and adjudicative functions. Prior to this time, the Antitrust Division of the Department of Justice (DOJ) was the sole enforcement agency in antitrust matters. The FTC Act also contained a section that outlawed “unfair methods of competition.”

These three laws—the Sherman Act of 1890 and the Clayton and FTC acts of 1914—together form the substantive framework for U.S. antitrust policy. (Key sections of these three statutes are reproduced in the appendix at the end of this chapter.) As already indicated in our brief description, the language of the acts is general, and interpretation has been left to the courts. Hence, to really understand what is legal and what is illegal in specific situations, one must be familiar with the important court decisions and the specific rules of law that have been developed in these decisions. In many situations there remains considerable uncertainty about what a future court might hold to be legal or illegal. This is true, for example, with regard to the term monopolization. If Microsoft has 90 percent of the general-purpose computer market, is it guilty of monopolization? As we shall see, the answer depends on the nature of the tactics Microsoft followed in winning and maintaining its large market share.

Economists generally view antitrust as a set of laws designed to promote competition and, therefore, economic efficiency. The basic idea is, of course, that certain types of business behavior can lead to an inefficient allocation of resources. At first glance, this view seems to be consistent with the language of the Sherman and Clayton acts. However, it should be observed that while economic analysis can and has influenced the development of antitrust doctrine, there are other important influences as well. One such influence is the political factor of protecting the small businessman. For example, because competition might lead to the bankruptcy of small, high-cost firms, in certain areas of antitrust the law has been interpreted
to protect small businesses, even if higher costs result. To illustrate, one important Supreme Court decision contained the following statement:

It is competition, not competitors, that the Act protects. But we cannot fail to recognize Congress’ desire to promote competition through the protection of viable, small, locally owned businesses. Congress appreciated that occasional higher costs and prices might result from the maintenance of fragmented industries and markets. It resolved these competing considerations in favor of decentralization. We must give effect to that decision.  

The viewpoint taken in this book is that economic efficiency should be the only objective in antitrust decisions and that antitrust policy should be exclusively concerned with protecting competition, not competitors. This position is consistent with the conclusion of the antitrust scholar Robert Bork:

Whether one looks at the texts of the antitrust statutes, the legislative intent behind them, or the requirements of proper judicial behavior, therefore, the case is overwhelming for judicial adherence to the single goal of consumer welfare in the interpretation of the antitrust laws. Only that goal is consistent with congressional intent, and equally important, only that goal permits courts to behave responsibly and to achieve the virtues appropriate to law. 

**Enforcement and Remedies**

As noted earlier, federal government enforcement is shared by the DOJ and the FTC. The states also have their own antitrust laws, enforced by the attorneys general of the individual states. In this book, we focus on the federal antitrust laws because they are far more important.

Antitrust laws are also enforced by private actions. For example, consumers or businesses that believe they have been harmed by price fixing or some other possible antitrust violations can bring a private antitrust suit. In fact, private suits have been the predominant form of antitrust enforcement for over fifty years. Table 3.2 shows the number of antitrust cases filed by the DOJ, the FTC, and private parties in U.S. District Courts from 1975 to 1997. As the table shows, private cases accounted for around 90 percent of the total.

Table 3.2 also indicates that the number of private cases reached a peak of 1,611 cases in 1977 and has been declining since. This decline might be due to new antitrust interpretations of the late 1970s that have increased the burden on plaintiffs to prove their cases. This is especially the case in areas such as vertical restraints and predatory pricing (to be discussed in chapters 8 and 9).

---


Table 3.2  
Antitrust Cases Filed in U.S. District Courts by Type of Case, 1975–2003

<table>
<thead>
<tr>
<th>Year</th>
<th>Total</th>
<th>U.S. Government Cases</th>
<th>Private Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Number</td>
<td>Percent</td>
</tr>
<tr>
<td>1975</td>
<td>1,467</td>
<td>92</td>
<td>6.3</td>
</tr>
<tr>
<td>1976</td>
<td>1,574</td>
<td>70</td>
<td>4.4</td>
</tr>
<tr>
<td>1977</td>
<td>1,689</td>
<td>78</td>
<td>4.6</td>
</tr>
<tr>
<td>1978</td>
<td>1,507</td>
<td>72</td>
<td>4.8</td>
</tr>
<tr>
<td>1979</td>
<td>1,312</td>
<td>78</td>
<td>5.9</td>
</tr>
<tr>
<td>1980</td>
<td>1,535</td>
<td>78</td>
<td>5.1</td>
</tr>
<tr>
<td>1981</td>
<td>1,434</td>
<td>142</td>
<td>9.9</td>
</tr>
<tr>
<td>1982</td>
<td>1,148</td>
<td>111</td>
<td>9.7</td>
</tr>
<tr>
<td>1983</td>
<td>1,287</td>
<td>95</td>
<td>7.4</td>
</tr>
<tr>
<td>1984</td>
<td>1,201</td>
<td>101</td>
<td>8.4</td>
</tr>
<tr>
<td>1985</td>
<td>1,142</td>
<td>90</td>
<td>7.9</td>
</tr>
<tr>
<td>1986</td>
<td>922</td>
<td>84</td>
<td>9.1</td>
</tr>
<tr>
<td>1987</td>
<td>858</td>
<td>100</td>
<td>11.6</td>
</tr>
<tr>
<td>1988</td>
<td>752</td>
<td>98</td>
<td>13.0</td>
</tr>
<tr>
<td>1989</td>
<td>738</td>
<td>99</td>
<td>13.4</td>
</tr>
<tr>
<td>1990</td>
<td>542</td>
<td>90</td>
<td>16.6</td>
</tr>
<tr>
<td>1991</td>
<td>743</td>
<td>93</td>
<td>12.5</td>
</tr>
<tr>
<td>1992</td>
<td>566</td>
<td>85</td>
<td>15.0</td>
</tr>
<tr>
<td>1993</td>
<td>724</td>
<td>86</td>
<td>11.9</td>
</tr>
<tr>
<td>1994</td>
<td>729</td>
<td>71</td>
<td>9.7</td>
</tr>
<tr>
<td>1995</td>
<td>819</td>
<td>75</td>
<td>9.2</td>
</tr>
<tr>
<td>1996</td>
<td>720</td>
<td>73</td>
<td>10.1</td>
</tr>
<tr>
<td>1997</td>
<td>632</td>
<td>62</td>
<td>9.8</td>
</tr>
<tr>
<td>1998</td>
<td>605</td>
<td>57</td>
<td>9.4</td>
</tr>
<tr>
<td>1999</td>
<td>684</td>
<td>76</td>
<td>11.1</td>
</tr>
<tr>
<td>2000</td>
<td>901</td>
<td>90</td>
<td>10.0</td>
</tr>
<tr>
<td>2001</td>
<td>751</td>
<td>44</td>
<td>5.9</td>
</tr>
<tr>
<td>2002</td>
<td>850</td>
<td>44</td>
<td>5.2</td>
</tr>
<tr>
<td>2003</td>
<td>773</td>
<td>44</td>
<td>5.7</td>
</tr>
</tbody>
</table>

The types of cases brought by the federal government also differ from those brought by private parties. The government cases tend to be of larger import. The most lengthy and costly monopolization cases are government cases. About two-thirds of the DOJ’s cases involve horizontal price fixing. The next most frequent cases have to do with monopolization and then mergers. The FTC concentrates on price fixing and mergers. Price fixing is also the most common private case. In contrast to government cases, private parties bring relatively fewer monopolization and merger cases. Private cases more often involve practices such as tying, exclusive dealing, dealer termination, and price discrimination.

The outcomes of antitrust cases are varied. By far the most common outcome is some form of settlement rather than a full-blown trial. Almost 90 percent of private cases are either settled or voluntarily dropped by the plaintiff. Settlements often take the form of agreements by the defendants to pay damages in order to avoid a trial. The damages are usually less than those claimed by the plaintiff, but the uncertainty of the outcome of a trial can make it in both parties’ interests to agree on some settlement amount and avoid a trial.

Government cases most frequently end by consent decrees, or orders. These are agreements between the government and the defendant that specify certain actions that the defendant will take. For example, in 1982 AT&T agreed to divest its telephone operating companies in return for the DOJ’s agreement to terminate its monopolization case.

Merger cases are often settled by the firms agreeing to “spin off” products or divisions in which there is an overlap in return for an agreement by the government not to prosecute the case. For example, in 1996 the DOJ required the Bank of Boston and BayBanks, both based in Boston, to divest twenty branch offices in Boston before merging.

In cases that proceed through the trial phase, the defendant may, of course, be found innocent or guilty. Various studies indicate that the plaintiff probably prevails less than one-third of the time. There are various penalties possible in cases where the defendant is found guilty. In monopolization or mergers, a guilty defendant may be forced to divest certain assets. For example, in a merger case the defendant would likely be forced to sell the acquired firm. Another remedy is an injunction. An injunction is a court order to prohibit an antitrust violator from some specified future conduct. For example, a firm may be prohibited from only leasing (and not also selling) its copying machines, or only selling film and development services as a package. In its recent 2001 settlement with the DOJ, Microsoft is now prohibited from using certain types of contracts with various types of companies, including computer manufacturers and Internet service providers.


6. Another possibility is a plea of nolo contendere in criminal proceedings. This means “I do not wish to contend,” and is not quite as strong as a guilty plea. It is advantageous to defendants because it cannot be used as proof of guilt in subsequent cases that might be brought by parties seeking damages.
Fines or prison sentences may be used in criminal cases brought under the Sherman Act. These are usually reserved for price-fixing cases, such as a famous one that occurred in the electrical equipment industry in the early 1960s. In that case, the judge sent seven defendants to jail for thirty days and fined the firms several million dollars. Historically, however, fines have been a very weak deterrent. In the 1960s the average fine per price-fixing conviction was $131,000. This represented about two-tenths of 1 percent of the sales involved in the conspiracy.

Beginning in the late 1990s, however, fines have drastically increased, owing to revisions of the federal sentencing guidelines and the corporate leniency program. The latter program offers amnesty to the first corporate coconspirator in a price-fixing case who confesses. One 1998 case began with a complaint to the DOJ by a steelmaker customer of firms that were fixing prices of graphite electrodes. Then, one of the conspirators, Carbide/Graphite Group, offered to tell all in return for amnesty. Next, Showa Denko Carbon agreed to pay a fine of $29 million. This was followed by a then record $110 million criminal fine that UCAR International agreed to pay.

In private cases successful plaintiffs can win treble damages, which can be a particularly strong remedy. For example, if firms are found guilty of fixing prices, damages can be measured as the excess payments made by customers over what the prices would have been in the absence of the conspiracy. While such damages are seldom easy to measure and are subject to further court litigation, the final amount is multiplied by three to determine the actual award. This trebling of actual damages can lead to very high awards. In the electrical equipment case referred to earlier, the total treble damages awarded to damaged customers in subsequent civil cases were approximately $400 million!

The trebling of damages is itself a controversial issue. On the one hand, it clearly stimulates the initiation of private antitrust enforcement (as compared to awards equal to the actual damages only). This should not be viewed as an unqualified virtue, however, in that it can lead to perverse results. Some argue that a customer might knowingly encourage antitrust violations with the intention of bringing a suit to recover three times the damages.

Another criticism of treble damages is the stimulation of so-called nuisance or extortion suits that have little merit in attacking likely antitrust violations. Rather, they are brought because they appear to be good investments. For example, assume that A can make up a story of plausible damage by B resulting from some possible antitrust violation. After trebling, the


8. In the 1950s the average fine was only $40,000, which represented less than one-tenth of the involved sales. These figures are from Richard A. Posner, *Antitrust Law* (Chicago: University of Chicago Press, 1976).

estimated award could be, say, $10 million. Given that the uncertainty of how the judge or jury might decide is often high in antitrust cases, it might pay B to offer A a settlement of, say, $500,000, even if B believes the chances of winning are relatively good. (Even if B thinks the probability of winning is 0.9, its expected damage payment is still $1 million with a trial. That is, in a statistical sense, the probability of losing, 0.1, times the $10 million is $1 million.)

Exemptions from Antitrust

Congress has granted certain industries and business activities exemptions from antitrust. These include labor unions, export cartels, agricultural cooperatives, regulated industries, and some joint research and development ventures.

Labor unions were exempted from antitrust in the Clayton Act itself. The reasoning for the exemption was to permit labor to match the bargaining power of employers. There are some limits to the exemption, however.

The Webb-Pomerene Act of 1918 exempted export associations. Hence, firms can combine in an association to fix prices on their foreign sales and to allocate markets. These practices would clearly violate the Sherman Act if done domestically.

The Capper-Volstead Act of 1922 authorized agricultural cooperatives of farmers, ranchers, and dairymen to market their commodities collectively. The rationale was to permit the cooperatives to offset the bargaining power on the demand side of the market.

The exemptions for regulated industries vary depending on the industry. The rationale is that regulation itself—such as regulation by public utility commissions of electricity prices—will serve to protect the public from antitrust practices. The insurance industry, for another example, has been investigated by Congress with regard to removing its exemption because of perceived ineffective regulation by the states.

Professional sports teams are treated somewhat more leniently under antitrust. Baseball was actually granted immunity by the Supreme Court in a 1922 decision. One reason for the lenient treatment is the view that a sports league is not simply a collection of independent firms, such as, say, the steel industry. A sports league must cooperate in various ways in order to produce its product—competitive sports contests. An illustration of the leniency allowed is the practice of drafting new players. The league does not permit its teams to bid against each other for new players graduating from college or high school. Rather, the players are allocated by the league rules to particular teams. The (controversial) rationale is that this practice is necessary to promote “competitive balance.”

Finally, some joint research and development ventures are exempt from antitrust. The rationale is that such ventures are needed to maintain the competitiveness of U.S. industry against foreign competition.
Summary and Overview of Part I

Antitrust law is intended to provide an environment for competition to thrive. This chapter has reviewed the primary laws, methods of enforcement, and exceptions to those laws. The field of industrial organization provides the tools for understanding imperfectly competitive markets and identifying when competitive processes may be thwarted. It can also be used to explore the implications of antitrust policy and to design more effective policies.

Subsequent chapters in part I will describe how economists model markets characterized by imperfect competition. This is largely the focus of chapters 5 and 6, though all of the chapters offer some coverage in this regard. This includes, for example, learning when collusion is likely to emerge, how a dominant firm can cripple competition in ways that reduce social welfare, and what features of a merger suggest that it should be prohibited.

The chapters on antitrust are organized by first stating the primary antitrust issue, providing some economic theory to understand where the market failure lies and why there is a role for government intervention, and then reviewing the development of antitrust case law, along with some of the more important recent cases. Chapter 5 focuses on collusive pricing and, in addition to the coverage just mentioned, describes recent innovations in enforcement policies. Chapter 7 introduces the topic of merger and focuses on horizontal merger—when two competitors combine to form a single firm—which is the type of merger of greatest concern to competition. We also discuss the FTC’s successful challenge of the proposed merger between Staples and Office Depot. Mergers between two firms that have a buyer-seller relationship, known as vertical mergers, are covered in chapter 8. We also review vertical restraints, which have long been a point of controversy. Such restraints include tying, exclusive dealing, resale price maintenance, and territorial restraints. Recent cases of note to be covered are Time Warner–Turner (vertical merger) and Visa–MasterCard (vertical restraint). Finally, chapter 9 covers monopolization practices such as predatory pricing and refusal to deal. Considerable attention will be given to two recent cases of importance involving Kodak and Microsoft.

Before going any further, however, chapter 4 describes how economists measure efficiency, both static and dynamic, so that we can judge where antitrust activities should be focused and what types of policies will lead to an improvement in social welfare.

Appendix: Antitrust Statutes

Sherman Act

1. Every contract, combination in the form of trust or otherwise, or conspiracy, in restraint of trade or commerce among the several States, or with foreign nations, is declared to be illegal. Every person who shall make any contract or engage in any combination or conspiracy hereby declared to be illegal shall be deemed guilty of a felony, and, on conviction thereof, shall be punished by fine not exceeding one million dollars if a corporation, or, if any
other person, one hundred thousand dollars or by imprisonment not exceeding three years, or by both said punishments, in the discretion of the court.

2. 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, shall be deemed guilty of a felony, and, on conviction thereof, shall be punished by fine not exceeding one million dollars if a corporation, or, if any other person, one hundred thousand dollars or by imprisonment not exceeding three years, or by both said punishments, in the discretion of the court.

Clayton Act

2. a. It shall be unlawful for any person engaged in commerce, in the course of such commerce, either directly or indirectly, to discriminate in price between different purchasers of commodities of like grade and quality, where either or any of the purchases involved in such discrimination are in commerce, where such commodities are sold for use, consumption, or resale within the United States or any Territory thereof or the District of Columbia or any insular possession or other place under the jurisdiction of the United States, and where the effect of such discrimination may be substantially to lessen competition or tend to create a monopoly in any line of commerce, or to injure, destroy, or prevent competition with any person who either grants or knowingly receives the benefit of such discrimination, or with customers of either of them: Provided, That nothing herein contained shall prevent differentials which make only due allowance for differences in the cost of manufacture, sale, or delivery resulting from the differing methods or quantities in which such commodities are to such purchasers sold or delivered.

b. Upon proof being made, at any hearing on a complaint under this section, that there has been discrimination in price or services or facilities furnished, the burden of rebutting the prima facie case thus made by showing justification shall be upon the person charged with a violation of this section, and unless justification shall be affirmatively shown, the Commission is authorized to issue an order terminating the discrimination: Provided, however, That nothing herein contained shall prevent a seller rebutting the prima facie case thus made by showing that his lower price or the furnishing of services or facilities to any purchaser or purchasers was made in good faith to meet an equally low price of a competitor, or the services or facilities furnished by a competitor.

c. It shall be unlawful for any person engaged in commerce, in the course of such commerce, to pay or grant, or to receive or accept, anything of value as a commission, brokerage, or other compensation, or any allowance or discount in lieu thereof, except for services rendered in connection with the sale or purchase of goods, wares, or merchandise, either to the other party to such transaction or to an agent, representative, or other intermediary therein where such intermediary is acting in fact for or in behalf, or is subject to the direct or indirect control, of any party to such transaction other than the person by whom such compensation is so granted or paid.

d. It shall be unlawful for any person engaged in commerce to pay or contract for the payment of anything of value to or for the benefit of a customer of such person in the course of such commerce as compensation or in consideration for any services or facilities furnished by or through such customer in connection with the processing, handling, sale, or offering for sale of any products or commodities manufactured, sold, or offered for sale by such person, unless such payment or consideration is available on proportionally equal terms to all customers competing in the distribution of such products or commodities.

e. It shall be unlawful for any person to discriminate in favor of one purchaser against another purchaser or purchasers of a commodity bought for resale, with or without processing, by contracting to furnish or furnishing, or by contributing to the furnishing of, any services or facilities connected with the processing, handling, sale, or offering for sale of such commodity so purchased upon terms not accorded to all purchasers on proportionally equal terms.

f. It shall be unlawful for any person engaged in commerce, in the course of such commerce, knowingly to induce or receive a discrimination in price which is prohibited by this section.

3. It shall be unlawful for any person engaged in commerce, in the course of such commerce, to lease or make a sale or contract for sale of goods, wares, merchandise, machinery, supplies, or other commodities, whether patented or unpatented, for use, consumption, or resale within the United States or any Territory thereof or the District of Columbia or any insular possession or other place under the jurisdiction of the United States, or fix a price charged therefor, or discount from, or rebate upon, such price, on the condition, agreement, or understanding that the lessee or purchaser thereof shall not use or deal in the goods, wares, merchandise, machinery, supplies, or other commodities of a competitor or competitors of the lessor or seller, where the effect of such lease, sale, or contract for
sale or such condition, agreement, or understanding may be to substantially lessen competition or tend to create a monopoly in any line of commerce.

7. No corporation engaged in commerce shall acquire, directly or indirectly, the whole or any part of the stock or other share capital and no corporation subject to the jurisdiction of the Federal Trade Commission shall acquire the whole or any part of the assets of another corporation engaged also in commerce, where in any line of commerce in any section of the country, the effect of such acquisition may be substantially to lessen competition, or to tend to create a monopoly. This section shall not apply to corporations purchasing such stock solely for investment and not using the same by voting or otherwise to bring about, or in attempting to bring about, the substantial lessening of competition. Nor shall anything contained in this section prevent a corporation engaged in commerce from causing the formation of subsidiary corporations for the actual carrying on of their immediate lawful business, or the natural and legitimate branches or extensions thereof, or from owning and holding all or a part of the stock of such subsidiary corporations, when the effect of such formation is not to substantially lessen competition.

Federal Trade Commission Act

5. a. (1) Unfair methods of competition in or affecting commerce, and unfair or deceptive acts or practices in or affecting commerce, are declared unlawful.

(2) The Commission is empowered and directed to prevent persons, partnerships, or corporations, except banks, common carriers subject to the Acts to regulate commerce, air carriers and foreign air carriers subject to the Federal Aviation Act of 1958, and persons, partnerships, or corporations insofar as they are subject to the Packers and Stockyards Act, 1921, as amended, except as provided in section 406 (b) of said Act, from using unfair methods of competition in or affecting commerce and unfair or deceptive acts or practices in or affecting commerce.
4 Efficiency and Technical Progress

As indicated in the preceding chapter, *economic performance* is the term used to measure how well industries accomplish their economic tasks in society’s interests. Clearly, to evaluate antitrust laws it is essential to have some well-defined objective. In order to evaluate a law that prohibits mergers between two rivals, it is important to have a conceptual tool that identifies the costs and benefits to society of that law.

The two dimensions of economic performance to be discussed here were referred to in the last chapter as *efficiency* and *technical progress*, but could also go by the labels of static efficiency and dynamic efficiency. The main distinction is that in discussing (static) efficiency it will be assumed that the technology is given, and in discussing technical progress the assumption is that resources are being allocated to developing new technologies (for producing old products more cheaply and for producing completely new products).

### Economic Efficiency

We begin by considering the theoretical world of perfect competition. Every microeconomics text devotes much attention to the perfectly competitive model. The key assumptions are these:

1. Consumers are perfectly informed about all goods, all of which are private goods.
2. Producers have production functions that rule out increasing returns to scale and technological change.
3. Consumers maximize their preferences given budget constraints; producers maximize profits given their production functions.
4. All agents are price takers, and externalities among agents are ruled out.
5. A competitive equilibrium, that is, a set of prices such that all markets clear, is then determined.

An important welfare theorem that follows from the preceding assumptions is that the competitive equilibrium is *Pareto optimal*. In short, the equilibrium cannot be replaced by another one that would increase the welfare of some consumers without harming others. An important property of the equilibrium is that *price equals marginal cost* in all markets.

Note that the ideal competitive world that we have described would have no need for government intervention in the marketplace, except for policies affecting income distribution. This book ignores problems of income distribution, leaving those problems to the field of public finance (which studies taxation and transfer payments).

Many of the listed assumptions will be relaxed and discussed in detail throughout this book. Of course, the key assumption to be discussed in this part of the book is the *price-taking*
assumption. That is, antitrust economics is concerned with the causes and consequences of firms’ abilities to set price above marginal cost.

Once we begin to relax these assumptions, it becomes clear that we need to develop partial equilibrium tools. That is to say, it becomes incredibly complex to deal with a general equilibrium model in which some markets are monopolies, externalities exist, imperfect information about product quality obtains, and so on. Hence we now turn to welfare economics concepts in the context of a single market, effectively ignoring the interactions with all other markets.

Partial Equilibrium Welfare Tools

The competitive model described by the list was said to satisfy the condition of Pareto optimality. This is also referred to as Pareto efficiency or simply economic efficiency. One tool for evaluating the effect of a policy change (say, breaking up a monopoly) is the Pareto criterion. That is, if everyone is made better off by the change (or no one is made worse off, and at least one person is made better off), then the Pareto criterion would say that the change is “good.” It is hard to argue with this criterion for evaluating public policies. The problem is that one is unlikely to find many “good” real-world policies. In most cases in the real world, at least some people will be harmed.

A generally accepted alternative standard in applied microeconomics is the compensation principle, which is equivalent to choosing policies that yield the highest total economic surplus. The basic idea is that if the “winners” from any policy change can, in principle, compensate the “losers” so that everyone is better off, then it is a “good” change. Note that actual compensation of the losers is not required. If it were required, of course, it would satisfy the Pareto criterion.

To illustrate, consider figure 4.1. The figure shows the market demand and supply curves for DVD players. Recall first a few facts about these two curves. The competitive industry’s supply curve is found by horizontal aggregation of the supply curves of individual firms. The individual firms’ supply curves are their marginal cost curves; hence we can think of the supply curve in figure 4.1 as the industry’s marginal cost curve.

Another useful point is to recognize that the area under the marginal cost curve represents the sum of the incremental costs for all units of output and, as a result, equals the total cost. Hence the total cost of producing $Q^*$ DVD players is the area $0Q^*DC$ (this is exclusive of any fixed costs).

81 Efficiency and Technical Progress

Under certain assumptions, the demand curve can be viewed as a schedule of the marginal \textit{willingness-to-pay} of customers.\footnote{This interpretation is most easily understood if the demand curve is assumed to be made up of many heterogeneous consumers with demands for at most one DVD player. Hence the individual with the highest valuation (or willingness-to-pay) for a DVD player is represented by the vertical intercept of the demand curve, 0A. The next highest valuation (for the second DVD player) is slightly less than 0A, and so forth. The person who actually has the marginal willingness-to-pay \( P^* \) is the person who obtains a zero (individual) consumer surplus; all others have positive surpluses. For example, the person with marginal willingness-to-pay of \( Q'F \) has to pay \( P^* \) and has a surplus of \( FG \). The key assumption necessary to make this interpretation generally valid is that the income effect for the good is “small.” See Robert D. Willig, “Consumer’s Surplus without Apology,” \textit{American Economic Review} 66 (September 1976): 589–607, for support for this interpretation.} For example, at the competitive equilibrium (price \( P^* \), output \( Q^* \)), the marginal willingness-to-pay \( P^* \) exactly equals marginal cost at the output \( Q^* \). Because the area under this schedule of marginal willingness-to-pay is total willingness-to-pay, consumers are willing to pay \( 0Q^*DA \) for output \( Q^* \). The difference between total

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4_1.png}
\caption{Demand and Supply Curves in Determination of Economic Surplus}
\end{figure}
willingness-to-pay and total cost is therefore the area $ACD$ and is referred to as the \textit{total surplus} generated in the DVD market. Finally, it is common to divide total surplus into \textit{consumer surplus} of $AP*D$ and \textit{producer surplus} of $P*CD$.

Consumer surplus is defined as the total willingness-to-pay $0Q*DA$ less what the consumers must actually pay. Because consumers must pay the rectangle defined by price $P*$ and the output $Q*$ (that is, area $0Q*DP*$), the area $AP*D$ in figure 4.1 is the consumer surplus. Producer surplus, defined in an analogous manner, is equal to the profit of the firms in the industry. Because firms receive revenues of price $P*$ times output $Q*$ (that is, area $0Q*DP*$) and they incur costs equal to the area under the marginal cost curve, $0Q*DC$, they earn a producer surplus of the difference, $P*CD$.

Notice that maximizing total surplus is equivalent to maximizing the sum of consumer and producer surplus. We next show that maximizing total surplus is equivalent to selecting the output level at which price equals marginal cost. In figure 4.1, assume that output $Q'$ is being produced and sold at price $Q'F$. Clearly, at the output $Q'$, the marginal willingness-to-pay $Q'F$ exceeds the marginal cost $Q'H$. Hence, a small increase in output of $\Delta Q$ would increase surplus by the area of the slender shaded region (approximately $FH$ height by $\Delta Q$ width). Output increases would continue to increase surplus up to output $Q^*$. Hence, maximizing surplus implies that output should be increased from $Q'$ to $Q^*$, adding an increment to total surplus of area $FHD$. Of course, by an analogous argument, we can show that output increases beyond $Q^*$ would reduce surplus, since marginal cost exceeds marginal willingness-to-pay. In short, equating price and marginal cost at output $Q^*$ maximizes total surplus.

It is useful to provide another interpretation for the area $FHD$ in figure 4.1. Recall that this area represents potential increases in total surplus if for some reason output is held at $Q'$. For illustrative purposes, assume that a cartel has agreed to restrict output to $Q'$, charging price $Q'F$. This results in a so-called \textit{deadweight loss} of surplus equal to area $FHD$. This is often referred to as the \textit{social cost of monopoly}. In other words, without the cartel, competition would cause price to equal marginal cost, yielding the higher total surplus of $ACD$ as compared to the surplus under the cartel case of $ACHF$. As before, it is sometimes said that there is a deadweight loss in consumer surplus of the triangle $FGD$ and a deadweight loss of producer surplus of the triangle $GHD$.

Now, consider the point made earlier about the compensation principle and the argument that if the winners can compensate the losers, the policy change is a good one. Using a simple monopoly-versus-competition example, we will show that additional insights can be obtained by considering consumers and producers separately.

**Monopoly-versus-Competition Example**

In figure 4.2 we show a monopoly equilibrium with price $P_m$ and quantity $Q_m$. For simplicity, we assume that average cost $AC$ is constant and therefore equal to marginal cost $MC$. Hence the monopolist chooses output $Q_m$, where marginal revenue $MR$ equals marginal cost.
Profit, or producer surplus, equals price minus average cost multiplied by quantity, or area $P_mC_B$. Consumer surplus equals the triangle $AP_mB$.

Next, consider a policy to break up the monopoly and replace it with a competitive industry. Let us assume no change in costs, so that the competitive industry supply is the horizontal line at the level of $MC$. (This assumption may not be satisfied in practice, inasmuch as one reason for the existence of a monopoly may be some technological superiority that achieves lower costs of production.) Hence the new equilibrium is price $P_c$ and output $Q_c$. Consumer surplus increases to the triangular area $AP_cD$, and producer surplus disappears.

In effect, the elimination of monopoly has led to a net gain in total surplus of triangle $BCD$. This triangle, the deadweight loss caused by the monopoly, is labeled $DWL$ in figure 4.2.

To reinforce the points we have made, we can use specific numerical demand and cost functions. In particular, assume
\[ Q = 100 - P \quad \text{Demand.} \]

\[ MC = AC = 20 \quad \text{Marginal and average cost} \]

The monopoly price is therefore \( P_m = 60 \), \( Q_m = 40 \), and the competitive equilibrium is \( P_c = 20 \), \( Q_c = 80 \).

**Monopoly**

Total surplus \( = AP_cCB = 2,400 \)

Consumer surplus \( = AP_mB = 800 \)

Producer surplus \( = P_mP_cCB = 1,600 \)

**Competition**

Total surplus \( = AP_cD = 3,200 \)

Consumer surplus \( = AP_cD = 3,200 \)

Producer surplus \( = 0 \)

The pro-competition policy leads to an increase in total surplus from \$2,400 to \$3,200. On this basis, it should be carried out. Notice, however, at the disaggregated level, producer surplus falls from \$1,600 to zero. The owners of the monopoly are therefore harmed. Consumers gain enough to compensate the monopoly owners and still be better off. That is, consumers gain by \$3,200 \( - \$800 = \$2,400 \). In principle, consumers could compensate the monopoly owners with \$1,600 to offset their loss, and still have a net gain of \$2,400 \( - \$1,600 = \$800 \). Of course, as discussed earlier, under the compensation principle the compensation need not be carried out. One can justify this outcome by noting that if the government is worried about the income level of the monopoly owners, it can handle this concern directly through the tax system.

**Oil Industry Application**

An interesting application of this type of analysis was performed on the oil industry in 1979. Kenneth Arrow and Joseph Kalt evaluated the benefits and costs of removing oil price controls in the United States. While the controls will be examined in detail in chapter 18, it is instructive to present their main findings here to illustrate efficiency losses and gains as compared with simple transfers of surplus from one group to another.

---

3. The monopolist sets marginal revenue \( MR \) equal to \( MC \). \( MR = 100 - 2Q \) and \( MC = 20 \). Equating and solving for \( Q \) gives \( Q = 40 \). The competitive equilibrium is found by setting \( P = MC \). So \( 100 - Q = 20 \) gives \( Q = 80 \). In each case, substitute the equilibrium value of \( Q \) into the demand function to obtain the value of \( P \).

In the 1970s the federal government, concerned with inflation, held oil prices in the United States below what prices would have been in the absence of the controls. This resulted in efficiency losses of approximately $2.5 billion per year. (A detailed analysis of these losses is provided in chapter 18.)

Our preceding analysis, shared by most economists, is that this is as far as economists can legitimately go in evaluating public policies. It then becomes a political decision as to whether the transfers among groups are viewed as supporting or offsetting the efficiency analysis. For example, in the hypothetical monopoly example, the transfer of surplus is from the monopoly owners to consumers, and this is presumably in the politically “correct” direction. That is, if one believes that consumers generally have lower incomes than monopoly owners, and that a more equal income distribution is good, breaking up the monopoly both eliminates efficiency losses and has politically correct distribution effects.

Arrow and Kalt took a further step by trying to evaluate the distribution effect of decontrolling oil prices. Roughly, the decontrol of oil prices would mean higher prices for consumers and higher profits for producers—a politically bad transfer. They were concerned with trying to compare the gain in efficiency with the loss in equity.

The transfer from consumers to producers was estimated to be about $2.8 billion. Arrow and Kalt then proposed, with numerous qualifications, that a dollar transfer from consumers to producers would lose about half its value. The resulting “equity cost,” as they termed it, would then be half of the $2.8 billion transfer, or $1.4 billion. Hence the efficiency gain of $2.5 billion\(^{5}\) exceeded the equity cost of $1.4 billion, and they therefore recommended that oil price decontrol was in the public interest.

The key to Arrow and Kalt’s analysis is their willingness to assign an “equity cost” of 50 cents per dollar transferred from consumers to producers. As noted earlier, the standard view of economists is that assigning an equity cost of this sort is arbitrary. Economic analysts currently have no empirical basis for assigning any specific value to these equity costs. Nevertheless, it is certainly true that the political process gives great weight to equity issues, and it is helpful for economists to at least set out the magnitude involved.

Some Complications

Economies of scale were implicitly assumed to be relatively small in the monopoly-versus-competition example. That is, we ignored the problem that arises when the representative firm’s long-run average cost curve reaches its minimum at an output level that is large relative to the market demand. In other words, in our monopoly example, we assumed that the single firm could be replaced with a large number of firms, with no effect on costs.

---

5. Actually, Arrow and Kalt noted that the $2.5 billion efficiency gain from decontrol should be reduced to $1.9 billion to reflect the fact that the efficiency gains would accrue primarily to producers. Thus the final comparison was a $1.9 billion gain and a $1.4 billion loss in favor of decontrol.
To take an extreme case, consider figure 4.3. Economies of scale are such that the long-run average cost curve $LRAC$ reaches its minimum at an output level that is large relative to market demand. Situations of this kind are referred to as natural monopolies, to reflect that production can be most cheaply carried out by a single firm. The profit-maximizing monopolist would set price equal to $P_m$ and output $Q_m$.

Suppose that it were known that in order to have a sufficient number of firms in the industry for competition to obtain, each firm would be able to produce an output of only $q$. As figure 4.3 shows, the average cost of output $q$ would be quite high and would result in a price of $P_c$, which exceeds the monopoly price.

Clearly, economies of scale can make monopoly the preferred market organization. Public utilities to provide electric power or sewage treatment are notable examples. In extreme cases of the type depicted in figure 4.3, the policy problem becomes one of regulating the natural monopolist. The approach usually followed in public utility regulation is to force the mono-
polist to price so as to earn a “fair” rate of return on its investment. An alternative, although not often followed in the United States, is to create a public enterprise, owned and operated by the government. These topics will be discussed in detail in part II.

More relevant to antitrust policy is the intermediate case, where economies of scale are more moderate relative to market demand. For example, it may be imagined that the size of the automobile market is only large enough to support three or four firms, each producing at the minimum point on its long-run average cost curve. This situation would give rise to an industry of three or four firms, or an oligopoly. The key factor differentiating oligopoly from perfect competition and monopoly is that the small number of firms creates a high degree of interdependence. Each firm must consider how its rivals will respond to its own decisions.

Oligopoly theory does not yield any definite predictions analogous to the \[ \text{price} = \text{marginal cost} \] prediction of perfect competition, or the \[ \text{price greater than marginal cost} \] prediction of monopoly. Nevertheless, most theories of oligopoly imply that price will exceed marginal cost, but by less than under monopoly.

Yet oligopoly is quantitatively very significant in most industrial economies, and it is therefore an important topic for study. It should be stressed, in addition, that the prevalence of oligopoly does not necessarily imply that large-scale economies are the cause. In fact, whether or not economies of scale explain the existence of particular oligopolies is a key public policy concern. We will return to oligopoly theory in chapter 5.

A second complication is the existence of product differentiation. Product differentiation refers to the situation in which some differences in the products of rival sellers are perceived by the buyers. The differences may be real differences, such as the differences in size, styling, horsepower, reliability, and so on, between Fords and Chevrolets, or they may be primarily the result of image differences conveyed through advertising. The main requirement is that consumers regard the differentiation sufficiently important that they willingly pay a somewhat higher price for their preferred brand.

E. H. Chamberlin\(^6\) constructed the theory of monopolistic competition in which many competitors produce differentiated products. All firms that produce products that are reasonably close substitutes are members of the product group. Given these assumptions and the assumption of free entry, the long-run equilibrium of a monopolistic competitor is given by the tangency of the firm’s demand curve with its average cost curve. This is shown in figure 4.4.

The monopolistic competitor earns zero profits in long-run equilibrium. This is a consequence of the assumption of free entry; the existence of a positive profit will attract entry until a firm’s own demand is reduced sufficiently to make profits zero. The product differentiation assumption gives the firm’s demand curve its slightly negative slope; that is, the firm can increase its price without losing all its sales to a competitor.

---

The relevant point here is that price exceeds marginal cost—the signal that there is a misallocation of resources. But consider Chamberlin’s argument:

The fact that equilibrium of the firm when products are heterogeneous normally takes place under conditions of falling average costs of production has generally been regarded as a departure from ideal conditions. . . . However, if heterogeneity is part of the welfare ideal, there is no prima facie case for doing anything at all. It is true that the same total resources may be made to yield more units of product by being concentrated on fewer firms. . . . But unless it can be shown that the loss of satisfaction from a more standardized product is less than the gain through producing more units, there is no “waste” at all, even though every firm is producing to the left of its minimum point.7

The key issue is the optimal amount of product variety, and this is a difficult theoretical problem. A large literature on this subject has developed since Chamberlin’s observation.8 In chapter 6 we present a simple model that illustrates the trade-offs involved.

X-Inefficiency

Other types of inefficiency may be important in monopoly. First, we consider X-inefficiency, so named by Leibenstein in his well-known 1956 article on the subject.9 Thus far, we have

---

assumed that both monopolists and perfect competitors combine their factors of production efficiently, thereby minimizing cost for each level of output. However, it can be argued that the pressures of competition force perfect competitors to be cost minimizers, whereas the freedom from competition makes it possible for the monopolist to be inefficient, or X-inefficient. That is, the monopolist may operate at a point above its theoretical cost curve.

Of course, X-inefficiency is inconsistent with the assumption that monopolists maximize profits. However, some economists have argued that the separation of ownership from control in large firms with market power permits the managers to substitute their own objectives for the profit objectives of the owners. Therefore, in such cases, X-inefficiency may arise.

**Monopoly-Induced Waste**

A third and final source of inefficiency created by monopoly is competition among agents to become a monopolist. Consider the example of a government-mandated monopoly in the form of a franchise. If figure 4.2 depicts the relevant demand and cost curves, then the franchise owner will earn profits equal to \( P_m P_c CB \). Knowing that the firm that receives this franchise will earn rents of \( P_m P_c CB \). firms will invest resources in lobbying the legislature or the regulatory agency in order to become the recipient of this franchise. This competition to earn monopoly profits uses up real resources in the form of labor by lobbyists and lawyers. These wasted resources represent a cost to society, just as do the traditional deadweight loss and any X-inefficiencies. Competition among firms for rents is appropriately referred to as rent-seeking behavior.10

How large is the welfare loss from rent-seeking behavior? We know that it cannot exceed the amount of monopoly profits (\( P_m P_c CB \) in figure 4.2). No firm would find it optimal to spend in excess of that amount in order to become a monopolist. In some simple models it has been shown that if rent-seeking is perfectly competitive (that is, there are many identical firms), then all rents will be competed away.11 In that case, the total welfare loss from monopoly is \( P_m P_c DB \) More generally, \( P_m P_c DB \) represents an upper bound on the welfare loss from monopoly (excluding any X-inefficiencies), while \( BCD \) is a lower bound.

There are a number of ways in which rent-seeking behavior may arise. As just mentioned, competition for rents could take the form of firms lobbying legislators in order to get favorable legislation passed, for example, entry regulation and import quotas. When these lobbying activities use up real resources, they represent a welfare loss associated with monopoly. Alternatively, if favorable government actions are achieved by bribing legislators or

---


regulators, then this is not a welfare loss but rather simply a transfer from the briber to the bribee. However, one could take the rent-seeking argument one step further and argue that agents will compete to become legislators or regulators in order to receive the rents from bribes. If real resources are used at that stage, then they represent a welfare loss.

Rent-seeking behavior can also arise in the form of excessive nonprice competition. Suppose firms are able to collude so that price exceeds cost. The lure of this high price-cost margin could generate intensive advertising competition as firms compete for market share. Depending on the particular setting, this advertising may have little social value and simply be the by-product of competition for rents. Historically, socially wasteful advertising has been thought to be a feature of the cigarette industry. As we will see in later chapters, nonprice rivalry among firms in a cartel or in a regulated industry can lead to excessive spending on product quality, product variety, and capacity, as well as on advertising.

Finally, unions have been found to be quite effective in extracting some of a firm’s profits in the form of higher wages. This higher wage results in the private marginal cost of labor exceeding its social marginal cost, so that a firm tends to use too little labor in the production process. This inefficient input mix represents yet another source of welfare loss associated with monopoly. Back when unions were more powerful than they are now, one study found that unions extracted in excess of 70 percent of monopoly rents.12

Estimates of the Welfare Loss from Monopoly

Having identified various sources of welfare losses due to price exceeding marginal cost, it is natural to wonder about the quantitative size of these losses in the U.S. economy. One method for estimating the traditional deadweight welfare loss (which we will denote $DWL$) is as follows. From figure 4.2, we know that $DWL$ equals $BCD$ when the monopoly price is charged. $BCD$ can be approximated by $\frac{1}{2}(P_m - P_c)(Q_c - Q_m)$, where this approximation is exact if the demand function happens to be linear. More generally, if $P^*$ is the price that firms charge and $Q^*$ is the resulting level of demand, then $DWL$ is approximated by $\frac{1}{2}(P^* - P_c)(Q_c - Q^*)$. Because $P^*$ and $Q^*$ are the actual price and quantity, one can collect data on $P^*$ and $Q^*$ for various firms or industries. However, we typically do not know the competitive price without estimating marginal cost. It is difficult to get a reliable estimate of marginal cost for just a single industry. To do so for a significant portion of the U.S. economy would be a gargantuan task. We then need to find some alternative way of estimating $DWL$ that does not require having data on $P_c$ and $Q_c$.

In his pioneering study, Arnold Harberger used the following approach. To begin, one can perform a few algebraic manipulations and show that

\[
\frac{1}{2}(P^* - P_c)(Q_c - Q^*) = \frac{1}{2} \eta d^2 P^* Q^*
\]  

(4.1)

where \( \eta \) is the absolute value of the market demand elasticity and \( d \) is the price-cost margin. More formally, \( d = (P^* - P_c)/P^* \) and \( \eta = |(\Delta Q/Q)/(\Delta P/P)| \) where \( \Delta Q = Q_c - Q^* \) and \( \Delta P = P^* - P_c \). Although data on industry revenue, \( P^* Q^* \), are available, one needs to come up with estimates of \( d \) and \( \eta \). In order to derive a ballpark figure of \( DWL \), Harberger used the difference between an industry’s rate of return and the average for the sample to estimate the price-cost margin \( d \), and simply assumed that \( \eta = 1 \). With this back-of-the-envelope technique, Harberger found that \( DWL \) was on the order of one-tenth of 1 percent of GNP. Though the assumption of unit elasticity is arbitrary, what is important is that the conclusion one draws from this estimate is robust to the value of \( \eta \). Even increasing it fivefold will mean that \( DWL \) is only one-half of 1 percent of GNP. Harberger concluded that the welfare losses from monopoly are very small indeed.

We thought it worthwhile to review Harberger’s work in order to show how one might go about estimating welfare losses from monopoly. However, there are several reasons to question the relevance and accuracy of his low estimate of \( DWL \). First, it is an estimate based on data from the 1920s. Whether such an estimate is relevant to today’s economy is questionable. Second, we know that there are sources of welfare loss from monopoly other than \( DWL \). Harberger estimated that the size of above-normal profits was around 3–4 percent of GNP. This leaves open the question of how much resources were used in competing for these rents. Depending on the extent of such competition, we know that the true welfare loss could be as high as 3–4 percent of GNP. The third and perhaps most important reason for questioning the validity of Harberger’s estimate is that later researchers have performed more careful analyses and found significantly higher values of \( DWL \).

One such study was performed by Keith Cowling and Dennis Mueller. They took a quite different approach to estimating \( DWL \). Their approach avoided having to make an arbitrary assumption on the demand elasticity by instead assuming that firms maximize profit. The first step in their analysis is to note that a firm’s profit-maximizing price \( P^* \) satisfies the following relationship:

\[
\frac{P^* - MC}{P^*} = \frac{1}{\eta}
\]  

(4.2)

where MC is marginal cost. In words, a firm sets price so that the price-cost margin equals the inverse of the (absolute value of the) firm demand elasticity. Note that in a competitive industry $\eta$ is infinity, so that (4.2) tells us that $P^* = MC$. Recall that Harberger showed that $\text{DWL}$ could be estimated by $\frac{1}{2} \eta d^2 P^* Q^*$ where $d = (P^* - MC)/P^*$ (and we have replaced $P_e$ with $MC$). From (4.2), it follows that $1/\eta = d$. Now substitute $d$ for $1/\eta$ in the expression that estimates $\text{DWL}$ (see equation 4.1):

$$\text{DWL} = \frac{1}{2} \eta d^2 P^* Q^* = \frac{1}{2} \left( \frac{1}{d} \right) d^2 P^* Q^* = \frac{1}{2} dP^* Q^*. \quad (4.3)$$

Substituting $(P^* - MC)/P^*$ for $d$ in equation 4.3, it follows that

$$\text{DWL} = \frac{1}{2} \left( \frac{P^* - MC}{P^*} \right) P^* Q^* = \frac{1}{2} (P^* - MC) Q^* = \frac{1}{2} \Pi^* \quad (4.4)$$

where $\Pi^*$ is firm profits. Because $\Pi^* = (P^* - AC) Q^*$, where $AC$ is average cost, the last equality in (4.4) uses the assumption that marginal cost is constant, so that $MC = AC$. Cowling and Mueller have then shown that the deadweight welfare loss created by a firm is approximately equal to half of its profits.

With this methodology, Cowling and Mueller collected data on $\Pi^*$ for 734 U.S. firms for 1963–1966. Remember that $\Pi^*$ represents economic profits, not accounting profits. Hence they used 12 percent as the normal return on capital in the economy and subtracted normal profits from accounting profits to estimate $\Pi^*$. Their estimate of $\text{DWL}$ was around 4 percent of GNP, considerably higher than that found by Harberger. If one includes advertising expenditures as wasted resources associated with rent-seeking behavior, their measure jumps to 13 percent of GNP. Of course, inclusion of all advertising expenditures assumes that all advertising lacks any social value. This assumption is clearly false, because some advertising reduces search costs for consumers. Thus one would expect Cowling and Mueller’s best measure of the welfare loss from monopoly to lie somewhere between 4 and 13 percent of GNP. It is interesting that under their most comprehensive measure, General Motors by itself created a welfare loss of one-fourth of 1 percent of GNP!

It is clearly important to understand the quantitative size of the welfare loss from price exceeding marginal cost, whether it is due to monopoly, collusion, or regulation. Unfortunately, estimating welfare losses is an inherently precarious task because of data limitations. One must then interpret these estimates with considerable caution. A final point is that even if we knew for certain that monopoly welfare losses were, say, only 1 percent of GNP, this would not be grounds for abolishing antitrust. The reason is that the 1 percent figure would apply to an economy with antitrust in place. Perhaps if antitrust did not exist, the monopoly losses would be much larger.
Efficiency in producing the desired bundle of known goods and services with a given technology is obviously important. Some argue, however, that economists place too much emphasis on this type of efficiency. They believe it is at least as important for industry to be efficient in generating new knowledge that saves resources in producing known products, as well as in creating new or higher-quality products. In short, industry should be technically progressive.

Importance of Technological Change


In figure 4.5, two production functions are shown. The functions apply to the economy as a whole and show that output per worker-hour, $Q$, rises (at a decreasing rate) with the amount of capital per worker-hour, $K$. The lower production function represents the best technology known at time $t = 1$. New knowledge at time $t = 2$ leads to a shift upward in the function, enabling society to obtain higher $Q$ for any given $K$. Thus the shift represents technological change between $t = 1$ and $t = 2$.

We can now indicate Solow’s method of analysis. Suppose that at $t = 1$ the amount of capital per worker-hour is $K_1$ and at $t = 2$ it is $K_2$. Furthermore, suppose that $Q_1$ and $Q_2$ are the observed outputs per worker-hour on these two dates. The total increase in $Q$ can be conceived as consisting of two parts: the movement from $A$ to $B$ (the effect of technological change) and the movement along the production function from $B$ to $C$ (the effect of increased capital per worker-hour). As stated earlier, Solow found that the amount of the total increase in $Q$ due to technological change (the movement from $A$ to $B$) was greater than that due to increased capital per worker-hour (the movement from $B$ to $C$).

The importance of new products is also clear. One has only to think of some examples: jet aircraft, DVDs, antibiotics, personal computers, nuclear power, and so forth. This dimension of technological change was not incorporated fully in Solow’s estimates.
Granted that technological change is important, we must now consider what determines it. At the industry level, it is reasonable to expect a number of factors to be influential in determining the rate of technical advance. Undoubtedly, the amount of resources devoted to research and development (R & D) is important. But the amount of private resources allocated will depend on profitability considerations, which in turn will depend on such things as the expected demand for the product and the technical feasibility of the project. And, what is particularly relevant in this book, the structure of the market should affect these profitability calculations, as well as government policy.

Some quite persuasive economists have argued that some monopoly power is necessary to provide incentives for firms to undertake research and development programs. The rationale for existing patent policy rests to some extent on this argument. Others, however, have taken the opposite position, namely, that it is competitive pressures that produce the higher rates of progressiveness.
The economist Joseph Schumpeter is usually credited with the view that some monopoly must be tolerated to obtain progressiveness:

But in capitalist reality as distinguished from its textbook picture, it is not [perfect] 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 strikes not at the margins of the profits and the outputs of the existing firms but at their very foundations and their very lives.17

Before turning to a rivalry model that provides some insight into these issues, it may be helpful to explain several terms that will be used in our discussions. At the beginning there is basic research, which seeks knowledge for its own sake. Most industrial firms engage in applied research, which is directed toward a particular product or process. If successful, invention takes place, which is the discovery of new knowledge. After invention, development must take place, leading to the commercial application of the invention, or innovation. The last phase of technical change is the diffusion of the product or process throughout the industry, or economy.

A Model of R & D Rivalry

F. M. Scherer and D. Ross have presented an instructive model of R & D rivalry in their book, Industrial Market Structure and Economic Performance. Their model is useful in illuminating the conflicting incentives that market structure provides for innovation: (1) more rivals tend to stimulate more rapid innovation in order to be first with a new product and benefit from the disproportionate rewards of being first, and (2) more rivals split the potential benefits into more parts, making each firm’s share less. Here we shall draw heavily on their expositional approach, which in turn is an attempt to simplify more mathematically complex models published elsewhere.

The model collapses innovative activity into a determination of the speed of new product development. That is, the model seeks to show what factors lead to the firm’s choice of the number of years from beginning R & D to the market introduction of the product. We should note that it is incorrect to equate a shorter time necessarily with “socially preferred.” While we often seem to identify higher rates of innovation as necessarily “good,” it is of course possible for innovation to take place too rapidly.18

The situation is one of oligopoly with each firm competing through improved products. To improve one’s product requires carrying out R & D for a certain time period prior to marketing. The time period can be compressed by expending more resources. Hence there is a cost-time trade-off that is shown in figure 4.6 as the curve CC’.

It is easy to explain the curve $CC'$ by example. Let one plan be to spend $400,000 per year for ten years. The present discounted value of this stream at 10 percent is $2.5 million. Hence this value is one point on $CC'$. Another plan is to spend $1 million per year for five years—with a present value of $3.8 million. This is a second point on $CC'$. Clearly the implication is that it costs more to shorten the time to innovation. There are several reasons for this. First, costly errors can be made when development steps are taken concurrently instead of waiting for the information early experiments supply. Second, parallel experimental approaches may be necessary to hedge against uncertainty. Third, there are diminishing returns in the application of additional scientific and engineering manpower to a given technical project.\(^{19}\)

---

It is assumed that firms choose the time to innovation $T$ in order to maximize the present discounted value of their profits. Hence the next step is to introduce the function $V$, which represents how the present value of net revenues varies with $T$. The net revenues are equal to revenues from the sale of the product minus the production and marketing costs incurred. As shown in figure 4.6, the $V$ functions (each $V$ function corresponds to a different number of rivals) slope down to the right. It is easy to explain the slope of $V_1$, which refers to a monopoly situation with no rivals.

Assume for simplicity that the net revenues from the product will be constant over time. Now, if somehow $T$ happened to be zero, the vertical intercept of $V_1$ would equal the present value of this constant stream of net revenues from $T = 0$ forever. If the flow is $1$ million per year, then the present value at 10 percent would be $10$ million. Now as $T$ increases, the early years of potential net revenues are lost, thereby reducing the present value and causing the $V_1$ function to slope down to the right. For example, if net revenues do not begin until year 3, the present value falls to $8.3$ million.

In this monopoly case, the profit-maximizing $T$ is easily found graphically. It is simply that value of $T$ that is associated with the largest vertical difference between the present value of net revenues and the present value of R & D costs. This is also found by locating the value of $T$ where the slope of $V_1$ equals the slope of $CC'$. The optimal $T$ is shown as $T_1$ in the figure.

Now consider a second situation in which there are, say, three rivals. This is represented with the function $V_3$. Two points should be noted about $V_3$ relative to $V_1$. It is lower, reflecting lower net revenues for each $T$, and it is steeper. Thus, $V_3$ is lower than $V_1$ simply because the total market potential net revenues must now be split three ways. That is, it is reasonable for a firm with two rivals to expect to share the market with the other two, to some degree. Notice that this shift downward reduces overall expected profits, but it does not eliminate them because $V_3$ still lies above $CC'$. This reduced expected appropriability of net revenues by the firm can lead to a situation in which the innovation is simply unprofitable—with a zero rate of innovation. Such a case is shown by the function $V_5$, which corresponds to five rivals. Presumably five rivals is “too many” and would result in too much imitation for R & D to be undertaken at all.

Return to the $V_3$ case and consider the second point made in the preceding paragraph. We see that $V_3$ is steeper than $V_1$. First note what this steeper slope implies about the optimal $T$. As the slope gets steeper, the optimal $T$ falls until the $V_3$ function’s slope equals that of $CC'$, at $T_3$. This steepness, in other words, leads to a faster speed of development as compared to the monopoly case. This effect of increasing the number of rivals is therefore a stimulating effect on the rate of innovation—as long as the number of rivals does not increase too much and cause a situation where innovation is completely unprofitable.

What causes the slope of $V_3$ to be steeper than $V_1$ can be explained as follows. The idea is that the proportionate payoff to being first, and enjoying the whole market until imitation,
Chapter 4

grows with the number of rivals. In monopoly, there is little loss as one innovates later and later—the monopoly still has the whole market in later years. This means the slope of $V_1$ is relatively flat. Now in a three-firm market, the first firm enjoys the whole market until imitation occurs. Let us say that when imitation occurs, the leader’s share falls to one-third—equal to the share of each of the two imitators. The relative size of the leader’s payoff to one of the two imitators’ payoffs is what determines the slope of $V_3$. Clearly, the relative payoff for a low $T$ (and being first) is greater than the case of monopoly. Furthermore, in some cases the pioneer firm is even relatively better off because of brand loyalty developed during the early years. This makes it possible to keep a proportionately greater share of the market than its imitators. For example, brand loyalty may make it possible for the pioneer to keep half the market, with each imitator getting one-fourth.

Hence the model that we have described points clearly to the influence of market structure on innovation. Though the complexity of the innovative process makes it difficult to obtain nice, neat results, one can infer that neither pole of perfect competition nor pure monopoly seems to be ideal. As Scherer and Ross put it in summarizing an extensive review of empirical work:

> What is needed for rapid technical progress is a subtle blend of competition and monopoly, with more emphasis in general on the former than the latter, and with the role of monopolistic elements diminishing when rich technological opportunities exist.\(^{20}\)

A more fundamental issue is that it may be naive to conceive of the public policy issue as one of choosing the optimal market structure to optimize the trade-off between static allocative efficiency and progressiveness. The reason is that structure itself should perhaps be viewed as evolving endogenously as technological change occurs through time. Thus, firms that are successful in the innovation game will grow while others decline or drop out. And, over time, the industry’s concentration will change as a result.

In chapter 24 we consider a special policy toward technological change—the granting of patents to provide incentives for inventive activity. Although the model of R & D rivalry implicitly assumed patents to be unimportant, chapter 24 goes to the other extreme and assumes that patents are essential. Most empirical studies conclude that the importance of patents varies greatly across industries, being especially important in pharmaceuticals and chemicals.

**Summary**

This chapter has examined two dimensions of economic performance: efficiency and technical progress. The major difference is that the efficiency section assumed a known technol-
ogy while the technical progress discussion focused on the allocation of resources to develop new knowledge (for producing new products, and for producing existing products more cheaply).

An important lesson that this chapter tries to teach is the usefulness of total economic surplus in assessing public policies. That is, if total economic surplus rises as a result of a policy change, then, under certain plausible assumptions, one can argue that the change is in the public interest. An example of such a change that was described was the decontrol of oil prices in the United States.

A hypothetical monopoly-versus-competition example was used to explain the concept of the deadweight loss caused by monopoly pricing. A short section discussed several empirical studies that have sought to estimate the social cost of monopoly in the United States.

In the technical progress section, a simple model of R & D rivalry was presented. The model illustrated how increasing the number of rivals can have two opposing effects on the speed of innovation. The key point of the model is that no simple relationship between the number of rivals and the rates of innovation exists—a larger number of rivals does not always produce better results for society.

Questions and Problems

1. Explain the difference between the Pareto criterion and the compensation principle as rules for deciding whether a particular policy change is in the public interest.

2. Assume, in the monopoly-versus-competition example in the text, where demand is \( Q = 100 - P \) and marginal cost \( MC = \text{average cost} AC = $20 \), that \( MC \) under competition remains at $20. However, assume that the reason the monopoly can continue to be a monopoly is that it pays $10 per unit of output to reimburse lobbyists for their efforts in persuading legislators to keep the monopoly insulated from competition. For example, the lobbyists may be generating (false) studies that demonstrate that competition results in higher costs.
   a. Calculate the prices and quantities under monopoly and competition.
   b. Calculate total economic surplus under monopoly and competition. The difference is the social cost of monopoly.
   c. The social cost of monopoly can be disaggregated into two distinct types of cost: the resources cost of rent seeking and the usual deadweight loss of output restriction. What are their respective magnitudes?

3. Discuss the concept of “equity cost” used in the oil industry study by Arrow and Kalt. Do you think it is generally true that “consumers” have lower incomes than “producers”? Does it matter to your answer that labor unions and senior citizens have large ownership interests in corporations through pension funds?

4. A (mini-) refrigerator monopolist, because of strong scale economies, would charge a price of $120 and sell forty-five refrigerators in Iceland. Its average cost would be $60. On the other hand, the Iceland Planning Commission has determined that five refrigerator suppliers would be
sufficiently competitive to bring price into equality with average cost. The five-firm equilibrium would yield a price of $100 and a total output of fifty refrigerators.

a. Consumer surplus under the five-firm industry organization would be larger than under monopoly. If the demand curve is linear, by how much is consumer surplus larger?

b. Producer surplus under monopoly is larger—by how much?

c. If the Planning Commission thinks that total economic surplus is the correct criterion, which organization of the refrigerator industry will they choose?

5. What is the best market structure for promoting technical progress?

6. A study in 1975 estimated the effect of monopoly on equity as opposed to efficiency (W. Comanor and R. Smiley, “Monopoly and the Distribution of Wealth,” Quarterly Journal of Economics [May 1975]). For 1962, the wealthiest 0.27 percent of the population accounted for 18.5 percent of wealth. If all industries were competitive, this study estimated that the wealthiest 0.27 percent would have only 13 percent of wealth in 1962. Can you explain this finding? Hint: The wealthiest 0.27 percent held 30 percent of business ownership claims.
In a recent decision, Supreme Court Justice Antonin Scalia wrote that collusion is the “supreme evil of antitrust.”¹ In this chapter, we will trace major judicial decisions from the passage of the Sherman Act in 1890 to the present to show the evolution of the current legal rules toward price fixing and thus how the government has strived to control this “evil.”

Before beginning this task, however, we shall discuss the theories of collusive and oligopoly pricing. Oligopoly, of course, refers to a market structure with a small number of sellers—small enough to require each seller to take into account its rivals’ current actions and likely future responses to its own actions. Price-fixing conspiracies, or cartels, are not limited to a small number of sellers, although it is generally believed that the effectiveness of a cartel is greater when the number of participants is small.

Our coverage will proceed in the following manner. In order to explore the theory of oligopoly and collusion, we will need to be properly tooled. Toward this end, an introductory discussion of game theory is provided. With that under our belts, we review the Cournot model and a model of collusive behavior. To bring theory to life, several price-fixing cartels are viewed through the lens of these models. The last section of this chapter discusses antitrust law, landmark price-fixing cases, and enforcement issues such as the U.S. Department of Justice’s successful revision of the corporate leniency program.

A very important assumption that underlies the analysis in this chapter is that potential entry is not a problem. We shall always assume that the number of active firms is fixed, so that our focus is on the internal industry problems of firms reaching an equilibrium when the only competition comes from existing firms. Allowing for competition from entrants is delayed until the next chapter.

### Game Theory

#### Example 1: Advertising Competition

Consider a duopoly in which firms do not compete in price because of collusion or regulation. Let the price be $15 and the quantity demanded be 100 units. If unit cost is $5, then profit per unit equals $10. That is, a firm receives revenue of $15 for each unit, and it costs the firm $5 to produce that unit. Though it is assumed that firms have somehow been able to avoid competing in price, it is also assumed that firms do compete via advertising. To simplify matters, a firm can advertise at a low rate (which costs $100) or at a high rate (which costs $200). Also for simplicity, assume that advertising does not affect market demand but only a firm’s market share; specifically, a firm’s market share depends on how much it advertises relative to its competitor. If both firms advertise an equal amount (whether low or high),

---

then firms equally share market demand—that is, each has demand of 50 units. However, if one firm advertises low and the other advertises high, then the high-advertising firm dominates the market with a market share of 75 percent.

Given these data, we can calculate each firm’s profits net of advertising expenditure for all of the possible advertising rates that firms can set. The result of these calculations is table 5.1, which is called a profit (or payoff) matrix. This matrix comprises four cells. In each cell there are two entries, the first of which is firm 1’s profit (net of advertising expenditure) and the second of which is firm 2’s profit (also net of advertising expenditure). If both firms advertise at a low rate, then each firm’s profit is 400. Each firm receives half of market demand, which is 50 units, and earns profit of $10 on each unit. This generates gross profit of $500. After subtracting the cost of advertising at a low rate, which is $100, we derive profit of $400. If instead firm 1 advertises at a low rate and firm 2 advertises at a high rate, then firm 1’s profit is 150 \[= (10)(100)(0.25) - 100\] and firm 2’s profit is 550 \[= (10)(10)(0.75) - 200\]. Finally, if both firms advertise at a high rate, then each receives profits of 300 \[= (10)(100)(0.5) - 200\].

Table 5.1
The Advertising Game

<table>
<thead>
<tr>
<th></th>
<th>FIRM 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low Advertising</td>
<td>High Advertising</td>
</tr>
<tr>
<td>FIRM 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low Advertising</td>
<td>400,400</td>
<td>150,550</td>
</tr>
<tr>
<td>High Advertising</td>
<td>550,150</td>
<td>300,300</td>
</tr>
</tbody>
</table>

Suppose that the two firms simultaneously choose how much to advertise. That is, each firm decides whether to spend $100 or $200 on advertising at the same time that its competitor decides. How much should each firm advertise if it wants to maximize its profits? Let’s look at this decision from the perspective of firm 1. First notice that how much profit it earns depends on how intensively firm 2 advertises. If firm 2 advertises at a low rate, then firm 1 earns $400 from choosing a low rate of advertising and $550 from choosing a high rate. The higher profit from advertising at a high rate comes from the higher market share that firm 1 would receive, which is more than sufficient to offset the higher cost of advertising. Thus, if firm 1 believes that firm 2 will not advertise very much, then it should advertise intensively. If instead firm 1 believes that firm 2 will advertise at a high rate, then it earns $150 from low advertising but $300 from high advertising. Therefore, profit-maximizing behavior by firm 1 is to advertise intensively, regardless of how much it believes firm 2 will advertise. By the symmetry of this setting, firm 2’s decision process is identical.
diction as to how profit-maximizing firms would behave in this setting is that both firms would invest heavily in advertising. This is a stable outcome, as each firm is maximizing its profits. Note, however, that joint profits are not maximized. Joint profits are $600 when both advertise intensively but are $800 if both pursue minimal advertising. The problem is that each firm has an incentive to advertise heavily, but when they both do so then each firm’s advertising is negated by the advertising of its rival.\(^2\)

**Example 2: Compatibility of Standards**

Let us now consider a quite different setting. Suppose it is 1980 and firm 1 is a supplier of videocassette recorders and firm 2 is a supplier of videocassettes. Each firm’s product can have either the Beta format or the VHS format. Further suppose that firm 1’s cost of producing a VHS VCR is slightly less than its cost of producing a Beta VCR, while firm 2’s cost of producing Beta cassettes is slightly less than producing a VHS cassette. These two firms are the sole exporters of these products to a country that currently has no VCRs or video-cassettes. Each firm must decide whether to manufacture its product with the Beta format or the VHS format. (Assume that it is too costly to manufacture both formats.) The consumers in this country are indifferent between the two technologies. All they care about is that the format of the videocassettes sold is the same as the format of the VCRs sold.

Assume that the profit matrix for these two firms is table 5.2. If both use VHS, then firm 1 earns $500 and firm 2 earns $200. If both use Beta, then firm 1 earns $400 and firm 2 earns $250. If they each supply a different format, then each earns zero profits, as demand is zero.

What should firms do in this setting? If firm 1 believes that firm 2 is going to supply Beta cassettes, then firm 1 earns higher profits by also supplying Beta cassettes (compare $400 and 0). If firm 2 believes that firm 1 will supply Beta VCRs, then firm 2 is better off supplying Beta cassettes (compare $250 and 0). Both firms using the Beta format is then a stable

\(^2\) This game is more commonly known as the Prisoner’s Dilemma and is the most widely examined game in game theory. There are literally hundreds of papers that investigate this game theoretically or test it experimentally in the laboratory using human subjects, typically college students. For a discussion of the Prisoner’s Dilemma, see Prajit Dutta, *Strategies and Games: Theory and Practice* (Cambridge, Mass.: MIT Press, 1999).
outcome, as each firm is choosing the format that maximizes its profits given the format chosen by the other firm. No firm has an incentive to change its decision.

There is another stable outcome, however, in which both firms choose the VHS format. Firm 1 prefers this stable outcome to the one in which the Beta format is chosen, while firm 2 prefers the outcome with the Beta format. We would predict that both firms would offer the same format, but which format that might be is unclear.

In contrast to Example 1, a firm’s decision in this setting depends on the decision made by the other firm. A firm’s profit-maximizing format is the one that matches the format that it believes the other firm is going to choose. This interdependence in firms’ decisions is an essential feature of oligopolies.  

The Strategic Form of a Game

The two preceding examples are known as games. We are all familiar with the term game, but in fact there is a branch of applied mathematics known as game theory. In game theory, a “game” is a well-defined object. In this section, we describe what a game is and how one analyzes it. The reason for spending time on game theory is that it is a tool designed for investigating the behavior of rational agents in settings for which each agent’s best action depends on what other agents are expected to do. As a result, game theory will prove to be very useful in investigating firm behavior in oligopolies and, more generally, in providing insight concerning the strategic behavior of firms.

Our goal is to understand the behavior of agents in a particular economic setting—for example, the decision by firms as to how much to advertise, how much to produce, or what technology to use. The first step in applying game theory is to define what the relevant game is (where “relevant” is defined by the researcher and the scientific community at large).

The strategic form (or normal form) of a game describes an economic setting by three elements: (1) a list of the agents who are making decisions, (2) a list of the possible decisions that each agent can make, and (3) a description of the way in which each agent evaluates different possible outcomes. Those agents who are making decisions are called players. The decision rule of a player is called a strategy. A strategy tells a player how to behave in the setting being modeled. A player’s strategy set includes all the possible strategies that a player could choose. Finally, a player’s payoff function describes how he evaluates different strategies. That is, given the strategies chosen by all players, a player’s payoff function tells him his state of well-being (or welfare or utility) from players having played those strategies.

In Example 2 the players are firms 1 and 2, a player’s strategy is a format, and the strategy set of a player contains the two available formats: {Beta, VHS}. When a player is a firm,
we typically assume that a player’s payoff is just its profit. In Example 2 the payoff functions of the two players are represented by table 5.2.\(^4\)

**Nash Equilibrium**

Having modeled a particular economic setting as a game, one can use it to recommend to players how they should play or to make predictions as to how they will play. While game theory is perhaps better at the former, it is more commonly used for the latter, and it is for the latter purpose that we will use it.

A strategy is a decision rule that instructs a player how to behave over the course of the game. A strategy may be very simple, like a rate of advertising, or very complicated, like what price to set at the beginning of each month in response to last month’s sales. On a conceptual basis, it is useful to think of players as choosing their strategies simultaneously at the beginning of the game. Once their strategies are chosen, the game commences and players act according to their strategies.

Assuming that players are rational, a player chooses the strategy that yields the highest payoff (if the player is a firm, this just means the firm is profit maximizing). As shown in tables 5.1 and 5.2, a player’s payoff depends not only on her strategy but also on the strategy of the other player. This relationship is an essential feature of games. Thus, in deciding which strategy is best, a player must take into account the strategies that she expects the other players to choose. To capture this interdependence, the concept of Nash equilibrium was developed. A list of strategies, one for each player, is a Nash equilibrium if each player’s strategy maximizes her payoff given the strategies chosen by the other players and if this condition holds simultaneously for all players.

In Example 1, the only Nash equilibrium is the strategy pair (High, High). It is a Nash equilibrium because given that firm 2 chooses High, firm 1 maximizes its profits (that is, its payoff) by also choosing High. In addition, given that firm 1 chooses High, High is optimal for firm 2. In other words, each firm is choosing a profit-maximizing strategy given the (profit-maximizing) strategy chosen by the other firm. At each of the other three strategy pairs, (Low, Low), (Low, High), and (High, Low), at least one firm can increase its profit by choosing a different strategy. For example, at (Low, High), firm 1 prefers to use High rather than Low, as it increases profit by 150. In Example 2, both (Beta, Beta) and (VHS, VHS) are Nash equilibria but (Beta, VHS) and (VHS, Beta) are not Nash equilibria. Note that the outcomes referred to as being stable in Examples 1 and 2 are Nash equilibria.

\(^4\) In Example 1, the set of players is firms 1 and 2, a strategy is a rate of advertising, a strategy set is \{Low, High\}, and the payoff functions of the players are represented by table 5.1. For an introduction to game theory, see Dutta, *Strategies and Games*. 
Oligopoly Theory

An oligopoly is an industry with a small number of sellers. How small is small cannot be decided in theory but only in practice. Nevertheless, in principle, the criterion is whether firms take into account their rivals’ actions in deciding on their own actions. In other words, the essence of oligopoly is recognized interdependence among firms. AT&T certainly considers the actions and likely future responses of MCI and Sprint and the other long-distance telephone service providers when it makes its decisions (whether concerning price, advertising, or other factors). However, a Kansas wheat farmer would be silly to worry about any effect of his planned output on the planned output of the farmer next door.

Since the pioneering work of Augustin Cournot in 1838, many theories of oligopoly have been developed. As it has turned out, the first model of oligopoly is still the most widely used one. In this section we will review Cournot’s modeling of the oligopoly problem. Afterward, a brief discussion of some other oligopoly models is provided.

The Cournot Solution

To simplify the analysis, consider the following numerical example. Let marginal cost be constant at $40 and assume that the inverse market demand function is

\[ P = 100 - Q \]  

(5.1)

If industry supply is \( Q \), then the price that equates supply and demand is \( 100 - Q \).

Prior to considering the oligopoly setting, let us review the monopoly case, as it will be a useful benchmark. Given (5.1), the marginal revenue curve of a monopolist is

\[ MR = 100 - 2Q \]  

(5.2)

Demand, marginal revenue, and marginal cost are depicted in figure 5.1. A monopolist maximizes its profit by setting \( Q \) so as to equate marginal revenue and marginal cost. This value for quantity is 30, and the price is $70. Monopoly profit equals $900.

Now suppose that there are instead two firms, denoted firm 1 and firm 2. Both firms have constant marginal cost of $40 and produce identical products. The distinguishing features of the Cournot model are that firms choose quantity (rather than price) and do so simultaneously. The price of the good is set in the market so as to equate industry supply (which equals


6. Because revenue is \( R = (100 - Q)Q \), marginal revenue equals the first derivative of \( (100 - Q)Q \) with respect to \( Q \): \( dR/dQ = 100 - 2Q \). Equating \( MR \) and \( MC \), \( 100 - 2Q = 40 \), and solving for \( Q \) yields the profit-maximizing quantity of 30.
the sum of the firms’ outputs) and demand. Thus, if \( q_1 \) and \( q_2 \) are the outputs of firms 1 and 2, respectively, then the resulting market price is

\[
P = 100 - q_1 - q_2.
\]  

(5.3)

Though the Cournot model was developed more than a century before game theory, one can interpret the Cournot model as a game. In the Cournot game, the set of players is comprised of firms 1 and 2. The strategy of a firm is its quantity. Finally, a firm’s payoff is simply its profits. For our numerical example, the profits of firms 1 and 2 are, respectively,

\[
\pi_1 = (100 - q_1 - q_2)q_1 - 40q_1
\]  

(5.4)

\[
\pi_2 = (100 - q_1 - q_2)q_2 - 40q_2.
\]  

(5.5)

These profits clearly demonstrate the interdependence that characterizes oligopoly. The profit of firm 1 depends not only on its own quantity but also on the quantity of firm 2 (and, similarly, \( \pi_2 \) depends on \( q_1 \) and \( q_2 \)). In particular, the higher \( q_2 \) is, the lower is \( \pi_1 \) (holding \( q_1 \)
fixed). That is, the more your competitor produces, the lower is the market price for the good, causing your revenue (and profit) to be lower.

Having formulated the Cournot model, we next proceed to determine the behavior of firms implied by profit maximization. However, it is not sufficient to determine the quantity that maximizes firm 1’s profit without also considering what quantity maximizes firm 2’s profit, as the former depends on the latter. We need to find a quantity for each firm that results in each maximizing profits given the quantity of its competitor. This means finding a Nash equilibrium.

As an initial step in deriving a Nash equilibrium, consider the problem faced by firm 1. It wants to select a quantity that maximizes $\pi_1$, taking into account the anticipated quantity of firm 2. Suppose firm 1 believes that firm 2 plans to produce 10 units. Using (5.3), the market price when firm 1 produces $q_1$ is then

$$P = 100 - q_1 - 10 = 90 - q_1.$$  (5.6)

When $q_2 = 10$, it follows that firm 1’s revenue is $(90 - q_1)q_1$ and its marginal revenue is $90 - 2q_1$. Firm 1 wants to choose $q_1$ to maximize its profits, which means equating firm marginal revenue with marginal cost. As shown in figure 5.2, the profit-maximizing quantity is 25.

One can go through the same exercise to find the profit-maximizing output for firm 1 for each possible value of $q_2$. For example, if $q_2 = 30$, then firm 1’s revenue is $(70 - q_1)q_1$ and its marginal revenue is $70 - 2q_1$. As shown in figure 5.2, the profit-maximizing output is now 15. Doing this calculation for all possible values of $q_2$, one finds that the value of $q_1$ that maximizes firm 1’s profit is

$$q_1 = 30 - 0.5q_2.$$  (5.7)

Equation 5.7 is known as firm 1’s best reply function because it gives the value of $q_1$ that is firm 1’s best (in the sense of maximizing profits) reply to firm 2’s output.\(^7\) In the same manner, one can derive the best reply function of firm 2 to be

$$q_2 = 30 - 0.5q_1.$$  (5.8)

In figure 5.3, firm 1’s best reply function is plotted. Note that it is downward sloping; the higher is firm 2’s quantity, the lower is the profit-maximizing quantity of firm 1. The intuition lies in figure 5.2. Note that when $q_2$ is raised from 10 to 30, firm 1’s demand and

\(^7\) To derive firm 1’s best reply function, find the value of $q_1$ that maximizes $\pi_1$. This is achieved where marginal profit is zero: $\partial \pi_1/\partial q_1 = 100 - 2q_1 - q_2 - 40 = 0$. Solving this expression for $q_1$ yields $q_1 = 30 - 0.5q_2$. An analogous method yields firm 2’s best reply function. What we are calling a firm’s best reply function, Cournot called a firm’s “reaction function.” In his original treatment, Cournot provided a dynamic story to his static model that entails each firm’s reacting to the other’s output. However, the term reaction function is a misnomer, as firms make simultaneous output decisions in the Cournot model so that there is no reaction.
marginal revenue curves shift in. This reflects the fact that for any value of \( q_1 \), a higher value of \( q_2 \) results in firm 1 receiving a lower price for its product. Because its demand is weaker, firm 1 produces less in response to firm 2’s producing more. Hence, firm 1’s best reply function is downward sloping; that is, firm 1 produces less, the more firm 2 is anticipated to supply.

A Nash equilibrium is defined by a pair of quantities such that each firm’s quantity maximizes its profit given the other firm’s quantity. The appeal of such a solution is that no firm has an incentive to change its output given what its competitor is doing. We have already shown that a firm maximizes its profits only when it produces according to its best reply function. A Nash equilibrium is then defined by a pair of quantities such that both firms are simultaneously on their best reply functions, which is shown in figure 5.4.

A Nash equilibrium is defined by each firm producing 20 units. That is, given that one’s rival produces 20, an output of 20 maximizes a firm’s profit. You should convince yourself that any other output pair is not a Nash equilibrium, as at least one firm is not maximizing its profit. For example, at \( q_1 = 30 \) and \( q_2 = 15 \), firm 2 is maximizing its profits (it is on its
best reply function), but firm 1 is not. Given \( q_2 = 15 \), firm 1’s profit-maximizing output is 22.5 (see figure 5.3).

To summarize, the Nash equilibrium of the Cournot game (which we also refer to as the Cournot solution, figure 5.4) is

\[
\begin{align*}
q_1 &= 20 \\
q_2 &= 20 \\
\pi_1 &= 400 \\
\pi_2 &= 400 \\
P &= 60.
\end{align*}
\]

Note that the price at the Cournot solution exceeds the competitive price of 40 (which equals unit cost) but is less than the monopoly price of 70. This is a standard feature of the Cournot solution and is not particular to our numerical example. The Cournot price exceeds marginal cost because firms do not act as price takers. Firms are not small in the sense that their quantity decisions affect price, and they realize this fact. They know that the more they produce, the lower is the market price. As a result, each firm supplies less than it would if it were a price taker, resulting in the Cournot price exceeding the competitive price.

\[\text{Figure 5.3} \]
Firm 1’s Best Reply Function

Instructor Evaluation Copy
Not for Classroom Use
The Cournot price is less than the monopoly price because each firm cares only about its own profits and not industry profits. When firm 1 considers increasing its quantity, it takes into account how this quantity increase affects \( p_1 \) but ignores how it affects \( p_2 \). As a result, in maximizing one’s own profit, each firm produces too much from the perspective of maximizing industry profit. Hence, the monopoly price (which is also the joint-profit-maximizing price under constant marginal cost) exceeds the Cournot price.

Though each firm is maximizing its own profit at the Cournot solution, both firms could simultaneously raise their profits by jointly reducing their quantity from 20 toward the joint-profit-maximizing output of 15. The problem is that neither firm has an incentive to do so. Suppose that firms were to communicate prior to choosing their quantity and agreed to each produce 15. If firm 1 actually believed that firm 2 would go along with the agreement and produce 15, firm 1 would do better by reneging and producing 22.5 (reading from its best reply function). Given \( q_2 = 15 \), \( q_1 = 22.5 \) yields profits of 506.25 for firm 1 versus profits of 450 from \( q_1 = 15 \). Of course, firm 2 is no fool and thus would never produce 15 anyway. Note the similarity with the problem faced by firms in the advertising game of Example 1.
In both games, the Nash equilibrium is Pareto inefficient in that firms could raise both of their profits by jointly acting differently.

In concluding this section, let us note that the Cournot solution predicts that the price-cost margin is inversely related to the number of firms, denoted $n$, and the absolute value of the elasticity of market demand, denoted $\eta$:

$$\frac{P - MC}{P} = \frac{1}{n\eta}.$$  \hfill (5.9)

Recall that the elasticity of demand measures how responsive demand is to a change in price. According to this formula, as the number of firms increases, the right-hand side expression in (5.9) becomes smaller, implying that the price-cost margin shrinks. The reader is referred to table 16.1 for a numerical example that establishes that the market price associated with the Cournot solution is decreasing in the number of firms. Equation 5.9 also tells us that as the industry becomes perfectly competitive (the number of firms approaches infinity), the price-cost margin converges to zero, or, in other words, price converges to the competitive price of marginal cost.\(^8\)

Other Models of Oligopoly

In the more than 150 years since Cournot developed his pioneering analysis of oligopoly, many alternative models of oligopoly have been put forth. In 1934, Heinrich von Stackelberg proposed a modification of the Cournot model based on the observation that some industries are characterized by one firm being a leader in the sense that it commits to its quantity prior to its competitors doing so.\(^9\) The Stackelberg model is a game with sequential moves in which the “leader” (say, firm 1) chooses quantity and then, after observing firm 1’s quantity, the “follower” (firm 2) chooses its quantity.

Equilibrium behavior in the Stackelberg model entails the follower, firm 2, acting the same way as in the Cournot model (though not choosing the same output). In place of firm 2’s conjecture as to what firm 1 will produce, firm 2 actually observes firm 1’s quantity. Given the observed output of firm 1, firm 2 chooses its output to maximize its profit, which is just given

\(^8\) The profit of firm $i$ is

$$\pi_i = P(q_i + \ldots + q_n)q_i - C(q_i).$$

The first-order condition is

$$\frac{\partial \pi_i}{\partial q_i} = P'(q_i + \ldots + q_n)q_i + P(q_i + \ldots + q_n) - C'(q_i) = 0.$$

Adding the first-order condition over $n$ firms gives $P'Q + nP - nMC = 0$, where $Q$ is industry supply. Dividing through by $n$ and using the formula for the absolute value of demand elasticity, $\eta = -(1/P)(P/q)$, one gets $-1/\eta + n - nMC/P = 0$. Rearranging, the expression in (5.9) emerges.

by its best reply function. In contrast, firm 1, being the leader, acts quite differently. Rather than take firm 2’s output as fixed (as it does in the Cournot model), firm 1 recognizes that firm 2 will respond to the output choice of firm 1. Taking into account the response of the follower, the leader chooses its quantity to maximize its profit.

Using the example from the preceding section, the Stackelberg leader’s problem is to choose output so as to maximize:

\[100 - q_1 - (30 - 0.5q_2)]q_1 - 40q_1.\]  

(5.10)

This expression is firm 1’s profit but where we have substituted firm 2’s best reply function, \(30 - 0.5q_1\), for its quantity, \(q_2\). This substitution reflects the fact that firm 1 influences firm 2’s quantity choice. Solving for the value of \(q_1\) that maximizes (5.10), one finds that the leader produces 30 units and the follower responds with a quantity of 15 (\(= 30 - 0.5 \cdot 30\)). Compared to the Cournot solution, firm 1 produces more and firm 2 produces less, and thus the leader ends up with a higher market share. Firm 1 produces above the Cournot quantity because it knows that firm 2 will respond by producing less (recall that firm 2’s best reply function is downward sloping). In other words, firm 1 takes advantage of moving first by committing itself to a higher quantity, knowing that it will induce its rival to produce less.

A second class of oligopoly models assumes that firms choose price rather than quantity. The first piece of work in this line is that of Joseph Bertrand. 10 In a critique of Cournot’s book, Bertrand briefly sketched a model in which firms make simultaneous price decisions. When firms offer identical goods and have constant marginal cost, there is a unique Nash equilibrium when firms choose price, and it entails both firms pricing at marginal cost. The Bertrand model yields the surprising result that oligopolistic behavior generates the competitive solution! As it turns out, this result is very special. If firms’ products are differentiated, price competition results in findings similar to those of the Cournot solution: each firm’s price lies between the competitive price and the monopoly price.

In concluding this brief sketch, let us note that our basic findings for the Cournot model with homogeneous goods are robust to allowing firms’ products to be differentiated as long as they are not too dissimilar. 11

Product Differentiation

No discussion of oligopoly theory would be complete without mentioning product differentiation. One of the most significant ways in which firms compete is by trying to make their

product unique relative to the other products in the market. The reason is that the more differentiated one’s product is, the more one is able to act like a monopolist. That is, a firm can set a higher price without inducing large numbers of consumers to switch to buying competitors’ products.

To consider the role of product differentiation, let us follow the suggestion of Bertrand and assume that firms make simultaneous price decisions with constant marginal cost—though, of course, we will assume that firms’ products are differentiated. This means that consumers perceive these products as being imperfect substitutes. That is, there are consumers who are willing to buy one firm’s product even though it is priced higher than its competitors’. It also typically means that a small change in a firm’s price causes a small change in its demand. For a market with two firms, let \( D_i(p_1, p_2) \) denote the number of units demanded of firm \( i \)’s product when the prices are \( p_1 \) and \( p_2 \) (\( i = 1, 2 \)).

An example of a firm demand curve when products are differentiated is

\[
D_1(p_1, p_2) = 100 - p_1 + 0.5p_2 \tag{5.11}
\]

\[
D_2(p_1, p_2) = 100 - p_2 + 0.5p_1. \tag{5.12}
\]

Note that a firm’s demand is decreasing in its own price but increasing in the price of its rival. The latter property reflects products being substitutes, so that when firm 2 raises its price, some consumers who previously purchased firm 2’s product decide to switch to buying firm 1’s product. Another notable property is that even if \( p_1 > p_2 \), firm 1 still has positive demand (as long as the difference between prices is not too great). Because firm 1’s product is distinct from that of firm 2, some consumers are willing to pay a premium for firm 1’s product. Finally, note that a firm’s demand is affected more by a change in its own price than by a change in the price of its rival. If we assume each firm has constant marginal cost of 20, then firms’ profit functions are

\[
\pi_1(p_1, p_2) = (p_1 - 20)D_1(p_1, p_2) = (p_1 - 20)(100 - p_1 + 0.5p_2) \tag{5.13}
\]

\[
\pi_2(p_1, p_2) = (p_2 - 20)D_2(p_1, p_2) = (p_2 - 20)(100 - p_2 + 0.5p_1) \tag{5.14}
\]

where firm \( i \) earns \( p_i - 20 \) on each unit sold.

To derive a Nash equilibrium for the differentiated products price game, one can use the same method that we used for the homogeneous goods quantity game (that is, the Cournot model). The first step is to derive each firm’s profit-maximizing price given the price of its competitor, that is, the best reply function. For the preceding profit functions, one can show that the best reply functions are

\[
p_1 = 60 + 0.25p_2 \tag{5.15}
\]
In contrast to the Cournot game, a firm’s best reply function is upward sloping (see figure 5.5). The reason is as follows. Firm 1’s demand rises in response to firm 2 charging a higher price, as some of firm 2’s consumers decide to buy from firm 1. Generally, the stronger a firm’s demand is, the higher is its profit-maximizing price. It follows that the higher firm 2’s price becomes, the higher is the price that maximizes firm 1’s profit, so that its best reply function is upward sloping.

As with the Cournot game, Nash equilibrium occurs where the best reply functions intersect, so that each firm is choosing a price that maximizes its profit, given the other firm’s price. Equilibrium then has firms pricing at 80. To convince yourself that this is true, if you plug 80 for \( p_2 \) into (5.15), you find that the resulting price for firm 1 is 80, and if you plug 80 for \( p_1 \) in (5.16), you find that the resulting price for firm 2 is 80. Given that one’s rival prices at 80, it maximizes a firm’s profit to also price at 80. Since this argument applies to both firms, both firms pricing at 80 is a Nash equilibrium.
An important result is that the more differentiated firms’ products are, the higher are equilibrium prices. To understand the underlying intuition, let us begin by considering the extreme case of homogeneous products. If firm 1 prices at the same level as firm 2, firms equally share market demand, inasmuch as consumers are indifferent about buying from the two firms. Because consumers buy from the firm with the lower price, firm 1 can induce all of firm 2’s consumers to buy from it by slightly undercutting firm 2’s price. Thus a small drop in firm 1’s price results in a doubling of its demand. This strong incentive to undercut a competitor’s price results, in equilibrium, in prices being driven down to marginal cost.

As products become differentiated, the rise in demand from undercutting a competitor’s price is reduced. Some consumers are willing to buy a competitor’s product even at a higher price (this is what it means for products to be differentiated). Because a firm does not get as much of its competitor’s demand from lowering its price, there is a reduced incentive for firms to set low prices. As a result, prices are higher, in equilibrium, when firms’ products are more differentiated. When firms’ products are so differentiated that consumers do not even perceive them as being substitutes, each firm is effectively a “local” monopolist and charges the monopoly price for its market.

Collusion

Using the Cournot solution, we found that oligopolistic competition results in firms jointly producing too much. Although each firm is individually maximizing its profits, both firms are aware that industry profits are not maximized. Going back to the example in the preceding section, the Nash equilibrium entails each firm producing 20 units and receiving profits of 400. If instead they each produced half of the monopoly output of 30, then each would receive profit of 450. Of course, no individual firm has an incentive to do so because producing 20 units is optimal, given the other firm is anticipated to supply 20 units. However, if they could find a way in which to coordinate and jointly reduce their production, they could increase profits for everyone. The lure of higher profits through coordination of their behavior is what collusion is all about.

Historically, there are many incidents of firms successfully coordinating their quantity and price decisions. Notable industries that have had cartels include tobacco, electrical equipment (including General Electric), and vitamins. Since the Cournot solution predicts that firms do not maximize joint profits, how are we to explain the fact that, in some industries, firms are able to collude? What keeps each firm from deviating from the collusive agreement by producing at a higher rate and thereby earning higher profits?
A Theory of Collusion

The inadequacy of the Cournot solution lies in the limitations of the Cournot model. A critical specification is that firms make quantity decisions only once. In reality, firms live for many periods and are continually making quantity decisions. To correct for this weakness in the Cournot model, consider an infinitely repeated version of that model. In each period, firms make simultaneous quantity decisions and expect to make quantity decisions for the indefinite future. For simplicity, assume that the demand curve and cost functions do not change over time.

There are several important differences between the standard (one-period) Cournot game and the infinitely repeated Cournot game. Because a firm chooses output more than once, a strategy is going to be a much more complicated object than simply some number of units. Rather than generally define a strategy, we will consider some examples. Second, and this is very important, each firm will receive information over time in the form of past prices and quantities. Though firms choose quantity simultaneously in a given period, each firm can observe the other firm’s past quantities as well as the resulting market price. Finally, each firm acts to maximize the sum of its discounted profits rather than just today’s profit. We shall let \( r \) denote the interest (or discount) rate for each firm.

Let \( q^1_t \) and \( q^2_t \) denote the period \( t \) quantity of firm 1 and firm 2, respectively, where \( t = 1, 2, 3, \ldots \). We first want to show that one Nash equilibrium for this game has each firm produce the Cournot quantity in every period: \( q^1_t = 20 \) and \( q^2_t = 20, t = 1, 2, 3, \ldots \). Recall that this example has each firm’s per-period profit as 400. A firm’s payoff, which is just the sum of its discounted profits, is then

\[
\frac{400}{1 + r} + \frac{400}{(1 + r)^2} + \frac{400}{(1 + r)^3} + \ldots = \frac{400}{r}.
\]  

(5.17)

For example, if the interest rate is 10 percent, then a firm’s payoff (and its market value) is $4,000.

Now consider firm 1 choosing a strategy in which its quantity differs from 20 for one or more periods. In all those periods for which \( q^1_t \neq 20 \), profits will be lower in those periods (as 20 units maximizes current profit), while profits in periods in which the firm continues to produce 20 remain the same. Hence the sum of discounted profits must be lower. It is then optimal for a firm to produce 20 in each period when it expects its competitor always to do

12. It is not necessary that a firm literally live forever but rather that it can potentially live forever. In other words, there is not any known date in the future for which firms are certain that they will no longer be around. In fact, firms can be around for quite a long time. Currently, the oldest recorded firm still in existence is the Swedish firm Stora Kopparberg (it means Great Copper Mountain). Documents show that it was in existence in 1288!

13. The present value of receiving $V$ every period equals \( V/r \), where \( r \) is the interest rate and \( V \) is received at the end of the period. For a derivation of this result, see chapter 2, note 7.
so. One Nash equilibrium for the infinitely repeated Cournot game is then just a repetition of the Cournot solution. Alternatively stated, repetition of a Nash equilibrium for the single-period game is a Nash equilibrium for the infinitely repeated game.

While a Nash equilibrium has been found for the infinitely repeated Cournot game, it does not put us any closer to understanding how firms can sustain a collusive outcome like the joint-profit maximum. Thus, let’s consider a very different strategy from the one just described. In particular, consider a strategy that allows a firm’s output to depend on how much its competitor produced in the past. An example of such a strategy (for firm 1) is as follows:

\[ q_1^t = 15 \]
\[ q_1' = \begin{cases} 15 & \text{if } q_1^\tau = 15 \text{ and } q_2^\tau = 15 \text{ for all } \tau = 1, \ldots, t-1 \\ 20 & \text{otherwise} \end{cases} \] \( t = 2, 3, \ldots \) (5.18)

This strategy says that firm 1 should produce 15 in period 1. Recall that 30 is the monopoly quantity, so that each firm producing 15 maximizes joint profit. In any future period, it should produce 15 if and only if both firms produced 15 in all periods prior to the current one. Alternatively, if one or more firms deviated from producing 15 in some past period, then firm 1 should produce 20 (the Cournot quantity) in all remaining periods. This strategy is called a “trigger strategy” because a slight deviation from the collusive output of 15 triggers a breakdown in collusion. The strategy for firm 2 is similarly defined:

\[ q_2^t = 15 \]
\[ q_2' = \begin{cases} 15 & \text{if } q_1^\tau = 15 \text{ and } q_2^\tau = 15 \text{ for all } \tau = 1, \ldots, t-1 \\ 20 & \text{otherwise} \end{cases} \] \( t = 2, 3, \ldots \) (5.19)

If both firms use these strategies, then each will produce 15 in period 1. Because each produced 15 in period 1, then, as prescribed by these strategies, each firm will produce 15 in period 2. By the same argument, the two firms will produce 15 in every period if they use these strategies. Hence the monopoly price is observed in all periods. If we can show that these strategies form a Nash equilibrium (that is, no firm has an incentive to act differently from its strategy), we will have a theory that explains how profit-maximizing firms can sustain collusion.

Given that firm 2 uses the strategy in (5.19), firm 1 receives profit of 450 in each period from using (5.18) so that its payoff is 450/r. Now consider firm 1 choosing a different strategy. Any meaningfully different strategy must entail producing a quantity different from 15 in some period. There is no loss of generality from supposing that this occurs in the first period. If \( q_1^1 \neq 15 \), then firm 2 learns after period 1 that firm 1 deviated from the collusive output. According to firm 2’s strategy, it will respond by producing 20 in all future periods.
Because firm 1 is aware of how firm 2 will respond, firm 1 will plan to produce 20 in all future periods after deviating from 15 in the current period, as doing anything else would lower firm 1’s payoff from period 2 onward. It follows that if $q_1 \neq 15$, then firm 1’s payoff is

$$\left[ (100 - q_1 - 15)q_1 - 40q_1^2 \right] \frac{1}{1+r} + \frac{400}{r(1+r)}.$$  \hspace{1cm} (5.20)

The first term is period 1 discounted profits, while the second term is the sum of discounted future profits. Given that firm 1 deviates output from the collusive level of 15 in the first period, (5.20) shows that the amount by which $q_1$ differs from 15 affects current profits but not future profits. This is because the punishment for cheating is independent of how much a firm cheats. Therefore, (5.20) is maximized by setting $q_1 = 22.5$, as that output maximizes current profits (reading off firm 1’s best reply function in figure 5.3 when $q_2 = 15$). Substituting 22.5 for $q_1$ in (5.20), one then derives the highest payoff that firm 1 can earn from choosing a strategy different from (5.18):

$$\frac{506.25}{1+r} + \frac{400}{r(1+r)}.$$  \hspace{1cm} (5.21)

Figure 5.6 shows the time path of profit from going along with the collusive agreement—earn 450 every period—and cheating on the cartel—earn 506.25 in period 1 and 400 every period afterward. Thus, cheating gives higher profits up front but lower profits in the future because it intensifies competition.

Given that firm 2 uses the strategy in (5.19), firm 1 earns $450/r$ from using the strategy in (5.18), while the highest payoff it can get from choosing a different strategy is that in (5.21). Therefore the strategy in (5.18) maximizes firm 1’s payoff if and only if

$$\frac{450}{r} \geq \frac{506.25}{1+r} + \frac{400}{r(1+r)}.$$  \hspace{1cm} (5.22)

Working through the algebra, (5.22) holds if and only if $r \leq 0.89$. In other words, if firm 1 sufficiently values future profits ($r$ is sufficiently small), it prefers to produce 15 each period rather than cheat and produce above 15. By the same argument one can show that firm 2’s strategy in (5.19) maximizes the sum of its discounted profits if and only if $r \leq 0.89$.

We conclude that (5.18) and (5.19) is a Nash equilibrium when the discount rate is sufficiently low.14

---

In contrast to the one-period game, we have shown that firms are able to produce at the joint-profit maximum without either firm having an incentive to cheat. What is critical for this result is that a firm’s quantity decision depends on the past behavior of its competitor. Note that if firm 1 ever cheats on the collusive agreement by producing different from 15 units, firm 2 will respond by giving up collusion forever and producing the Cournot output of 20. This response to a deviation lowers firm 1’s future profits (and also firm 2’s future profits). A firm that cheats raises its current profits, but the punishment of future collusion breaking down lowers future profits. As long as a firm sufficiently values future profits, it will prefer not to cheat.

Of course, if \( r > 0.89 \), then these trigger strategies do not form a Nash equilibrium, which means that they cannot sustain the joint-profit maximum. In that event, each firm prefers to cheat and receive higher profit today. However, trigger strategies with a higher collusive output should form a Nash equilibrium. For example, if \( 0.89 < r \leq 1.33 \), one can show that firms can use trigger strategies to support an output of 16 though not an output of 15. Although price is below the monopoly level, it is still above the Cournot price.

Finally, let us note that it can be shown that collusion becomes more difficult as the number of firms grows; that is, as the number of firms increases, each firm must have a lower discount rate in order for cheating to be unprofitable at the joint-profit maximum. Intuitively, as the number of firms increases, two things happen. First, each firm has a smaller share of the market at the joint-profit maximum. This provides a firm with a bigger increase in current

![Figure 5.6: Profits from Colluding versus Cheating](image)
profit from cheating on the collusive agreement. On the other hand, the punishment from cheating—the Cournot solution—is more competitive as there are more firms. Because price falls more after cheating, the loss from cheating is greater when the number of firms is higher. This works in the opposite direction to make collusion easier. It has been shown that the first effect dominates, so that the net gain to cheating goes up as the number of firms increases. Hence, collusion is more difficult the greater the number of firms. Consistent with the static setting, we conclude that having more firms results in a lower price.

Challenges to Collusion

In developing a concise and instructive theory of collusion, many important issues related to collusion were ignored. This section will discuss some of the real-world complications associated with collusion that were not taken account of in our previous discussion.

Coordination and Bargaining

The previous section described a mechanism by which firms could sustain high prices. Each firm restricts its supply—below that which maximizes current profit—under the threat that any overproduction will cause collusion to collapse, with a return to the lower profit associated with competition. As long as firms sufficiently care about future profit, this mechanism is self-enforcing in the sense that each firm finds it optimal to abide by it, given that other firms do so. What was left unsaid is how firms come to deploy such a mechanism. Indeed, there are typically many equilibria of the form just described. For example, one can show that the rule in equations 5.18–5.19 is an equilibrium not just for a collusive quantity of 15 but for any quantity between 15 and 20. If firms are currently not colluding—and thereby each is producing 20—how do they coordinate in reducing supply? And among all those collusive outcomes that they could support, how do they settle upon a particular one? The first question is a problem of coordination, while the latter throws in an element of bargaining.

In describing how firms, in practice, have solved the coordination problem, it is important to draw a distinction between two methods of collusion. *Explicit collusion* refers to when firms directly communicate about plans to coordinate their behavior in term of price, production, allocation of customers, and the like. As described later in our discussion of antitrust law and policy, the very act of communicating with such an intent is illegal, and therefore, colluding firms work hard to maintain secrecy. *Tacit collusion* is when firms coordinate their behavior without explicitly communicating. The legality of that activity is far more problematic and typically eludes prosecution. Though the ultimate outcome on welfare may be the same—price is higher whether achieved through either brand of collusion—they are treated far differently in the eyes of the law, but more on that later.

A number of coordination mechanisms to support tacit collusion have been observed. It is important to keep in mind that the task of such a mechanism is not just to coordinate on price
and production once but also to respond to cost and demand changes. One common institution is price leadership. The price leader might be the largest or lowest-cost firm, though this is not always so. The identity of the leader may even change over time. The essential idea is that the leader openly announces its intention to change its price, and the other firms normally follow with similar price changes. Of course, the leader must accurately assess what price is likely to be acceptable to its rivals. If the leader is too far off the mark, some rivals may elect not to follow.

Price leadership through public announcements of fare changes had been commonly practiced in the airline industry and, in 1994, the Department of Justice settled a case involving such practices. A particular concern was that the airlines were using a commonly owned computer network to signal each other for the purpose of sustaining noncompetitive prices. The computer system provided instantaneous transmission of more than 100,000 domestic fare changes daily. The network makes it easier for the airlines to enter the data in their computer reservation systems for travel agents.

It was conjectured that one of the pricing rules that airlines sought to establish and sustain was the rule that each airline gets to set the fares in its own hub. A 1989 encounter between America West and Northwest is evidence in support of this conjecture. America West set a low round-trip fare of $258 for the Minneapolis–Los Angeles route. This low fare would largely attract passengers from Northwest, as it has a hub in Minneapolis. Rather than lowering its $308 fare to match America West, Northwest set a new fare that struck directly at America West’s hub in Phoenix. Northwest cut its $208 fare between Phoenix and New York to $168 and, most interestingly, initially made the fare available for only two days. Apparently, America West got the message. Five days after setting its low Minneapolis–Los Angeles fares, America West rescinded them. It has been reported that some airline executives have gone so far as to communicate their personal feelings with respect to being undercut by prefixing new fares with the letters “FU.”

There are other practices and customs in particular industries that aid in reducing the uncertainties of collusive pricing. One example is the use of conventional markup pricing rules. If all firms in an industry become accustomed to calculating prices with the same formula, then pricing changes that reflect industry-wide cost changes are more likely to be understood and accepted.

Another practice is the use of a basing point system of pricing. This system is used in some industries where freight costs are relatively large and consumers are located at various distances from producers. An example is the Pittsburgh-plus system used in the steel industry until it was abandoned in 1924 because of an antitrust order. Under that system, every steel producer could easily quote the same price to a customer in, say, Durham, North Carolina. The price would equal the price at the Pittsburgh mill plus the freight charge from Pittsburgh to Durham—even if the actual producer’s mill was in Birmingham. Hence, basing point systems reduce what would otherwise be a very complicated price quotation problem to a
relatively simple formula. Firms wishing to collude need only agree on a single price at the Pittsburgh mill.

In the case of explicit collusion, coordination is not difficult, as firms can directly communicate their intentions over the phone, through faxes and e-mails, and at face-to-face meetings (though let us not forget there is still the issue of whether those intentions are genuine in the sense that each firm will do what it has agreed to do.) For example, one of the tasks conducted by the global lysine cartel in their meetings was setting the price of lysine in different countries that they served.\(^{15}\) Complicating matters was that exchange rates fluctuate, which required regular adjustment by the cartel.

For example, if, going back to 1993, the U.S. price is set at 90 cents per pound and the price in Japan is set at 112.50 Yen per pound, then, at an exchange rate of 125 Yen to $1, the prices are equal (that is, \(\frac{112.5}{125} = 0.9\)). But now suppose the yen depreciates by 20\% so that the exchange rate is 150 Yen to $1. If the stated lysine price in Japan remains at 112.5 yen, Japanese manufacturers are effectively selling at 75 cents (\(\frac{112.5}{150}\)). This difference in the price could create all sorts of problems in maintaining collusion. For example, some U.S. buyers may choose to buy in Japan and transport it to the United States, which would result in the allocation of sales across cartel members differing from the agreed-upon allocation. Cartel members then need to communicate and realign prices, and that is what they did. Over a period of less than three years, the lysine cartel held ten formal meetings to set prices and conducted almost two dozen other face-to-face meetings.\(^ {16}\)

Even if coordination issues are solved under explicit collusion, challenges remain in that firms may differ in what they wish to coordinate upon. There are typically vast differences between firms in cost and capacity (among other traits) and this translates into different preferences over the collusive outcome. To illustrate how cost differences can create difficulties in agreeing on a particular price, consider the following duopoly setting. Suppose that firms sell identical products but firm 1 has higher cost than firm 2. Assume that they agree to split the market equally. As shown in figure 5.7, each firm faces the same demand curve \(d(P)\), which is just half the market demand curve \(D(P)\).

The profit-maximizing price for the low-cost firm, \(P_L\), is determined by the intersection of marginal revenue and its marginal cost \(MC_L\). In contrast, the high-cost firm prefers the higher price of \(P_H\). Because both firms must set the same price (otherwise one of the firms would have zero demand), there is an obvious conflict. Firms must negotiate in order to resolve this difference. Thus, if firms have different cost functions or products, there is a bargaining problem that the cartel must solve. This compounds the usual coordination problem described earlier.

---

15. Lysine is an organic chemical used in animal feed to promote the growth of muscle tissue.

More important than differences in the desired price are differences among firms as to the appropriate allocation of supply at a given price. While all firms can see the merit in raising price, they will differ as to how much each firm ought to constrain supply in order to achieve that higher price. To flesh out this important challenge to colluding firms, let us return to the model used in the earlier section, A Theory of Collusion, but suppose that the firms’ now have different costs: firm 1 has a constant marginal cost of 35 and firm 2 has a constant marginal cost of 45. The best reply functions are now:

\[ q_1 = 32.5 - 0.5q_2 \]  \hspace{1cm} (5.23)

and

\[ q_2 = 27.5 - 0.5q_1, \]  \hspace{1cm} (5.24)

which are depicted in figure 5.8. The noncollusive solution then has production levels of 25 and 15 for firm 1 and firm 2, respectively, and, with an inverse market demand curve of \(100 - Q\), the price is 60.
Suppose firms consider colluding using the general mechanism described in equations 5.18–5.19, though there’ll be different quantities. Rather than specify the collusive quantities yet, let \( q_1 \) denote what firm 1 is to produce and \( q_2 \) denote what firm 2 is to produce.\(^{17}\) The punishment quantities are now 25 for firm 1 and 15 for firm 2 (instead of 20). The first question to pose is, For what values of \( q_1 \) and \( q_2 \) is this an equilibrium? The analogous equation to (5.22) for firm 1 is:

\[
\frac{(100 - \tilde{q}_1 - \tilde{q}_2 - 35)}{r} \geq \frac{(100 - (32.5 - 0.5\tilde{q}_2) - \tilde{q}_2 - 35)(32.5 - 0.5\tilde{q}_2)}{1 + r} + \frac{625}{r(1 + r)} \tag{5.25}
\]

and for firm 2 is:

\[17. \text{ In equations 5.18–5.19, these variables take the value 15.}\]
One can show that the set of values \( \tilde{q}_1 \) and \( \tilde{q}_2 \) that satisfy equations 5.25–5.26 looks like the gray area in figure 5.8.\(^{18}\) That is, if \( \tilde{q}_1 \) and \( \tilde{q}_2 \) lie in the gray area, then firms can sustain those quantities as part of a collusive equilibrium.

But which pair of quantities will firms settle upon? To make the task easier, suppose firms have decided on raising the price from 60 to 70, which means that industry supply must be reduced from 40 to 30. The line running from the horizontal axis at \( q_1 = 30 \) to the vertical axis at \( q_2 = 30 \) represents all pairs of quantities for which the total supply is 30. The bold portion of that line is the part that intersects the gray area and thus represents all of the values for \( \tilde{q}_1 \) and \( \tilde{q}_2 \) that are sustainable and produce a price of 70. We have then narrowed the problem down to firms choosing from a pair of quantities from the bold line. As one moves from the upper left end to the lower right end of that line, firm 1 is producing a larger share of those 30 units, something which is more preferred by firm 1 but less preferred by firm 2.

We have now reached the point of conflict among colluding firms. Many cartel meetings wrestle over this issue and many a cartel has collapsed from an inability to reach a consensus. One common solution (which, as we shall see later, was used for the citric acid cartel) is historical precedent; that is, collusive market shares are set equal to recent (noncollusive) market shares. If, prior to forming a cartel, the noncollusive solution prevailed, then any pair of quantities on the line running from \((0,0)\) to \((25,15)\) results in the historical market share of 62.5% for firm 1 and 37.5% for firm 2. If firms have agreed on a collusive price of 70, the solution is then the intersection, which yields \( \tilde{q}_1 = 18.75 \) and \( \tilde{q}_2 = 11.25 \).

The lysine cartel exemplifies the difficulty in achieving an agreement on market shares. Beginning in November 1992, lysine producers were able to agree on price, but their failure to reach a consensus as to the allocation of supply ultimately led to the breakdown of collusion. As a result, price fell from a high of 98 cents per pound to 62 cents by June 1993. A stable collusive outcome was only obtained later that year in October, when firms hammered out an agreement on market shares. This persisted until June 1995, when, in response to information from an informant, the FBI raided the offices of Archer Daniels Midland, the U.S. member of the lysine cartel.

**Imperfect Monitoring**

A simplification in our model of collusion is that firms could perfectly observe the past quantity decisions of their competitors. In fact, one of the serious problems faced by some cartels is in determining whether or not someone cheated.

---

18. We set \( r = 0.05 \) in deriving the set in figure 5.8.
Consider the following modification to the infinitely repeated Cournot game. Assume that each firm knows its own past quantities and past market prices but does not observe the past quantities of its competitors. Also suppose that the market price depends on firms’ quantities and that there are random (and unobserved) forces that shift the market demand curve.

The monitoring problem faced by a cartel is as follows. If a firm cheats by producing above the collusive output, price will tend to be lower. However, if price is low, firms cannot infer for certain that someone cheated. A low price is consistent with no one cheating and random forces having shifted the market demand curve (for example, recessionary forces). Because monitoring is imperfect, there is a stronger incentive to cheat because a firm may be able to get away with it. Although collusion is still supportable in such a setting, it is more difficult to sustain.\(^{19}\)

The famous electrical equipment price-fixing cases of the early 1960s provide an example in which imperfect monitoring resulted in cheating and ultimately the breakdown of collusion. The four companies involved were General Electric, Westinghouse, Allis-Chalmers, and Federal Pacific. Their arrangement was to rotate the business on a fixed percentage basis among the four firms. Sealed bids were made so that each firm would submit the low bid a sufficient number of times to obtain the agreed-on market share.

In 1957, the chairman of Florida Power & Light Company, Mac Smith, was buying about a million dollars’ worth of circuit breakers that year and was pressuring G.E. and Westinghouse for price breaks: “Westinghouse had proposed to Florida Power that it add all its circuit-breaker orders to its order for Westinghouse transformers. In return, Westinghouse would take 4 percent off circuit-breaker book and hide the discount in the transformer order.”\(^{20}\) The problem arose when Smith decided to split the circuit-breaker order between G.E. and Westinghouse. G.E. had somehow discovered the attempt by Westinghouse to cheat, and had matched the discount. This action quickly led to a series of deeper and deeper price cuts. Westinghouse offered Baltimore Gas & Electric 5 percent off. A week later Allis-Chalmers reduced its price to Potomac Electric by 12 percent. The process escalated until discounts of 60 percent were reached in the winter of 1957–1958.

**Other Challenges**

In concluding, let us briefly mention two other complications that a cartel faces. We assumed that the demand curve and cost functions do not change over time. In fact, this assumption is far from the truth for some industries. For example, many industries’ demand curves shift

---


over time with the business cycle and seasonal cycles. When demand and cost conditions are changing, the collusive price ought to adjust. If it doesn’t, then the incentive to cheat could change to the point that a firm may find it optimal to cheat. Historically, collusion is more likely to break down during recessions.\(^{21}\)

Thus far the role of entry in destabilizing cartels has been ignored. Our focus has been on internal stability (does any firm have an incentive to cheat?) to the exclusion of external stability (does a potential entrant have an incentive to enter?). It is very important to consider the role of entry when examining collusion. Unless there is some factor that prevents entry, one would expect collusion to attract entrants. In the case of the vitamins cartel of the 1990s, Chinese entry into vitamin C posed a serious problem, as the cartel’s market share dwindled in response to growing Chinese imports. This led to a decline in the extent of collusion in vitamin C, while in those vitamin markets for which entry was difficult due to technological barriers, collusion remained strong and prices high.

**Collusion in Practice**

To bring the preceding analysis to life, three episodes of collusion are examined. The railroad cartel of the nineteenth century exemplifies how, even under explicit collusion, cartels can routinely break down. Though the source of periodic reversions to competition is not known with certainty, observed behavior appears to conform to what happens when there is imperfect monitoring of collusive agreements. More than a century later, the citric acid cartel exemplifies how explicit collusion is conducted. The degree of sophistication and detail is impressive as we observe firms striving to create a well-functioning cartel. Finally, Nasdaq market makers offer an example of tacit collusion and how such can emerge even in markets with many firms.

**Railroads in the 1880s**

Because cartels were legal (though not enforceable by the courts) in the United States prior to 1890, the railroads entered a cartel agreement in 1879 in order to stabilize price. This agreement created the Joint Executive Committee (JEC). The role of the JEC was to set the rail rate of eastbound freight shipments from Chicago to the Atlantic seaboard.

In an empirical analysis, Robert Porter sought to determine whether the JEC maintained collusion through a mechanism similar to the trigger strategy described earlier.\(^{22}\) The collusive device considered by Porter was distinct in two ways. First, the punishment for cheat-


ing is reversion to the Cournot solution (that is, a breakdown of collusion) for some finite number of periods rather than forever. Second, it was assumed that the JEC could only imperfectly monitor the firms’ actions. Because of imperfect monitoring, this theory of collusion predicts that there will be episodes of what looks like cheating. As a result, there will be periodic reversions to the Cournot solution. One would then expect to find periods in which firms are colluding, so that price is high, and periods in which collusion breaks down and firms revert to something more competitive.

Porter examined rail rates for grain during the period 1880–1886 and assessed whether there were periodic switches between collusion and competition. Figure 5.9 shows the movement in the rate from 1880 to 1886. The evidence is that price was relatively high in weeks 0–80 and 120–220. There appeared to have been breakdowns in collusion (“price wars”) in weeks 80–120 and periodically over weeks 220–360. The dashed line below the grain rate indicates periods in which Porter concluded that collusion broke down. The railroad industry appears to have been colluding, but not perfectly. Collusion was intermixed with periodic breakdowns, resulting in intense competition and lower prices.

![Figure 5.9](image_url)

**Figure 5.9**
Cartel Pricing of Rail Rates for Shipment of Grain from Chicago to the Atlantic Seaboard, 1880–1886
Citric Acid in the 1990s

Citric acid is an organic chemical used as an additive in foods and detergents. In that it is a fairly homogeneous product purchased by industrial buyers who will shop around for the best deal, competition can be intense. Toward circumventing such competition, Terrance Wilson and Barrie Cox, high-level managers of Archer Daniels Midland (ADM), flew to Europe in January 1991 to meet with representatives of the three largest European manufacturers to discuss the possibility of forming a cartel. This trip paid off two months later with the first meeting of the lysine cartel.23

At this and ensuing meetings, firms agreed to raise prices and allowed each company to offer a 3 percent discount to their five biggest customers. Market shares were set at each firm’s historical market share over 1988–1990, which meant that Haarman & Reimer/Bayer was allocated 34 percent, ADM received 27.5 percent, and Jungbunzlauer and Hoffman-La Roche got 24 percent and 14.5 percent, respectively. In order to monitor the agreement, each firm was required to submit monthly sales volumes to a representative of Hoffman-La Roche, who would then compile these statistics and distribute the information among the other cartel members. In the event that realized market shares did not match targeted shares, a buy-back system was put in place to even things out.

During the four years in which the cartel was operating, full-scale meetings occurred about every eight weeks, with some of them taking place during the meetings of the European Citric Acid Manufacturers’ Association (ECAMA). It is quite typical to use trade association meetings as a pretense for firms’ representatives to get together, though the cartel meeting is certainly not on the formal agenda of the association! At these meetings, there would be a discussion of the latest cartel sales report—toward monitoring the agreement—and a discussion of trends in demand and cost, which would lead up to a decision on the new cartel price. The representatives would also discuss any problems with the cartel’s operation, such as accusations about cheating.

As shown in figure 5.10, the cartel was successful. The price of citric acid rose by 35 percent, though it took about two years of gradual price increases to reach that level. The rising price did bring forth a significant supply response from Chinese manufacturers of citric acid who were not members of the cartel. Though of lower quality, the volume of Chinese imports to the United States rose by 90 percent over 1991–1993. This rise of noncartel supply, along with overproduction by cartel members, led to a price decline toward the end of the cartel’s life.

Although collusion was deteriorating on its own, the citric acid cartel was formally shut down by the antitrust authorities. The U.S. Department of Justice learned about collusion in

citric acid from its investigation of the lysine cartel. ADM employee Mark Whitacre, who was a high-level executive participating in the lysine cartel, had turned FBI informant in November 1992. This permitted the FBI to tape meetings of the lysine cartel. At one of those meetings, Terrance Wilson mentioned ADM’s participation in the citric acid cartel as a “model” of how successful collusion could be achieved with lysine. He urged them to form an association like the ECAMA and to use an independent accounting firm to calculate members’ sales, as was done in citric acid. It was at one of those lysine cartel meetings that he uttered the now infamous statement: “The competitor is our friend, and the customer is our enemy.”

Figure 5.10
List and Contract Prices of Citric Acid, 1987–1997

Nasdaq in the 1990s

On the New York Stock Exchange, stocks are traded by physically bringing together traders and having prices set through an auction. In contrast, the National Association of Securities
Dealers Automated Quotation (Nasdaq) system operates in a decentralized manner through a network of market makers (also called dealers). At least two market makers are associated with each Nasdaq stock. Each market maker posts the price at which it is willing to buy a stock (known as the bid price) and the price at which it is willing to sell a stock (known as the ask price). Bid and ask prices are posted on a network and are observed by all market makers. The difference between the lowest ask price and the highest bid price quoted by dealers forms the “inside spread” and is how dealers earn a return on their services. For example, if the lowest ask price is $10\frac{3}{8}$—so that the next dealer sale will be at $10\frac{3}{8}$—and the highest bid price is $10$—so that the next dealer purchase will be at $10$—then the spread is $\frac{3}{8}$. Dealers make a gross return of $\frac{3}{8}$ on each share transacted. Each dealer competes for investors’ buy and sell orders by posting lower ask prices and higher bid prices than competing dealers. More intense competition should result in more compressed spreads. By the rules of Nasdaq (prior to June 1997), the minimum spread was $\frac{1}{8}$ for stocks whose bid price exceeded $10$ in that Nasdaq required price fractions for bid and ask quotes to be multiples of $\frac{1}{8}$. For example, quotes could be $10\frac{1}{8}$ and $10\frac{1}{4}$ but not $10\frac{3}{16}$.

While engaging in some academic research on Nasdaq prices, William Christie and Paul Schultz noticed something odd (or, as we will see, not odd enough). 25 Many of the Nasdaq stocks they were examining were rarely quoted in prices with odd eighths. For these stocks, almost all quotes were in even eighths, for example, $10$ and $10\frac{1}{4}$ but not $10\frac{1}{8}$. Examining the 100 most actively traded Nasdaq stocks in 1991, the market makers for seventy of them rarely used odd-eighth quotes. The affected markets included such highly traded stocks as Intel, Amgen, and Microsoft, whose stocks often attracted up to sixty market makers at any one time.

What could cause such an unusual pattern of behavior? One possibility is that it was simply a product of the transactions technology. Perhaps dealers had to earn a spread of at least $\frac{1}{4}$ to cover their costs and, for whatever reason, an arbitrary rule of dealing in quarters became the norm. Further study raised doubts about this explanation. Comparing these hundred Nasdaq companies with a hundred companies trading on the New York and American Stock Exchanges that were of similar price and market capitalization (and where odd-eighth price

fractions were used with about the same frequency as even eighths), Christie and Schultz found the average spread on the Nasdaq stocks to be significantly higher. This finding suggested another hypothesis, that the dealers of these stocks had developed a collusive norm of avoiding odd-eighth quotes with the intent of imposing a minimum inside spread of \( \frac{1}{4} \).

Evidence consistent with this hypothesis was the response of spreads to the public release of this study in the *Los Angeles Times* on May 26, 1994. On the very next day, the number of market makers of Microsoft who exclusively used even eighths dropped from forty-one to one.26 Figure 5.11 shows the precipitous drop in the average inside spread. Although it had

been consistently above $\frac{1}{4}$, it was now well below it. One could argue that market makers, recognizing that their collusive practices had been revealed, chose to return to normal competitive behavior out of fear of adding to the damages they would already be likely to pay.

While this pattern of pricing is consistent with tacit collusion by Nasdaq market makers, there is no evidence that they explicitly communicated about avoiding odd-eighth quotes. Are industry conditions ripe for collusion? It is not clear. For many of these markets, the number of dealers is quite large. In 1995 the average number of market makers per Nasdaq security was around ten, and some of the large, most actively traded stocks had in excess of sixty. Such a large number of firms could make collusion quite difficult to sustain. Compounding this factor is the ease of entry that makes the arrival of a new market maker another source of instability for a collusive arrangement. However, a feature conducive to collusion is the near-perfect monitoring of firm behavior. Bid and ask quotes are instantly disseminated to all dealers, so that any deviation from the collusive norm would be immediately detected and, in principle, punished. Also, given the perfect substitutability of products (the shares bought and sold by one dealer are the same as the shares bought and sold by another dealer), punishment could be severe if the other dealers decided to price a deviator’s business away.

A class action suit was filed in 1994 by investors alleging that thirty-seven Nasdaq dealers colluded to keep spreads artificially high. In December 1997, thirty-six of those dealers agreed to an out-of-court settlement of around $1 billion while not admitting any wrongdoing.

---

**Antitrust Law and Policy toward Price Fixing**

Section 1 of the Sherman Act outlaws “every contract, combination . . . or conspiracy in restraint of trade.” If interpreted literally, this language would make nearly every type of business agreement or contract illegal. Two lawyers who decide to form a partnership would be illegally in restraint of trade since they eliminate competition between themselves. In the 1911 *American Tobacco* case, however, the Supreme Court stated that “the words ‘restraint of trade’ . . . only embraced acts or contracts or agreements or combinations . . . which, either because of their inherent nature or effect or because of the evident purpose of the acts, etc., injuriously restrained trade.” This view persists to this day and has been more recently acknowledged by the Supreme Court in the 1997 *State Oil Company v. Khan* decision: “[A]lthough the Sherman Act, by its term, prohibits every agreement ‘in restraint of trade,’

---

27. Probes by the Securities and Exchange Commission and the Department of Justice did conclude that this pattern of pricing violated Section 1 of the Sherman Act.
In this Court has long recognized that Congress intended to outlaw any unreasonable restraints.\(^{28}\)

Subsequent opinions have developed tests for identifying unreasonable restraints which have become known as the *per se rule* and the *rule of reason*. When a practice can have no beneficial effects and only harmful effects, the “inherent nature” of the practice is injuriously restraining trade. Price fixing by a cartel seems to fit this description and is now illegal per se. This means that the behavior need only be proved to have existed; there is no allowable defense. If a certain practice does not qualify as a per se offense, the rule of reason applies. This term refers to the tests of “inherent effect” and “evident purpose.” For example, a merger between two firms in the same market is not necessarily harmful or beneficial. Hence the court must then look to the “inherent effect” of the merger and its “evident purpose” or intent. A merger would be judged as legal or not depending on an evaluation of the evidence concerning the actual intent of the firms to monopolize the market and their ability to realize any such intent as measured by the market shares involved and the ease with which entry can reverse monopolization.

Per se rule antitrust cases tend to be much shorter than the lengthy deliberations often found in rule of reason cases. The IBM monopolization case, an extreme example of a rule of reason antitrust case, lasted thirteen years and involved 950 witnesses, 726 trial days, 17,000 exhibits, and 104,400 trial transcript pages. Justice Thurgood Marshall has described the rationale for the per se category:

> Per se rules always contain a degree of arbitrariness. They are justified on the assumption that the gains from imposition of the rule will far outweigh the losses and that significant administrative advantages will result. In other words, the potential competitive harm plus the administrative costs of determining in what particular situations the practice may be harmful must far outweigh the benefits that may result. If the potential benefits in the aggregate are outweighed to this degree, then they are simply not worth identifying in individual cases.\(^{29}\)

### Economic Analysis of Legal Categories

These categories are also generally consistent with economic analysis. A simple example should make this point clear. Consider the model developed by Oliver Williamson to illustrate the possible trade-offs created when two firms in the same market seek to merge.\(^ {30}\)

In figure 5.12 the initial price is \(P_0\) and output is \(Q_0\). The degree of competition is assumed to be sufficient to force price down to \(AC_0\). Now assume that a merger takes place that creates

---


both cost savings and market power. Hence the postmerger equilibrium results in a price increase to \( P_1 \) and a cost reduction to \( AC_1 \). Output falls from \( Q_0 \) to \( Q_1 \).

The merger results in a deadweight loss in consumers’ surplus equal to triangle \( A_1 \) in figure 5.12. (Recall the discussion of deadweight loss in chapter 4 at figure 4.2.) On the other hand, there is a gain to society as a result of the cost savings, which is shown by the rectangle \( A_2 \) in figure 5.12. That is, \( A_2 \) represents the cost savings in producing output \( Q_1 \) at an average cost of \( AC_1 \) rather than \( AC_0 \).

The analysis here follows that of Robert Bork.\(^{31}\) He makes the point that the diagram can be used to illustrate all antitrust problems, because it shows the relationship between efficiency gains and losses. Bork also stresses that efficiency gains through cost savings are but one form of increased efficiency—and that area \( A_2 \) should be taken to symbolize any efficiency gain and not merely cost reductions. An example might be an improvement in competitive effectiveness achieved through combining a firm with marketing expertise with one possessing research and development skills.

Of course, not all mergers produce both gains and losses. Some may result in only one or the other, or neither. It is, however, appropriate for the courts to investigate this issue rather than simply declaring mergers to be illegal per se. Ideally, then, a rule of reason decision would find mergers with no gains but actual losses to be illegal. Similarly, mergers with positive gains and no losses would be legal.

The cases where both gains and losses occur are more difficult. The argument could be made that the antitrust authority should compare the loss of \( A_1 \) with the gain of \( A_2 \). If \( A_2 \) exceeds \( A_1 \), the merger should be permitted, and not otherwise. In chapter 7, where we discuss mergers in depth, we will return to this issue and describe the Federal Trade Commission’s approach in the context of the Staples–Office Depot merger case.

A cartel, in contrast to a merger that integrates the productive activities of the firms, can lead only to the area \( A_1 \) losses. Cost savings are quite unlikely without actual integration. Hence it is sensible to place cartels that attempt to fix prices or allocate markets in the per se category. The “inherent nature” of price fixing is to suppress competition, and there are no beneficial effects.

**Per Se Rule Cases**

As we have noted, the courts have taken a clear position with regard to overt conspiracies by competitors to fix prices or share markets. A number of early railroad cases struggled with the issue, as did an interesting case involving cast-iron pipe manufacturers. This latter case, *Addyston Pipe*, was decided on appeal by the Supreme Court in 1899.\(^{32}\)

The opinion of the Circuit Court of Appeals, written by Judge (later President) William Howard Taft, is regarded by some as one of the greatest antitrust opinions in the history of the law. According to antitrust expert Robert Bork, Taft made “a remarkable attempt to settle the issue of goals and to provide the Sherman Act with a workable formula for judging restraints. The opinion is one of almost unparalleled suggestiveness, and yet its potentialities, after more than seventy years, remain almost entirely unexploited.”\(^{33}\)

The per se rule toward price fixing was firmly established in the Supreme Court’s *Trenton Potteries* decision in 1927. Some twenty-three manufacturers of sanitary pottery belonged to an association that attempted to fix the prices of their products. The defendants had roughly 82 percent of the market. In its decision, the Court concluded:

The aim and result of every price-fixing agreement, if effective, is the elimination of one form of competition. The power to fix prices, whether reasonably exercised or not, involves power to control the market and to fix arbitrary and unreasonable prices. The reasonable price fixed today may through economic and business changes become the unreasonable price of tomorrow. . . . Agreements which create such potential power may well be held to be in themselves unreasonable or unlawful restraints, without

---

\(^{32}\) *United States v. Addyston Pipe & Steel Co.*, 175 U.S. 211 (1899).

the necessity of minute inquiry whether a particular price is reasonable or unreasonable as fixed and without placing on the Government in enforcing the Sherman Law the burden of ascertaining from day to day whether it has become unreasonable through the mere variation of economic conditions.  

Notwithstanding its rather strong statement in Trenton Potteries, the Supreme Court six years later seemed to refute its earlier opinion. The Appalachian Coals case concerned some 137 companies that joined together to form Appalachian Coals, Inc. The new company was to act as exclusive selling agent for the companies. These 137 firms accounted for 12 percent of all bituminous coal produced east of the Mississippi River and 54 percent of the production in the Appalachian region. A key fact was that the country was in the midst of the Great Depression, and coal mining was especially depressed. Prices of coal fell by 25 percent from 1929 to 1933, and the majority of coal-mining firms suffered losses.

The government brought suit, and the district court in 1932 found Appalachian Coals in violation of the Sherman Act. However, upon appeal, the Supreme Court reversed the decision. The opinion observed that “a close and objective scrutiny of particular conditions and purposes is necessary in each case. . . . The mere fact that the parties to an agreement eliminate competition among themselves is not enough to condemn it.” This is clearly in opposition to the per se rule enunciated in Trenton Potteries. It can be explained perhaps by the “deplorable” conditions of the industry and the desire on the part of the Court to assist in bringing about a “more orderly” marketing system. And, as Richard Posner has observed, “Faith in the policy of competition was deeply shaken by the depression of the 1930s; this more than anything may explain the outcome in the Appalachian Coals case.”

This apparent inconsistency was remedied by the Court in a 1940 decision, when it once again issued a strong per se rule toward price fixing. We should note that the composition of the Court had changed significantly from the 1933 version. Also, by 1940 there was little remaining public support for cartelization as a remedy for depressions. (The National Industry Recovery Act, which promoted this idea, was ruled unconstitutional in 1935.)

The 1940 decision, Socony-Vacuum, involved the gasoline industry. Independent refiners were dumping gasoline at very low prices. During 1935 and 1936 more than a dozen major oil refiners, including Socony-Vacuum (now known as Mobil), agreed to a coordinated purchasing program to keep prices up. Each major refiner selected “dancing partners” (that is, independent refiners) and was responsible for buying the surplus gasoline placed on the market by its partners. The Supreme Court, on appeal, sustained the verdict of guilty. The Court stated flatly that “price-fixing agreements are unlawful per se under the Sherman Act

and that no showing of so-called competitive abuses or evils which those agreements were designed to eliminate may be interpreted as a defense.”

Since the 1940 decision, the per se rule toward price fixing has been perhaps the most unambiguous antitrust rule of law. Even so, there are difficult, gray areas of collective activity that pose problems. Trade association activities are an example. Trade associations collect and disseminate information on a wide variety of topics to their members. While normally information helps make markets work better, price information can lead to diminished price competition, especially in oligopolistic markets. A detailed examination of trade association cases is beyond the scope of this book. Suffice it to say that the courts have adopted a rule of reason approach, and that no clear principles of what is legal or illegal have emerged.

Another interesting issue on the applicability of the per se rule arises with professions. Until the 1970s, doctors, lawyers, engineers, and other professions were apparently viewed as outside the jurisdiction of the Sherman Act. In the 1975 Goldfarb case, however, the Supreme Court found a bar association in violation of Section 1. The Virginia Bar Association, as was common practice at the time, had circulated a list of suggested minimum attorney’s fees for various services. For example, the fee for title search in connection with the sale of real estate was 1 percent of the value of the property. The Bar Association issued an opinion that “evidence that an attorney habitually charges less than the suggested minimum fee schedule adopted by his local Bar Association, raises a presumption that such lawyer is guilty of misconduct.” (Generally, professional associations argued that price competition would lead to reduced quality of services and that consumers, unable to judge quality, would be harmed.)

The Goldfarbs, planning to buy a house in Fairfax County, Virginia, believed the 1 percent fee for title search to be too high. Hence they consulted some thirty-six lawyers and none would give them a lower price. As a result, the Goldfarbs brought a price-fixing suit against the Bar Association. The Supreme Court ultimately decided in the Goldfarbs’ favor. Subsequent cases have reinforced the interpretation that professional organizations cannot restrict price competition.

Interestingly, the United States stands alone in its per se rule against price fixing. Other industrial nations have adopted a rule of reason approach. For example, the United Kingdom has set forth eight “gateways” that can enable a price-fixing agreement to escape illegality. One gateway is that the agreement can be legal if the court considers the agreement necessary to avoid serious and persistent unemployment effects.

We conclude this section on per se rule cases with a case involving college football. The Supreme Court argued that because of the special characteristics of sports, the rule of reason should apply, even though the alleged offense was horizontal price fixing.

The National Collegiate Athletic Association (NCAA) consists of some 850 voting members—colleges and universities with athletic programs. It serves as a regulatory body that sets playing rules, standards of amateurism, rules about recruitment of athletes, the size of athletic squads and coaching staffs, and so forth. The issue in the case here relates to the NCAA’s restraint on its members’ rights to negotiate their own television contracts for football games. In 1981 the NCAA negotiated contracts with two television networks, ABC and CBS, that limited the total number of games that could be televised and the number by each institution. It also effectively set the price that each institution could receive per telecast. A number of major college football programs decided to negotiate their own telecasts with NBC. In response to a publicly announced threat of disciplinary action by the NCAA, the major football programs filed an antitrust suit against the NCAA.

While a lower court held that the NCAA television plan constituted illegal per se price fixing, the Supreme Court held that the rule of reason should apply. The Court reasoned that college football is “an industry in which horizontal restraints on competition are essential if the product is to be available at all.” Hence, unlike many industries like the steel industry, in which firms need not coordinate their activities, sports leagues need to do so to schedule games, set standards, and so on. In such cases, horizontal restraints are not, on balance, always harmful.

Despite opting for the rule of reason approach, the Court decided against the NCAA. The main justification put forth by the NCAA was that its television plan would protect live attendance at games. Fans who could watch Notre Dame on television might not go to a Podunk College game in their own city. The Court decided that protecting an inferior product from competition was inconsistent with the basic policy of the Sherman Act. “The rule of reason does not support a defense based on the assumption that competition itself is unreasonable.”

Tacit Collusion

A much more difficult antitrust problem is what is termed *tacit collusion* or *conscious parallelism*. Suppose that there is no evidence that firms got together and made overt agreements to fix prices; however, the firms did behave in parallel fashion by charging identical prices. The key question is whether the conspiracy can be inferred from such behavior. From the viewpoint of economic performance, the outcome of a higher price is just as harmful as if the firms operated a cartel. Should not then both types of collusion be treated equally under the antitrust laws? One obvious problem is that the remedies would probably need to differ.

A cartel could stop communicating, but tacit understandings might require a structural dissolution of the industry to be made ineffective. 

As we will see by reviewing the leading cases, the courts have taken what Judge Richard Posner calls a “cops and robbers” approach to price fixing. By this he means that the weapons that the law had developed to deal with conspiracies in other areas were focused on price fixing. Hence the inquiry became limited to the question of whether the defendants had met or communicated with one another. “Once the conspiracy approach to explicit collusion became firmly ensconced in the minds of bench and bar, it was perhaps inevitable that tacit collusion would be considered beyond the reach of the antitrust laws because, by definition, it did not involve explicit, detectable acts of agreement or communication.”

Several early cases indicated that the courts would infer conspiracy from parallel behavior if some additional evidence existed. A 1939 case, Interstate Circuit, involved the manager of a motion picture exhibition chain in Texas and eight motion picture distributors. The exhibitor sent identical letters to the distributors (Paramount, RKO, and so on), naming all eight as addressees, and demanding certain restrictions. For example, the manager demanded that the distributors not release their first-run films to theaters charging less than 25 cents admission. After the letters were mailed, the distributors did exactly what the exhibitor had demanded. However, there was no evidence of meetings or other communications among the distributors. The parallel behavior of the distributors, plus the letter, was sufficient for the Supreme Court to find illegal conspiracy. According to the Court:

It taxes credulity to believe that the several distributors would, in the circumstances, have accepted and put into operation with substantial unanimity such far-reaching changes in their business methods without some understanding that all were to join, and we reject as beyond the range of probability that it was the result of mere chance.

In the 1946 American Tobacco decision, the Court seemed to state that the cigarette industry was guilty of conspiracy based solely on its parallel pricing behavior. It observed that “no formal agreement is necessary to constitute an unlawful conspiracy. Often crimes are a matter of inference deduced from the acts of the person accused.” William Nicholls interpreted this

40. However, one possible direction that remedies, short of dissolution, might take is illustrated by the U.S. turbo-generator industry. The two rivals, G.E. and Westinghouse, were able to maintain stable prices for many years through the use of a pricing book and a “price protection” plan. Under the “price protection” plan, the sellers guaranteed that any discount on new orders would apply retroactively on all orders taken in the past six months. The effect of the plan was to make the incentive to “cheat” through secret price cuts much less. In 1977 the Justice Department got the companies to agree to drop the price protection plan; the hope is, of course, that price competition will be stimulated.


decision as “a legal milestone in the social control of oligopoly” because it permitted “the inference of illegal conspiracy from detailed similarity of behavior.”

Briefly, the facts in the case that were viewed as particularly significant were as follows:

1. On June 30, 1931, Reynolds announced an increase in its wholesale price for 1,000 cigarettes from $6.40 to $6.85. The other two major firms, American and Liggett & Myers, followed upward to the same price within twenty-four hours. This increase occurred in the midst of the Great Depression, when leaf prices and labor costs were falling.

2. In November 1932, after a loss of 23 percent of the market to smaller firms selling “economy brands,” the Big Three dropped their prices, almost in unison, to $5.50. This reduction rapidly forced many of the economy brand suppliers out of business.

3. The Big Three bought large amounts of low-grade tobacco, thereby bidding up its price, even though they did not use it for their own cigarettes. It was the type of tobacco used for economy brands.

4. The Big Three declined to participate in leaf tobacco auctions unless buyers from all three companies were present, and they refrained from buying tobacco grades in which the others had a special interest.

These facts were sufficient for the courts to infer the existence of a conspiracy. However, more recent cases do not seem to support the view that parallel oligopoly pricing alone will be found to be illegal. In the 1954 Theatre Enterprises decision the Supreme Court concluded:

The crucial question is whether respondents’ conduct toward petitioner stemmed from independent decision or from an agreement, tacit or express. To be sure, business behavior is admissible circumstantial evidence from which the fact finder may infer agreement . . . But this court has never held that proof of parallel business behavior conclusively establishes agreement or, phrased differently, that such behavior itself constitutes a Sherman Act offense. Circumstantial evidence of consciously parallel behavior may have made heavy inroads into the traditional judicial attitude toward conspiracy; but “conscious parallelism” has not yet read conspiracy out of the Sherman Act entirely.

And, in a widely publicized case involving parallel pricing of tetracycline by five pharmaceutical manufacturers, a district court found the firms to be innocent. A key fact in this case was that Pfizer, Cyanamid, Bristol, Upjohn, and Squibb each charged the same price of $30.60 for its brand of tetracycline continuously from November 1953 until July 1960. Much attention was given to the low manufacturing cost of around $3.00, and therefore the extremely high profit margins.

In their defense, the companies argued that it would make no sense for any one of them to cut its price. The market demand for tetracycline was price inelastic, so that a general price reduction would not expand the total market. Also, a price cut by one would be matched immediately by the others, making such a tactic self-defeating. Furthermore, entry was barred by virtue of Pfizer’s patent; hence, there was no need to lower price to limit entry. Therefore the maintenance of a common price of $30.60 by the five firms did not imply a conspiracy. The Court apparently agreed with this line of argument, concluding that “the parallel pricing among the tetracycline producers, standing alone, does not indicate price fixing.”

How can we reconcile these more recent decisions with the earlier American Tobacco case? Or, more to the point, how are the courts likely to draw the line in determining what circumstantial evidence warrants an inference of conspiracy? The accumulated precedents have been summarized as “parallelism plus.” In the tobacco case it can be argued that the “plus” could be inferred from the fact that the firms had advance knowledge of impending rival actions that could hardly have been gained without covert communications.

Though tacit collusion remains largely immune to prosecution, some scholars and judges are not content for it to remain that way. Judge Posner argues that tacit collusion can, in some instances, be thought as a contractual arrangement and, on those terms, warrants prosecution:

...one seller communicates his “offer” by restricting output, and the offer is “accepted” by the actions of his rivals in restricting their outputs as well. It may therefore be appropriate in some cases to instruct a jury to find an agreement to fix prices if it is satisfied that there was a tacit meeting of the minds of the defendants on maintaining a noncompetitive pricing policy. ... What is being proposed is less the alteration of the substantive contours of the law than a change in evidentiary requirements to permit illegal price fixing to be found in circumstances in which an actual meeting of the minds on a non-competitive price can be inferred even though explicit collusion cannot be proved.

As Judge Posner recognizes, the “devil lies in the details.” Most methods we can think of for inferring tacit collusion are plagued with the serious possibility of wrongly convicting firms that did nothing more than take account of their rivals’ behavior as part of the normal process of competition. For example, while the apparent response of reducing quantity in

46. Pfizer had licensed the other four firms to sell tetracycline.
48. A case decided in 1984 by the U.S. Court of Appeals is consistent with the recent decisions in finding for the firms. Unlike the cases discussed, in which the legal issue was one of inferring conspiracy from parallel behavior, the FTC charged the firms with competing unfairly under the FTC Act. The charge was that the firms engaged in “facilitating practices” that made tacit collusion easier to achieve. The facilitating practices included quoting prices on a uniform delivered price basis, announcing price changes to customers well in advance of the effective date, and including a “price protection plan” similar to the one described in note 34. For a study of this case, see George A. Hay, “Practices That Facilitate Cooperation: The Ethyl Case,” in John E. Kwoka, Jr., and Lawrence J. White, eds., The Antitrust Revolution, 2nd ed. (New York: HarperCollins, 1994).
response to a decrease in a rival’s quantity is contrary to the usual formulation of competition (recall from figure 5.3 that a firm’s best reply function is downward sloping), such an observation can be rationalized as being consistent with competition in a number of ways, including an inward shift of market demand and a common increase in input prices. Whether we can expect to have the data and the methods to distinguish those possibilities to a reasonable level of confidence is the key open question.

**Enforcement Policy**

Within the federal government, price-fixing cases are handled by the Antitrust Division of the U.S. Department of Justice (DOJ), though they can also be pursued through private litigation. If the DOJ takes a case and succeeds in establishing guilt (either through a judicial decision or a plea settlement with the firms) then private litigation typically ensues to collect compensation for the victims. However, a defendant may plead *nolo contendere*, which means that it neither admits nor denies involvement in a crime but accepts punishment as though it were guilty. With such a plea, plaintiffs must then establish guilt prior to arguing for compensation.

**Penalties**

When an employee for a corporation violates antitrust laws, he puts himself and the corporation at risk of receiving serious penalties. For corporations, they take the form of financial penalties and include both fines levied by the government and damages awarded to victims in association with a private suit. In the latter case, Section 4 of the Clayton Act of 1914 specifies that any individual “who shall be injured in his business or property by reason of anything forbidden in the antitrust laws” may sue to recover “threelfold the damages by him sustained. . . .” Under this condition, known as *treble damages*, victims collect three times their calculated harm. Individuals guilty of price fixing can be subject to fines and also prison sentences.

Before delving into the specifics of how fines and damages are calculated, let us think about how one would want to calculate them if the intent is to deter collusion. For this purpose, consider a simple model in which two firms may collude and, if they do, may be discovered. Each firm has a constant marginal cost of $c$. If they collude, then the price is $\hat{P}$ with resulting industry quantity $\hat{Q}$, so that a firm’s collusive profit is $\pi = \frac{1}{2}(\hat{P} - c)\hat{Q}$, because as each produces half of $\hat{Q}$. If firms do not collude, then price is $\hat{\hat{P}}$ (which is less than $\hat{P}$), with noncollusive industry supply of $\hat{Q}$ and firm profit of $\pi = \frac{1}{2}(\hat{\hat{P}} - c)\hat{Q}$. Suppose that the probability of the cartel being discovered and successfully prosecuted is $z$, where $0 < z < 1$. If the penalty for each firm is $X$, then the expected profit from colluding is $\pi - zX$, where the expected penalty is $zX$. 
Collusion is then deterred if the profit from not colluding exceeds that from colluding, which is the condition:

\[ \hat{\pi} > \pi - zX. \]

Rearranging, it is equivalent to

\[ X > \frac{1}{z} (\pi - \hat{\pi}) \]

so that deterrence is achieved when the penalty is sufficiently high. If we interpret the “damage” as the increase in profit due to collusion, \( \pi - \hat{\pi} \), deterrence then requires the penalty to be a multiple of damages where this multiplicative factor is at least \( 1/z \). For example, when the probability of the cartel being discovered and successfully prosecuted is only 10 percent, then the multiple must be 10. Since the cartel will only have to pay a penalty with some probability, if it was forced only to return its ill-gotten profit, then it would always be optimal to form a cartel. Only by making the penalty sufficiently large relative to the gain in profit can collusion be deterred. This is a rationale for the Clayton Act setting the damage multiple well above one.

In practice, reality departs considerably from the above calculus and, in particular, as to how damages are calculated. Standard antitrust practice is to calculate damages not as the gain in profit from colluding but rather as the additional revenue on the units sold: \((\hat{P} - \hat{P})\hat{Q}\). \( \hat{P} \) is known as the “but for price” which is the price that would have occurred had it not been for collusion, while \( \hat{P} - \hat{P} \) is referred to as the “overcharge.” To see how this formula differs from the gain in industry profit from collusion, let us refer to figure 5.13. The gain to firms from colluding is rectangle \( A \) (which measures the higher profit earned on the \( \hat{Q} \) units produced while colluding) minus rectangle \( B \) (which measures the forgone profit from producing \( \hat{Q} - \hat{Q} \) fewer units). However, damages are calculates to include only rectangle \( A \) and thus overestimate the gain to firms by rectangle \( B \). Looking at this from the perspective of consumers, the cost to consumers from collusion is rectangle \( A \) plus triangle \( C \), which measures the forgone surplus from consuming \( \hat{Q} - \hat{Q} \) fewer units. Interestingly, we then have that calculated damages exceed the amount by which firms benefit from collusion but fall short of how much consumers are harmed.

It is important to recognize that if the courts find firms guilty of price fixing, the firms are required to pay the victims treble damages. However, most cases are settled out of court and, in that case, single damages are quite common.\(^{50}\) In practice, deterrence may then be quite weak.

---

Damages are generally much higher than government fines. Indeed, historically, fines have been trivial. As recently as the 1980s, the average government fine in antitrust cases was only $368,000, which is paltry for most corporations.\(^{51}\) In recent years, however, the significance of fines has grown tremendously, the average fine having exceeded $4.75 million during the 1990s.\(^{52}\)

Table 5.3 shows the largest fines associated with Sherman Act violations; they now reach into hundreds of millions of dollars. Two policy changes have been instrumental here. The revision of the federal sentencing guidelines in 1991 allowed the government to set much higher fines, and the revision of the corporate leniency program in 1993 (which is discussed later) has induced firms to agree to pay larger fines by enhancing the bargaining position of the government.

In practice, the sentencing guidelines allow a base fine to equal 20% of the sales of the cartel members during the time of the conspiracy. With this base, a range for the fine is

\[ \text{Demand} \]

\[ A \]

\[ B \]

\[ C \]

\[ \hat{P} \]

\[ \dot{P} \]

\[ \hat{Q} \]

\[ \dot{Q} \]

\[ \text{Figure 5.13} \]

Calculating Damages in a Price-Fixing Case

52. Posner, \textit{Antitrust Law}, 2nd ed., p. ••.
calculated by deriving a maximum and minimum multiplier. These multipliers depend on an organization’s “culpability score,” which is calculated by taking into account aggravating behavior (such as a history of misconduct and high level employees being involved) and mitigating behavior (such as accepting responsibility). Here is an example from the vitamins cartel. For Hoffman-La Roche, the DOJ calculated their sales during the conspiracy to be $3.28 billion, which meant a base fine of $656 million (= 3,280,000,000 \times 0.2). Starting with the standard culpability score of 5, there was a 5-point upward adjustment as high-level executives participated and the organizational unit committing the offense had more than 5,000 employees. Another 2 points were added because the firm had a history of collusion, and yet another 3 points for obstructing the government’s investigation. It received a 2-point downward adjustment for accepting responsibility and fully cooperating. The net result was a culpability score of 13, which resulted in a minimum and maximum multiplier of 2 and 4, respectively. This meant that the fine should lie in the range of $1.3 billion (= 656,000,000 \times 2) and $2.6 billion (= 656,000,000 \times 4). What then often happens is that there is a final fudge factor, which, in the case of Hoffman-La Roche, allowed it to pay a fine of only $500 million.

---

Corporate Leniency Program

In 1978, the DOJ established a program whereby corporations and individuals who were engaging in illegal antitrust activity (such as a price fixing) could receive lenient treatment if they fully cooperated in an investigation. Leniency means not being criminally charged for the activity being reported, which allows a corporation to avoid government fines (though it is still liable for private damages) and an individual to escape fines and prison sentences. In spite of the potential appeal of amnesty, the program was rarely used, and one likely reason is that its design left considerable uncertainty as to whether an application for leniency would be approved. In particular, leniency would be denied if the government could have “reasonably expected” that it would have learned of the cartel without the applicant’s assistance.

Under the 1978 program, leniency applications averaged about one per year. Soon after its revision in 1993, applications were coming in at the rate of two per month. What led to such a radical increase? The DOJ made several substantive changes. It laid out a much clearer set of conditions for a leniency application to be approved, which served to reduce uncertainty. In addition, it allowed amnesty in cases for which an investigation had been started. While firms are unlikely to apply for leniency when the authorities do not even have a hint that collusion is occurring, there is a much stronger incentive if the authorities suspect a cartel exists, in which case the prospect of prosecution may be imminent. Finally, one of the conditions for leniency is that the DOJ “has not received information about the illegal activity being reported from any other source.” This meant that amnesty is limited to one firm per cartel, which can create a “race to the courthouse,” as a firm may apply for leniency simply out of fear that another firm will beat them to it.

To appreciate the power of the leniency program, let us examine the incentives of a firm to apply for amnesty. Using our previous notation, let $\bar{\pi}$ and $\hat{\pi}$ denote the profit a firm earns from colluding and competing, respectively. Breaking the penalty into two parts—fines (denoted $F$) and damages (denoted $D$)—amnesty means avoiding $F$ though still paying $D$. Suppose the market has two firms, they form a cartel, and the DOJ has become suspicious about collusion. Assume that, in the absence of a firm coming forward with evidence, each firm believes the DOJ will be able to successfully prosecute with probability $w$, where $0 < w < 1$. Each firm must decide whether or not to apply for leniency.

The situation the two firms face is summarized by the payoff matrix in table 5.4, where the first number in a cell is the payoff to firm 1 and the second number is firm 2’s payoff. If one or both apply for leniency then collusion falls apart and otherwise is maintained. If both choose not to apply, then they are not prosecuted with probability $1 - w$, in which case they

continue to earn collusive profit, and are successfully prosecuted with probability \( w \), in which case they receive noncollusive profit and pay damages and fines. This yields an expected payoff for a firm equal to \((1 - w)p\hat{\pi} + w(p\hat{\pi} - D - F)\). If both apply for leniency, then let us suppose the DOJ gives each a 50 percent reduction in the fine (or, alternatively, flips a coin to decide who gets full amnesty), so that a firm can expect to pay \( D + (F/2) \). Finally, if only one firm applies for leniency, then that firm pays penalties of \( D \), while the other firm is stuck paying \( D + F \).

In analyzing the Nash equilibria for this game, first note that both applying for leniency is always an equilibrium. If the other firm applies then a firm, by also applying, reduces its penalty from \( D + F \) to \( D + (F/2) \) and earns noncollusive profit in either case. In other words, if the other firm is expected to “rat to the Feds,” then it is best to rat as well, since the “gig is up.” Next note that it is never an equilibrium for only one firm to apply for leniency. This leaves as the remaining possibility that neither applies for leniency, in which case either collusion persists or both firms end up paying a penalty of \( D + F \). Both not applying for leniency is an equilibrium when

\[
(1 - w)p\hat{\pi} + w(p\hat{\pi} - D - F) > \hat{\pi} - D,
\]

which can be rearranged to

\[
(1 - w)(\hat{\pi} - \hat{\pi} + D) > wF.
\]

If damages \((D)\) are sufficiently large relative to fines \((F)\) or if the incremental profit from collusion is sufficiently high \((\hat{\pi} - \hat{\pi})\), then it is an equilibrium for firms not to turn evidence. But when that condition doesn’t hold, the only solution is for both firms to apply for leniency, so there is a race to the courthouse. This occurs when the probability of successful prosecution without a firm acting as a witness is sufficiently high \((w)\) or when the government fine \((F)\) is sufficiently high. The power of the leniency program lies in taking advantage of each

<table>
<thead>
<tr>
<th>FIRM 2</th>
<th>Do Not Apply</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FIRM 1</strong></td>
<td><strong>Apply</strong></td>
</tr>
<tr>
<td><strong>Apply</strong></td>
<td>( \hat{\pi} - D - (F/2) ),</td>
</tr>
<tr>
<td><strong>Do Not Apply</strong></td>
<td>( \hat{\pi} - D - (F/2) ),</td>
</tr>
<tr>
<td></td>
<td>( \hat{\pi} - D - F, \hat{\pi} - D )</td>
</tr>
<tr>
<td></td>
<td>( (1 - w)p\hat{\pi} + w(p\hat{\pi} - D - F) ),</td>
</tr>
<tr>
<td></td>
<td>( (1 - w)p\hat{\pi} + w(p\hat{\pi} - D - F) )</td>
</tr>
</tbody>
</table>
firm’s fear that the other firm may apply for amnesty first. The recent record suggests that this dynamic has played a critical role in the successful prosecution of many price-fixing cartels, including those in vitamins, graphite electrodes, and fine arts auctions.

Summary

This chapter examined a variety of issues related to the behavior of firms in oligopolistic industries. An oligopoly is characterized by having a relatively small number of firms. The low number results in it being appropriate for each firm to take into account the actions and future responses of its competitors in deciding how much to produce or what price to set.

Though there are many models of oligopoly, this chapter focused upon the Cournot model, which specifies that firms make simultaneous output decisions. We chose the Cournot model for several reasons. First, it is widely used by industrial organization economists. Second, many of the qualitative results of the Cournot solution are intuitively plausible and consistent with some empirical evidence. Finally, a number of the most important results generated by the Cournot model are representative of results from many other oligopoly models. Though the Cournot model is idiosyncratic in specifying that firms choose quantity and not price, its results are quite general.

It was shown that the Cournot solution entails a price that exceeds the competitive price but falls short of the monopoly price. Firms jointly produce too much relative to the joint-profit maximum. While each firm is individually maximizing its profit at the Cournot solution, firms could raise all of their profits by jointly reducing output and moving price toward the monopoly level.

It is this lure of higher profits that provides firms with the desire to collude. The problem that firms face in colluding is that each can increase its current profit by deviating from the agreed-on output or price. To explain and understand collusive behavior in practice, an infinite horizon extension of the Cournot model was developed. It was found that collusion is consistent with each firm acting to maximize its sum of discounted profits. A firm is deterred from deviating from the collusive outcome by the threat that cheating will induce a breakdown in collusion. While cheating raises current profits, it lowers future profits by inducing greater competition in the future. Though there are many challenges to having an effective cartel—monitoring firm behavior to ensure that there is no cheating and coming to an agreement as to what price to set and how to allocate supply perhaps being the most important—many industries have succeeded in overcoming them.

With an understanding of how firms can collude, the chapter then turned to exploring antitrust law with respect to collusion or, as it is often called, price fixing. Although price fixing was made illegal with the Sherman Act (1890), the interpretation of this law took place only with key early cases like Addyston Pipe and Steel (1899) and Trenton Potteries (1927).
These cases established the per se rule with respect to price fixing. This rule says that price fixing is illegal regardless of the circumstances. There is no allowable defense.

Current law is such that to prove that firms are guilty of price fixing one needs a “smoking gun,” for example, a memo from a CEO to his competitor stating what the collusive price is to be. It is insufficient to show that the price decisions of firms are consistent with firms acting collusively. Although firms that collude without overtly communicating result in the same welfare losses as does a cartel that does overtly communicate, only the latter are typically prosecuted under Section 1 of the Sherman Act.

Of course, a well-developed law against price fixing is for naught if cartels are not discovered and properly punished. Recent enforcement policy has become more effective with the revision of the corporate leniency program and new sentencing guidelines. The former works to induce a firm to become a witness for the prosecution—thereby enhancing the prospects of a successful prosecution—and the latter permits the government to levy more severe penalties. In spite of the increase in these fines, private damages awarded to the plaintiffs in an antitrust case generally remain the most serious financial penalty imposed on a corporation due to price fixing.

Questions and Problems

1. In 1971 the federal government prohibited the advertising of cigarettes on television and radio. Can you explain why this ban on advertising might have raised the profits of cigarette manufacturers? Hint: Use the Advertising Game.

2. The inverse market demand for mineral water is \( P = 200 - 10Q \), where \( Q \) is total market output and \( P \) is the market price. Two firms, A and B, have complete control of the supply of mineral water and both have zero costs.

   a. Find the Cournot solution.
   
   b. Find an identical output for each firm that maximizes joint profits.

3. Continuing with problem 2, assume that each firm can choose only two outputs—the ones from parts a and b in question 2. Denote these outputs \( q_a \) and \( q_b \).

   a. Compute the payoff/profit matrix showing the four possible outcomes.
   
   b. Show that this game has the same basic properties as the Advertising Game. In particular, each firm’s optimal output is independent of what the other firm produces. Now consider firms playing an infinitely repeated version of this game and consider the following strategy for each firm: (i) produce \( q_a \) in period 1, (ii) produce \( q_b \) in period \( t \) if both firms produced \( q_b \) in all preceding periods, and (iii) produce \( q_a \) in period \( t \) if one or more firms did not produce \( q_b \) in some past period. Assume each firm acts to maximize its sum of discounted profits where the discount rate is \( r \).

   c. Find the values for \( r \) such that this strategy pair is a Nash equilibrium.
4. Consider a duopoly with firms that offer homogeneous products where each has constant marginal cost of \( c \). Let \( D(P) \) denote market demand. Firms make simultaneous price decisions. Letting \( p_1 \) and \( p_2 \) be the prices of firms 1 and 2, respectively, the demand function of firm 1 is specified to be
\[
D(p_1, p_2) = \begin{cases} 
D(p_1) & \text{if } p_1 < p_2 \\
\frac{D(p_1)}{2} & \text{if } p_1 = p_2 \\
0 & \text{if } p_1 > p_2.
\end{cases}
\]
If firm 1’s price is lower than firm 2’s price, then all consumers buy from it, so that its demand equals market demand. If both firms charge the same price, then they equally split market demand. If firm 1’s price is higher than firm 2’s price, then all consumers go to firm 2. Firm 2’s demand function is similarly defined. Each firm chooses prices to maximize its profit.
   a. Show that both firms pricing at marginal cost is a Nash equilibrium.
   b. Show that any other pair of prices is not a Nash equilibrium.
   Suppose that we limit firms to choosing price equal to \( c \), \( 2c \), or \( 3c \).
   c. Compute the payoff/profit matrix.
   d. Derive all of the Nash equilibrium price pairs.

5. In its rivalry with Westinghouse, General Electric instituted a “price protection” plan. This plan stated that if G.E. lowered its price, it would rebate the price difference to its past customers. Show that this plan makes collusion between G.E. and Westinghouse easier. Hint: G.E.’s gain from deviating is lower with the price protection plan in effect.

6. Assume an industry with two firms facing an inverse market demand of \( P = 100 - Q \). The product is homogeneous, and each firm has a cost function of \( 600 + 10q + 0.25q^2 \). Assume firms agree to equally share the market.
   a. Derive each firm’s demand curve.
   b. Find each firm’s preferred price when it faces the demand curve in part a. Now assume that firm 1’s cost function is instead \( 25q + 0.5q^2 \) while firm 2’s is as before.
   c. Find each firm’s preferred price when it faces the demand curve in part a.
   d. Compute each firm’s profit when firm 1’s preferred price is chosen. Do the same for firm 2’s preferred price. Which price do you think firms would be more likely to agree upon? Why?
   e. Show that neither price maximizes joint profits.
   f. Find the price that maximizes joint profits. Hint: It is where marginal revenue equals both firms’ marginal cost.
   g. Would firm 1 find the solution in part f attractive? If not, would a side payment from firm 2 to firm 1 of $500 make it attractive?

7. In the NCAA case, the Supreme Court held that the rule of reason was applicable even though horizontal price fixing was involved. Explain the rationale.

8. What are the benefits and costs of the per se rule?

9. What is the law toward parallel business behavior? Assume, for example, that three firms charge identical prices for a product and it is agreed by all observers that the price is unusually high
153 Oligopoly, Collusion, and Antitrust

compared to cost. Would this alone constitute a Sherman Act offense? Cite a relevant case to support your answer.

10. In March 2004, a class action suit was filed in Madison, Wisconsin, that accused twenty-four bars and the Madison-Dane County Tavern League of conspiring to fix prices on beer and liquor by agreeing to eliminate weekend happy hours. The suit, filed on behalf of three University of Wisconsin students, contends that the University of Wisconsin encouraged bars to collude as part of an antidrinking campaign. There is no evidence that the bar owners ever directly communicated. Do you think this is a violation of Section 1 of the Sherman Act? How would you go about calculating the damages incurred by the elimination of half-price drinks?

Appendix

Game Theory: Formal Definitions

The strategic (or normal) form of a game is defined by three elements: (1) the set of players, (2) the strategy sets of the players, and (3) the payoff functions of the players. The set of players comprises those individuals making decisions. Let \( n \) denote the number of players. The strategy of a player is a decision rule that prescribes how she should play over the course of the game. All of the strategizing by a player takes place with regard to her selection of a strategy. The strategy set of player \( i \), which we denote \( S_i \), comprises all feasible strategies. A player is constrained to choosing a strategy from her strategy set. The payoff function of player \( i \) gives player \( i \)'s utility (or payoff) as a function of all players' strategies. An \( n \)-tuple of strategies, one for each player, is referred to as a strategy profile.

Player \( i \)'s payoff function is denoted \( V_i(\cdot) \), where \( V_i(s_1, \ldots, s_n) \) is the payoff to player \( i \) when player 1's strategy is \( s_1 \), player 2's is \( s_2 \), \ldots, player \( n \)'s strategy is \( s_n \).

A strategy profile \( (s_1^*, \ldots, s_n^*) \) is a Nash equilibrium if and only if each player's strategy maximizes her payoff given the other players' strategies. Formally,

\[
V_i(s_1^*, \ldots, s_n^*) \geq V_i(s_1^*, \ldots, s_{i-1}^*, s_i, s_{i+1}^*, \ldots, s_n^*)
\]

for all \( s_i \) in \( S_i \) and for all \( i = 1, \ldots, n \).
6

Market Structure and Strategic Competition

There are two key sources of competition in markets—existing firms and potential entrants. Chapter 5 focused on the behavior of existing firms while taking their number as being exogenously determined. This chapter extends this analysis in two important ways. First, we consider the determinants of the number of sellers. The factors analyzed are scale economies and entry conditions. Second, we consider the role of potential competition—specifically, the effect that the threat of entry has on the price-cost margin.

In the section on market structure, our discussion begins with the problem of measuring the concentration of an industry. The analysis then turns to investigating how scale economies and entry conditions affect actual and potential competition. In the discussion of dominant firm theory, we begin an exploration of how established firms can influence the future competitiveness of the industry. Although entry conditions are treated as being exogenous to firms in the section on market structure, the section on strategic competition explores how established firms can affect entry conditions and thereby deter entry.

Market Structure

This section investigates two key elements of market structure—concentration and entry conditions. Our discussion of concentration is a continuation of our coverage of oligopoly theory and collusion in that its focus is on the role of actual competition. We then investigate the determinants of concentration by considering scale economies and entry conditions. Our discussion of entry conditions is concerned with understanding their role in determining the extent of actual competition (that is, the number of firms) and the extent of potential competition. The discussion of entry conditions in this section presumes that they are exogenously determined. Later in this chapter, we will investigate ways in which existing firms can influence entry decisions.

Concentration

In our analysis of oligopoly theory in the preceding chapter, we assumed that firms were identical—having the same products and same cost functions. As long as the industry is symmetric, a single number—the number of sellers—accurately measures market concentration. While the abstract world is often specified to be symmetric in order to reduce the complexity of the analysis, in the real world there is typically great heterogeneity among firms. The implication of firms having different products and cost functions is that firms have very different market shares. As a result, a simple count of the number of firms can be a very misleading measure of the degree of concentration.

One of the traditional tasks in industrial organization has been to develop a statistic that allows a single number to measure reasonably the concentration of an industry. In constructing a useful index of concentration, one first needs to understand the purpose that a
concentration index is to serve. From a welfare and antitrust perspective, a concentration index should measure the ability of firms to raise price above the competitive level. A higher value for a concentration index should indicate a higher price-cost margin or a higher likelihood of firms being able to collude successfully. Note that a concentration index is exclusively concerned with actual competition and ignores potential competition. For this reason, a concentration index cannot fully assess the competitiveness of a particular industry. Some other information will have to be provided so as to take account of the degree of potential competition.

**Definition of the Market**

To measure concentration, one must first define the limits of the market. Computation of firm market shares requires knowledge of the total sales of the market, which itself requires defining which products sold to which consumers constitute a well-defined market.

The issue of how markets should be defined from the viewpoint of economic theory has never been answered definitively. Economic theorists generally take the market as a given. However, when one engages in empirical work, it becomes necessary to make difficult judgments about what products and sellers constitute the market. For example, the outcome of one famous antitrust case hinged on whether the relevant market was cellophane or whether the correct market was “flexible wrapping materials” (that is, Saran, aluminum foil, brown paper, cellophane, and the like).

Most economists agree that the ideal market definition must take into account substitution possibilities in both consumption and production. George Stigler has expressed this point as follows:

An industry should embrace the maximum geographical area and the maximum variety of productive activities in which there is a strong long-run substitution. If buyers can shift on a large scale from product or area B to A, then the two should be combined. If producers can shift on a large scale from B to A, again they should be combined.

Economists usually state this in an alternative form: All products or enterprises with large long-run cross-elasticities of either supply or demand should be combined into a single industry.

A further difficulty on the supply side is the distinction between “substitution” and “new entry.” That is, where does the market stop and potential entry begin? Consider the airline industry. Should the market be, say, the “New York–Los Angeles market” or the entire United States? If the market is the New York–Los Angeles route, concentration would be relatively high, given this tight market definition. Entry would be “easy,” though, as airlines serving

Miami–San Francisco, for example, could easily switch. Alternatively, if the market is defined as the entire United States, concentration would be low but might include some airlines not well suited for the New York–Los Angeles route, for instance with aircraft designed for short hops and a small volume of traffic.

Of course, some definition must be followed. No harm is done if it is recognized that what is important is the competitive constraint on potential monopoly pricing. That is, one would get the same answer in analyzing pricing on the New York–Los Angeles route by viewing it either as a highly concentrated market with easy entry or as part of the unconcentrated United States market.

F. M. Scherer has offered the following proposal: “At the risk of being somewhat arbitrary, we should probably draw the line to include as substitutes on the production side only existing capacity that can be shifted in the short run, i.e., without significant new investment in plant, equipment, and worker training.”

A similar, though more specific, definition has been put forth by the Justice Department. It will be discussed in chapter 7, where we examine merger cases.

**Concentration Ratio**

Although economists have devised many indices to measure concentration, the most widely used measure is the concentration ratio. The $m$-firm concentration ratio is simply the share of total industry sales accounted for by the $m$ largest firms.

A fundamental problem with concentration ratios is that they describe only one point on the entire size distribution of sellers. Consider the size distributions of two hypothetical industries, as shown in table 6.1. Clearly, industries X and Y have the same four-firm concentration ratio, namely, 80 percent. However, noting the other information in table 6.1, most economists would regard the two industries as likely to exhibit quite different patterns of

<table>
<thead>
<tr>
<th>Firm</th>
<th>Industry X</th>
<th>Industry Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>60</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>85</td>
</tr>
</tbody>
</table>

competitive behavior. Suppose, now, that we calculate the three-firm ratios: Industry Y is now seen to be “more concentrated” (75 percent versus only 60 percent for Industry X). The basic problem is simply that this type of measure wastes relevant data. Nevertheless, the concentration ratio is superior to a simple count of sellers.

To pursue this point, consider the size distributions of the two industries shown as concentration curves in figure 6.1. The height of a concentration curve above any integer \( n \) on the horizontal axis measures the percentage of the industry’s total sales accounted for by the \( n \) largest firms. In general, the curves will rise from left to right, and at a continuously diminishing rate. In the limiting case of identical shares, such as in Industry X, the curve becomes a straight line. The curves reach their maximum height of 100 percent where \( n \) equals the total number of firms in the industry. If the curve of Industry Y is everywhere above the curve of X, then Y is more concentrated than X. However, when the curves intersect, as they do in figure 6.1, it is impossible to state which is the “more concentrated” industry unless we devise a new definition.

The most widely available concentration ratios are those compiled by the U.S. Bureau of the Census. Ideally, these ratios should refer to industries that are defined meaningfully from the viewpoint of economic theory. However, the census classifications of industries were developed over a period of years “to serve the general purposes of the census and other

![Figure 6.1](image-url)

**Figure 6.1**
Concentration Curves for Industries X and Y
government statistics” and were “not designed to establish categories necessarily denoting coherent or relevant markets in the true competitive sense, or to provide a basis for measuring market power.”4 The census frequently includes products that are not close substitutes, and it sometimes excludes products that are close substitutes. An example of the latter is the existence of two separate “industries” for beet sugar and cane sugar. The census ignores both regional markets (for example, all bakeries are combined into a single national market) and foreign competition (steel imports are excluded from the steel industry).

Because we will refer occasionally to studies that have used census concentration ratios, it should be helpful to provide a brief description of their procedure for classifying industries. Their classification system (known as the Standard Industrial Classification, or SIC) makes use of a series of numbers in which each succeeding digit represents a finer degree of classification.5 Thus, in the manufacturing sector of the economy, there are only 20 two-digit industries. An example is industry 20, the “Food and Kindred Products” industry. Within this two-digit industry there are nine three-digit industries, such as industry 201, the “Meat Products” industry. Within this three-digit industry there are three four-digit industries, such as industry 2015, “Poultry Dressing Plants.” The Census Bureau has computed concentration ratios for the top four, eight, and twenty firms for some 450 four-digit industries; these are the ratios most often used in statistical studies of industrial organization.

In table 6.2 we show four-firm concentration ratios for selected four-digit industries in 1992. The industries were chosen to cover a wide range of concentration ratios. It is immediately apparent that the theoretical market structures of perfect competition and monopoly do not provide useful categories for our real-world industries. Nor is it clear which industries should be classified as oligopolistic or competitive. Of course, most would agree that the industries at the top of table 6.2 are oligopolies, but how far down the list should we descend?

**HHI**

In 1992 the Antitrust Division of the Justice Department and the Federal Trade Commission issued new guidelines concerning their policy toward mergers.6 These guidelines are expressed in terms of the HHI (named for its inventors, O. C. Herfindahl and A. O. Hirschman). It is then important to define and discuss the HHI.

---


5. As a result of the North American Free Trade Agreement (NAFTA), a new industrial classification system is being put in place. The North American Industry Classification System (NAICS) is designed to aid in the exchange of statistics between the United States and its two North American trading partners, Canada and Mexico. Starting in 1999, the Census Bureau will issue statistics using the NAICS.

6. These guidelines are discussed in detail in chapter 7.
The HHI has the advantage of incorporating more information about the size distribution of sellers than the simple concentration ratio does. If we let $s_i$ denote firm $i$’s proportion of total industry sales (that is, its market share), then the HHI is defined as

$$\text{HHI} = \left(100s_1\right)^2 + \left(100s_2\right)^2 + \ldots + \left(100s_n\right)^2$$

where $n$ equals the number of firms. The HHI is the weighted average slope of the concentration curve (recall from figure 6.1 that industries with more steeply sloped curves are more concentrated). The weight for the slope of each segment of the curve is the corresponding $s_i$ for that segment.

If an industry consists of a single seller, then HHI attains its maximum value of 10,000. The index declines with increases in the number of firms and increases with rising inequality among a given number of firms. Hence, referring to our example in table 6.1, while assuming firms 6, 7, and 8 in Industry Y each have 5 percent of the market, we can calculate their HHIs:

Industry X: $\text{HHI} = 20^2 + 20^2 + 20^2 + 20^2 + 20^2 = 2,000$.
Industry Y: $\text{HHI} = 60^2 + 10^2 + 5^2 + 5^2 + 5^2 + 5^2 + 5^2 + 5^2 = 3,850$.

Thus the HHI would indicate that Industry X would be more likely to exhibit competitive behavior.

The Justice Department regards an HHI of 1,000 as critical. That is, if a merger leaves the HHI for the industry at 1,000 or less, the merger is unlikely to be challenged as violating antitrust laws. Note the HHIs in table 6.2.

<table>
<thead>
<tr>
<th>Industry</th>
<th>CR4</th>
<th>HHI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cigarettes</td>
<td>99</td>
<td>NA</td>
</tr>
<tr>
<td>Breweries</td>
<td>91</td>
<td>NA</td>
</tr>
<tr>
<td>Cereal Breakfast Foods</td>
<td>87</td>
<td>2,774</td>
</tr>
<tr>
<td>Turbines and Turbine Generators</td>
<td>79</td>
<td>2,404</td>
</tr>
<tr>
<td>Flat Glass</td>
<td>76</td>
<td>1,781</td>
</tr>
<tr>
<td>Gasoline Engine and Engine Parts</td>
<td>68</td>
<td>1,425</td>
</tr>
<tr>
<td>Farm Machinery and Equipment</td>
<td>53</td>
<td>1,707</td>
</tr>
<tr>
<td>Petroleum Refineries</td>
<td>41</td>
<td>625</td>
</tr>
<tr>
<td>Iron and Steel Mills</td>
<td>39</td>
<td>560</td>
</tr>
<tr>
<td>Pharmaceutical Preparations</td>
<td>37</td>
<td>489</td>
</tr>
<tr>
<td>Audio and Video Equipment</td>
<td>30</td>
<td>415</td>
</tr>
<tr>
<td>Women’s &amp; Girls’ Cut and Sew Dresses</td>
<td>14</td>
<td>111</td>
</tr>
<tr>
<td>Ready Mixed Concrete</td>
<td>6</td>
<td>29</td>
</tr>
</tbody>
</table>

Notes: CR4 = four-firm concentration ratio, HHI = Herfindahl-Hirschman Index.
One of the attractive features of the HHI is that it has foundations in oligopoly theory. Suppose that firms have homogeneous products and engage in Cournot competition. Let us allow firms to have different cost functions. Thus, \( c_i \) will denote the (constant) marginal cost of firm \( i \), where \( i = 1, \ldots, n \). One can show that the Cournot solution has a firm’s market share being negatively related to its marginal cost. The lower is firm \( i \)'s marginal cost, the higher is its profit-maximizing output and thus the higher is its share of the market. The important result is that the HHI is directly related to a weighted average of firms’ price-cost margins from the Cournot solution:

\[
\frac{s_1 (P^c - c_1)}{P^c} + \frac{s_2 (P^c - c_2)}{P^c} + \ldots + \frac{s_n (P^c - c_n)}{P^c} = \frac{HHI}{10,000 \cdot \eta}
\]

where \( P^c \) is the Cournot price, \( s_i \) is firm \( i \)'s market share, and \( \eta \) is the absolute value of the price elasticity of market demand. The higher is the HHI, the higher is the industry price-cost margin.\(^7\)

Using Concentration Indices in Antitrust Policy

Empirical evidence has shown that a high concentration index for an industry is signal of a high price-cost margin.\(^8\) Thus the relationship is true not only in theory but also in practice.

An important question is, What policy implications are to be drawn? In answering this question, one must have a theory for why concentration and price-cost margins (or profits) are positively related. Traditionally, there have been two main hypotheses. The \textit{collusion hypothesis} states that the more concentrated an industry is, the less competitive are firms and thus the higher the price-cost margin. This finding we established in the preceding chapter. For the Cournot solution, the smaller the number of firms (and thus the greater the concentration), the higher is the price-cost margin. In addition, collusion was found to be easier as the number of firms decreases. A reasonable policy implication from this theory is that one should break up highly concentrated industries.

In contrast to the collusion hypothesis, Harold Demsetz’s \textit{differential efficiency hypothesis} (or superior efficiency hypothesis) argues that high concentration does not cause a high price-cost margin. Instead, high concentration tends to be observed with high price-cost margins. The argument is as follows. In some industries there are apt to be a few firms that have a differential advantage over their competitors. This advantage could be due to lower cost or better products. In those industries these superior firms will tend to dominate the market—so that

---


concentration is high—and be able to price considerably above cost—so that the price-cost margin and industry profit are high. This argument underlies the relationship between the HHI and the industry price-cost margin for the Cournot solution. When firms have different costs, the result is skewed market shares and a high HHI. While the differential efficiency hypothesis is really a statement about firms—those firms with high market share will tend to have a high price-cost margin—it also implies that at the industry level, after aggregating individual firm data, one will tend to observe high industry concentration with high industry price-cost margins. According to the differential efficiency hypothesis, one does not want to go around breaking up highly concentrated industries. To do so would be to penalize firms for being superior and thereby deter them from doing what we want them to do—provide better products at a lower cost.

Note that the prediction of the differential efficiency hypothesis is that firms with high market shares will tend to have high price-cost margins and profits. An implication of this relationship is that industry profit will tend to be positively related to industry concentration. In contrast, the collusion hypothesis predicts that higher concentration causes higher price-cost margins and profits. A number of empirical tests of these two competing hypotheses have been performed. The empirical evidence strongly supports the differential efficiency hypothesis.9 The evidence shows that a firm’s profit is strongly positively associated with its market share. There is typically a weak positive association between industry profit and concentration.

Scale Economies

Perhaps the most important explanation of why some industries are more concentrated than others is the magnitude of economies of scale relative to total market demand. In other words, what fraction of the market’s output is needed by a firm to achieve minimum long-run average cost? For example, the automobile industry depends on large-scale production to achieve low unit costs, and there is “room” for only a small number of such large-scale firms in the market. In contrast, gasoline service stations do not gain cost advantages significantly beyond a certain volume, and comprise a relatively unconcentrated industry.

The specialization of labor and equipment that can be achieved as a result of larger size is an important source of economies of scale. The classic example is again automobile assembly. As the rate of output increases, workers can specialize more narrowly and become highly efficient in a number of tasks. Rather than install a complete engine, a worker might be responsible for attaching one small part. Similarly, if the output is sufficiently large,

specialized automatic screw machines rather than general-purpose lathes might be used in producing ball bearings.

Turning to the sources of diseconomies of scale, we find only one major explanation in the economics literature. This explanation is that as a firm increases in size, it becomes more and more difficult for the top management to exercise control over the entire organization. Thus this “control loss,” as the number of layers of management increases, is a possible source of diseconomies of scale.

However, according to some experts, this problem can be partially mitigated by a decentralized multidivisional form of corporate organization. That is, by delegating authority and responsibility to operating divisions (as General Motors did in creating Chevrolet, Buick, Oldsmobile, Pontiac, and so on), control loss can be offset.

While scale economies have been measured in various ways, many economists consider that the best way is to use engineering cost estimates, because engineers’ cost estimates usually embody assumptions quite similar to those underlying the long-run average cost curve of economic theory. An engineering study can hold fixed relative factor prices, product homogeneity, location, technology, volume, and so forth, thereby isolating the effect of increases in the rate of output on cost.

The results of one such study—which estimated both efficient plant and efficient firm sizes—are given in table 6.3. This 1975 study, conducted by Scherer, Beckenstein, Kaufer, and Murphy, obtained its estimates by interviews with technically qualified personnel working in the industries studied.\(^{10}\) To obtain the estimates for their twelve industries, they interviewed personnel in 125 companies in six countries.

If we take the figures in table 6.3 at face value, the second column (which gives “efficient firm” size as a percentage of the total market) indicates that scale economies are not particularly severe in most of the twelve industries. Only in the refrigerator and freezer industry is as much as 20 percent of the market necessary for efficiency. The third column indicates that actual market shares of the leading firms are generally considerably greater than necessary to attain efficient size. For example, the average market share of the four leading firms in the storage battery industry is about 15 percent, yet the “efficient firm” market share is only 2 percent.

Two other studies using the engineering approach, one by Bain\(^ {11}\) and the other by Pratten,\(^ {12}\) found similar results for different samples of industries. That is to say, in only a few industries in each study did the estimated efficient shares approximate the actual shares of the leading firms.


Table 6.3
Minimum Efficient Scale of Plants and Firms as Percentage of U.S. National Market, 1967

<table>
<thead>
<tr>
<th>Industry</th>
<th>Minimum Efficient Scale Plant as Percentage of Total Market</th>
<th>Minimum Efficient Scale Firm as Percentage of Total Market</th>
<th>Four-Firm Concentration Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beer Brewing</td>
<td>3.4</td>
<td>10–14</td>
<td>40</td>
</tr>
<tr>
<td>Cigarettes</td>
<td>6.6</td>
<td>6–12</td>
<td>81</td>
</tr>
<tr>
<td>Cotton Synthetic Fabrics</td>
<td>0.2</td>
<td>1</td>
<td>36</td>
</tr>
<tr>
<td>Paints, Varnishes, and Lacquers</td>
<td>1.4</td>
<td>1.4</td>
<td>22</td>
</tr>
<tr>
<td>Petroleum Refining</td>
<td>1.9</td>
<td>4–6</td>
<td>33</td>
</tr>
<tr>
<td>Shoes, Except Rubber</td>
<td>0.2</td>
<td>1</td>
<td>26</td>
</tr>
<tr>
<td>Glass Containers</td>
<td>1.5</td>
<td>4–6</td>
<td>60</td>
</tr>
<tr>
<td>Cement</td>
<td>1.7</td>
<td>2</td>
<td>29</td>
</tr>
<tr>
<td>Steel Works</td>
<td>2.6</td>
<td>3</td>
<td>48</td>
</tr>
<tr>
<td>Ball and Roller Bearings</td>
<td>1.4</td>
<td>4–7</td>
<td>54</td>
</tr>
<tr>
<td>Refrigerators and Freezers</td>
<td>14.1</td>
<td>14–20</td>
<td>73</td>
</tr>
<tr>
<td>Storage Batteries</td>
<td>1.9</td>
<td>2</td>
<td>61</td>
</tr>
</tbody>
</table>


Needless to say, the validity of such studies of economies of scale is not universally accepted. John McGee, for example, argues that there are some serious problems:

Such estimates cannot be forward-looking and are stale when done. They must implicitly embody some unspecified but homogeneous quality of production management, organization, and control, and must assume that some unspecified but given quality of overall management is imagined both to choose and to use the hypothetical physical plant that someone constructed on paper. Also, business problems are not solely engineering problems—which partly explains why not all successful businesses are run by practicing engineers.13

An extensive discussion of the various techniques used to estimate economies of scale—such as engineering cost estimates or econometric analysis of accounting data—is contained in a book by Scherer.14 Unfortunately, space limitations prevent further discussion of this important topic here.

**Entry Conditions**

Thus far in our analysis, we have measured the competitiveness of an industry by the number of firms or some other measure of concentration. An equally important factor, however, is the ease with which entry can take place. Entry conditions are important for two reasons. First, the number of active firms is partially determined by the cost of entry as well as other

factors such as economies of scale. Thus, entry conditions play an important role in determining concentration. Second, entry conditions determine the extent of potential competition. It is generally believed that a credible threat of entry will induce active firms to compete vigorously. If they do not, so that the industry has a high price-cost margin, entry will take place and drive price down. According to this argument, entry conditions are then important because the cost or difficulty of entry affects the effectiveness of potential competition. Later in this chapter we will examine the proposition that when entry is relatively costless, active firms must compete vigorously.

Defining the relevant set of entry conditions has proved to be a difficult and controversial subject in industrial organization. Nevertheless, here are some questions one needs to ask in order to assess entry conditions: How many prospective firms have the ability to enter in a reasonable length of time? How long does it take to enter this industry? How costly is entry? Will a new firm be at a disadvantage vis-à-vis established firms? Does a new firm have access to the same technology, the same products, the same information? Is it costly to exit the industry? You might wonder why the last question relates to entry conditions. Since an entrant is uncertain as to whether it will succeed, the cost of exit can be an important factor in the original decision to enter.

**Equilibrium under Free Entry**

Entry into an industry means acquiring the ability to produce and sell a product. In almost every industry, there is some cost to entry. This cost may represent investment in a production facility. If entry requires a license (as it does in many professional occupations), then the cost of this license contributes to the cost of entry. Entry into a consumer market often entails extensive advertising of one’s product in order to introduce it to the market. Or it might mean the issuance of free samples. As an initial step in considering entry conditions, we begin by investigating the relationship between the cost of entry and the number of competing firms.

Consider an industry in which all active and prospective firms have access to the same production technology and input prices so that each firm has the same cost function. Further assume that all firms produce the same product. Let \( \pi(n) \) denote each firm’s profit per period when there are \( n \) active firms in the industry. For example, if we model active firms as simultaneously choosing output, then the Cournot solution applies, so that \( \pi(n) = P^*q^* - C(q^*) \), where \( q^* \) is the Cournot firm output (where a higher \( n \) implies a lower \( q^* \)), \( P^* \) is the Cournot price, and \( C(q^*) \) is each firm’s cost of producing \( q^* \) (not including the cost of entry). We will assume that firm profit, \( \pi(n) \), is decreasing in the number of firms, \( n \). This assumption makes sense because the presence of more firms generally means a more competitive environment exists.

Assume that active firms operate in this industry forever and \( r \) is each firm’s discount rate. If the industry has reached an equilibrium and it entails \( n \) active firms, then each active firm’s
The sum of discounted future profits is \( \pi(n)/r \). Using the condition that firm profit is decreasing in the number of firms, figure 6.2 plots the present value of a firm (before netting out any cost of entry).

Suppose that this is a new industry and prospective firms simultaneously decide whether or not to enter. A free-entry equilibrium is defined by a number of entrants, denoted \( n^e \), such that entry is profitable for each of the \( n^e \) entrants and entry would be unprofitable for each of the potential entrants who chose not to enter. If \( K \) denotes the cost of entry, the free-entry equilibrium number of firms is defined by

\[
\frac{\pi(n^e)}{r} - K > 0 > \frac{\pi(n^e + 1)}{r} - K.
\]  

(6.1)

For the case in figure 6.2, \( n^e \) equals five.

Suppose condition 6.1 does not hold; in particular, consider the first inequality not holding. In figure 6.2 under this supposition six or more firms would decide to enter. Because \( [\pi(n)/r] - K < 0 \) when \( n > 5 \), this could not be a free-entry equilibrium. Each of the prospective firms...
expects to have a negative present value from entering the industry. One or more of these entrants would prefer not to enter. Thus, six or more active firms is too many. Alternatively, suppose that the second inequality in (6.1) does not hold; that is, \( n \) firms plan to enter and \([\pi(n + 1)/r] - K > 0\) (for figure 6.2, \( n < 5 \)). In that case, entry of an additional firm is profitable, so that we would expect additional entry. Hence this is not a free-entry equilibrium either. Only when both inequalities in (6.1) are satisfied is an equilibrium achieved. In that case, entrants have a positive present value and nonentrants would have a negative present value if they entered.\(^{15}\)

The relationship between the cost of entry and the number of active firms at a free-entry equilibrium is quite straightforward. If the cost of entry rises, fewer entrants find entry profitable. As shown in figure 6.2, if the cost of entry is higher, say, at \( K^0 \), there would be less entry, as only three firms would enter.

This model of entry is useful in revealing how entry conditions influence the number of competing firms. However, it is unsatisfactory in that it ignores several important factors relevant to the role of entry. This model does not allow for asymmetries between firms, and it is these asymmetries that explain the observed inequality in market shares. In particular, asymmetries between existing firms and potential entrants are ignored in that this model only examines initial entry into an industry. In fact, the concern of antitrust analysis is largely with the entry conditions faced by potential entrants for a currently active industry. Of key importance is the effectiveness of the threat of entry in keeping price close to cost. The remainder of this chapter will focus on disadvantages that potential entrants might face compared to existing firms and, in particular, the ability of existing firms to create such disadvantages.

Prior to moving on in our analysis, we can at least touch upon this issue with the use of figure 6.2. Suppose that, as shown in figure 6.2, the cost of entry is \( K \), so that five firms enter. Further suppose that, after some number of periods, an unanticipated and permanent shift outward of the market demand curve occurs. This rise in demand causes each firm’s profit to rise so that the present value schedule shifts up (see figure 6.2). As a result, each active firm is earning profit of \( \hat{\pi}(5) \) rather than \( \pi(5) \). If the cost of entry is still \( K \), this rise in demand will induce two additional firms to enter. However, for reasons that will be considered shortly, the entry cost of a new firm may be higher than that for earlier entrants. If instead it costs \( K^0 \) for a firm to enter the industry today, additional entry will not occur in response to the rise in demand. As a result, the price-cost margin is higher but entry does not occur to stay its rise.

\(^{15}\) This theory ignores the difficult coordination problem faced by firms simultaneously making entry decisions. If all potential entrants are identical, how do they decide which \( n \) should enter? In fact, there is another free-entry equilibrium for this model in which each potential entrant enters with some probability. In that case, there could be too little entry or too much entry.
Barriers to Entry

The traditional wisdom in industrial organization is that serious and persistent monopolistic deviations of price from cost are likely only when two conditions coexist: sufficiently high seller concentration to permit (collusive) pricing, and high barriers to entry of new competition.  

We have reviewed collusive pricing, but what are barriers to entry? There is perhaps no subject that has created more controversy among industrial organization economists than that of barriers to entry. At one extreme, some economists argue that the only real barriers are government related. Examples include a franchise given by government to a local cable television company and the requirement that to operate a New York City taxicab one must own a government-issued medallion. A patent is another example in that it gives a firm a twenty-year monopoly. At the other end of the spectrum, some economists argue that almost any large expenditure necessary to start up a business is a barrier to entry. Given this state of affairs, we cannot hope to provide a definitive answer. Our objective will be to discuss the various views and definitions and try to evaluate each of them.

A pioneer in this area (and a source of much of the controversy), Joe Bain, defined a barrier to entry as “the extent to which, in the long run, established firms can elevate their selling prices above minimal average costs of production and distribution . . . without inducing potential entrants to enter the industry.” One immediate problem with this definition is that it is a tautology: a barrier to entry is said to exist if existing firms earn above-normal profit without inducing entry. In other words, Bain defines a barrier to entry in terms of its outcome.

One gets a better idea of what Bain has in mind when he states what he considers to be barriers to entry. These include scale economies, the capital cost requirements of entry, government restrictions like tariffs and patents, and absolute cost advantages of existing firms. Sources of the latter include a better technology (protected through patents or trade secrets), control of low-cost raw material supplies, and the learning curve. The learning curve refers to the idea that a firm with greater experience (typically measured by total past output) has discovered more ways of improving its production process.

These barriers are quite diverse and certainly entail very different welfare implications. A government restriction like a tariff is typically welfare-reducing. It is then a “bad” barrier to entry. In contrast, superior efficiency of existing firms due to a better technology is a “good” barrier to entry. No reasonable economist believes that society is better off if existing firms are made less efficient. However, at least in the short run, welfare would be higher if existing firms were forced to share their know-how with new entrants. The important point to

---

make here is that a barrier to entry, as defined by Bain, need not imply that its removal would raise welfare.\(^\text{18}\)

A very different definition was put forth by Nobel laureate George Stigler: “a barrier to entry may be defined as a cost of producing (at some or every rate of output) which must be borne by firms which seek to enter an industry but is not borne by firms already in the industry.”\(^\text{19}\) The emphasis of this definition is on differential costs between existing firms and entrants. For example, suppose that later entrants have to advertise their product to consumers while existing firms do not. This cost of advertising is a barrier to entry according to Stigler’s definition (and also Bain’s). We believe it is correct to say that Stigler’s definition is narrower than Bain’s. That is, some things are barriers according to Bain but not according to Stigler (for example, scale economies), although the reverse is not true.\(^\text{20}\)

A third definition comes from Christian von Weizsäcker: “Barriers to entry into a market . . . can be defined to be socially undesirable limitations to entry of resources which are due to protection of resource owners already in the market.”\(^\text{21}\) This definition certainly is the best motivated. However, like Bain, it defines a barrier to entry by a particular outcome. Ideally, we would like a definition to point out those specific factors in industries that are reducing social welfare. Whether one can construct a welfare-based operational definition that antitrust economists and lawyers could then use is an open question.

In the remainder of this section, we want to focus on the controversy related to the definition of entry barriers. The discussion has largely revolved around what Bain has labeled barriers to entry. For example, a large amount of capital necessary for entry is often cited as a source of cost disadvantage faced by new entrants. Richard Posner strongly disagrees with this position:

Suppose that it costs $10,000,000 to build the smallest efficient plant to serve some market; then, it was argued, there is a $10,000,000 “barrier to entry.” a hurdle a new entrant would have to overcome to serve the market at no disadvantage vis-à-vis existing firms. But is there really a hurdle? If the $10,000,000 plant has a useful life of, for example, ten years, the annual cost to the new entrant is only

---

18. Although one can remove government restrictions, how can one remove scale economies? Scale economies are due to existing technological know-how, and knowledge cannot be removed (unless one lives in the world created by George Orwell in *Nineteen Eighty-Four*). However, one can talk about breaking up firms so that each firm takes less advantage of economies of scale.
20. For a discussion of entry barriers, see Harold Demsetz, “Barriers to Entry,” *American Economic Review* 72 (March 1982): 47–57. Demsetz criticizes the definitions of Bain and Stigler for focusing solely on the differential opportunities of existing firms and potential entrants. He argues that this approach ignores legal barriers to entry, such as the requirement that one must have a license in order to operate but where licenses are traded freely (so that the incumbent firm and the potential entrant have the same opportunity costs). It is a matter of interpretation whether Bain ignores such barriers.
$1,000,000. Existing firms bear the same annual cost, assuming that they plan to replace their plants. The new entrant, therefore, is not at any cost disadvantage at all.\(^{22}\)

Posner does agree with a somewhat more subtle view of the capital requirements barrier. This is the view that the uncertainty of a new entrant’s prospects may force the entrant to pay a higher risk premium to borrow funds than existing firms must pay. Others have observed that when truly huge amounts of capital are required, the number of possible entrants who can qualify is greatly reduced. And, although this risk premium may not bar entry, it could delay it.

Perhaps the most controversial entry barrier is that of scale economies. To get a feel for this controversy, we have constructed a fictitious conversation between Joe Bain and George Stigler.\(^{23}\) The conversation concerns a market for which the firm average cost function is as shown in figure 6.3. Average cost is declining until an output of \(\hat{q}\) is reached, after which average cost is constant. Here \(\hat{q}\) is the minimum efficient (or optimal) scale. Suppose that


\(^{23}\) It is important to emphasize that this conversation is not based on any stated opinions of Bain and Stigler but on the current authors’ interpretation of their writings.
there is a single firm in the industry and it is producing $q^0$ and pricing at $P^0$. Note that its price exceeds average cost.

**Joe:** As is apparent from figure 6.3, scale economies are a barrier to entry. The existing firm is pricing above cost, yet entry does not occur, as it would be unprofitable.

**George:** But why do you say that entry is unprofitable?

**Joe:** The reason is quite obvious. If a new firm comes in and produces at minimum efficient scale, the total industry supply would be $\hat{q} + q^0$. Because price falls below average cost, entry is unprofitable. Of course, a new firm could instead produce at a low rate and thereby reduce the extent by which price is depressed. However, because average cost is declining, the new firm would be at a considerable cost disadvantage and once again incurs losses. In either case, entry is unprofitable.

**George:** Why do you assume that the new firm expects the existing firm to maintain its output? Why can’t the new firm enter and slightly undercut the existing firm’s price of $P^0$? It would then get all of market demand and earn profit approximately equal to $[P^0 - AC(q^0)]q^0$. Entry is profitable!

**Joe:** Don’t be silly, George. Do you really believe that a new firm could induce all consumers to switch to buying its product by setting a slightly lower price?

**George:** Why not?

**Joe:** There are lots of reasons. For example, it has been found empirically that consumers are hesitant to switch brands. Such brand loyalty makes sense as consumers have a lot less information about a new brand’s quality, since they lack personal experience. In order to offset brand loyalty, a new firm would have to offer a considerable price discount in order to lure consumers away.

**George:** Joe, we finally agree on something. What you’ve done is pointed out the real barrier to entry—that consumers have a preference for the existing firm’s product. In order to overcome brand loyalty, a new firm must initially sell at a discount or perhaps even give away the product in the form of free samples. The cost of getting consumers familiar with its product is a measure of the true barrier to entry, and it meets my definition. Scale economies are just a red herring.

Well, someone had to have the last word, and we let it be the economist with the Nobel Prize. In any event, we hope this dialogue gave you a taste of the controversy and difference of opinions related to scale economies as a barrier to entry.

Having mentioned brand loyalty as a barrier to entry, let us point out its relevance in the ReaLemon case. The ReaLemon brand of reconstituted lemon juice commanded a premium price over its rivals. The apparent reason was that consumers had greater experience with ReaLemon and were unwilling to experiment with unknown brands (without large price
discounts). Several policies for eliminating this barrier are conceivable, but consider the one proposed by a Federal Trade Commission official: “For competition to enter the processed lemon juice industry, the barriers to entry which inhere in the ReaLemon trademark must be eliminated. As a consequence... the only effective relief under the facts shown by the record in this case requires the licensing of the ReaLemon brand name to others wishing to enter the production, marketing and sale of processed reconstituted lemon juice.”24 The important question here is whether consumers will truly gain as a result of trademark licensing. Who will ensure that all licensees selling ReaLemon will maintain a high-quality product? In brief, the ReaLemon trademark has served to convey information to consumers, and licensing to remove the barrier to entry probably would raise information costs.

If you were paying attention, you should be quite confused about entry barriers. Join the crowd! The concept of barriers to entry lacks clarity, and one is never sure what to do with it. It is certainly not clear what are the welfare implications of any particular thing called a barrier to entry. The most unfortunate part is that some economists and antitrust lawyers throw the term entry barrier around as if there were one accepted and meaningful definition when there is not. The best advice we can offer is to perform a two-stage inquiry. In the first stage, carefully examine the assumptions underlying the particular argument that something is a barrier. Determine whether it is indeed true that existing firms can maintain price above cost while deterring entry. In the second stage, consider whether there is a policy that could “remove” the barrier and improve social welfare.

**Contestability and Sunk Costs**

A different perspective on entry conditions is provided by the theory of contestable markets due to William Baumol, John Panzar, and Robert Willig.25 A market is *perfectly contestable* if three conditions are satisfied. First, new firms face no disadvantage vis-à-vis existing firms. This condition means that new firms have access to the same production technology, input prices, products, and information about demand. Second, there are zero *sunk costs*; that is, all costs associated with entry are fully recoverable. A new firm can then costlessly exit the industry. If entry requires construction of a production facility at cost $K$, then sunk costs are zero if, on exiting the industry, a firm can sell it for $K$ (less any amount due to physical depreciation). If there is no market for such a facility and it must be sold for scrap at price $R$, then sunk costs equal $K - R$. The third condition is that the entry lag (which equals the time between when a firm’s entry into the industry is known by existing firms and when the new


firm is able to supply the market) is less than the price adjustment lag for existing firms (the time between when it desires to change price and when it can change price).

The central result is that if a market is perfectly contestable, then an equilibrium must entail a socially efficient outcome. For example, suppose that there are scale economies as in figure 6.3 and a firm prices at \( P^0 \), which exceeds \( P^* \), the price associated with average cost pricing.\(^{26}\) If the market is perfectly contestable, then a new firm could enter, undercut the price of \( P^0 \) by a small amount, and earn profit of approximately \( (P^0 - AC(q^0))q^0 \). Implicit in it earning that profit is that its product and cost function are the same as the incumbent firm’s and the incumbent firm is unable to adjust price prior to the new firm selling. Furthermore, when the incumbent firm eventually does respond by adjusting its price, the new firm is assured of being able to costlessly exit the industry with its above-normal returns intact because there are zero sunk costs. This type of hit-and-run entry will occur unless the incumbent firm prices at \( P^* \). Contestability is a theory for which potential competition plays the dominant role in generating competitive behavior.

If there are positive sunk costs, then a new firm cannot costlessly exit. If the exit cost is sufficiently large, then it may swamp the above-normal profit earned before the incumbent firm responds. However, if sunk costs are close to zero, the price that the incumbent firm charges must be close to \( P^* \). Baumol, Panzar, and Willig then interpret sunk cost as a barrier to entry.

This definition has close ties to Stigler’s. It is also related to Posner’s view about risk premiums being higher for potential entrants. The more expenditures that are sunk, the higher you would expect a new firm’s cost of capital (due to the risk premium). Because there is a chance that this new firm will fail, higher sunk costs mean more costs that cannot be recovered on exit. For example, the cost of advertising one’s product: Who would pay much of anything for the trademark of a failed product?

In concluding, we should note that the theory of contestable markets is quite controversial. One reason is that the theory is not robust to small changes in some of the assumptions. In particular, if the entry lag exceeds the price adjustment lag (which seems very plausible), then price could be considerably above \( P^* \).\(^{27}\) Although the theory is robust to the amount of sunk costs, it appears that most industries have considerable sunk costs. The relevance of the theory is then put into question. However, if nothing else, contestability has been instrumental in causing antitrust analyses to reduce their emphasis on concentration and take proper account of potential competition.

26. The socially efficient solution referred to here is defined as the social welfare optimum, subject to the constraint that firms earn at least normal profits. This solution entails a single firm pricing at average cost and meeting all demand. This analysis, however, does ignore the use of nonlinear pricing schemes like two-part tariffs. For details see chapter 11.

Dominant Firm Theory

Though the theory of monopoly is given a prominent place in textbooks, in fact there have been very few monopolies. There have been, however, many industries in which one firm was dominant. At the beginning of the twentieth century, U.S. Steel (now USX) commanded 65 percent of the market and was more than ten times the size of its largest competitor. More recent examples include IBM and AT&T. IBM has long been the dominant firm in the mainframe computer market, where its market share exceeded 60 percent in the 1980s. Years after entry was allowed into the long-distance communications market, AT&T remains the dominant player, with a market share around 60 percent (see chapter 15 for details).

Static Analysis

The standard dominant firm model assumes that there is one big firm and a large number of small price-taking firms, typically referred to as the “competitive fringe.” Because of its position, the dominant firm is modeled as selecting a price that the fringe firms take as given in deciding how much to supply. Given that the fringe comprises price takers, their profit-maximizing output can be represented by a supply function, $S(P)$.

Being strategic, the dominant firm knows that when it prices at $P$, the fringe will supply $S(P)$ to the market. The dominant firm’s demand is specified to be what is left over of market demand after fringe supply is sold. Denoted $D_d(P)$, the dominant firm’s demand function is then defined by

$$D_d(P) = D(P) - S(P)$$

where $D(P)$ is the market demand function. Figure 6.4 depicts the dominant firm’s demand function along with the resulting marginal revenue curve $MR_d$. Because the fringe supply curve lies above the market demand curve when price exceeds $\hat{P}$, the fringe supplies all of demand when the dominant firm prices above $\hat{P}$. Therefore, $D_d(P) = 0$ when $P > \hat{P}$; the dominant firm prices itself out of the market. On the other hand, the dominant firm prices the fringe out of the market when its price is less than $P^0$, as $S(P) = 0$ for all $P < P^0$. In that case, the market demand curve and the dominant firm demand curve coincide. For prices between $P^0$ and $\hat{P}$, the market is supplied by both the dominant firm and fringe firms.

If the dominant firm’s (constant) unit cost is $c_d$, then its profit is

$$\pi_d = (P - c_d)D_d(P) = (P - c_d)[D(P) - S(P)].$$

28. Because each fringe firm is small, its output decision does not affect price. Hence the additional revenue from one more unit is just price, so that marginal revenue equals the market price for a fringe firm. Hence, a fringe firm chooses its output $q$ so as to equate its marginal cost $MC(q)$ to price. It follows that an individual fringe firm’s supply function $s(P)$ is defined as that quantity which equates a fringe firm’s marginal cost with price. If there are $n$ fringe firms (where $n$ is large), the fringe supply function is then $S(P) = ns(P)$. 
The dominant firm maximizes its profit by equating its marginal revenue $MR_d$ with its marginal cost $c_d$. The intersection, shown in figure 6.4, occurs at a dominant firm output of $q^*$. The price that the dominant firm charges is $P^*$, where it is read off its demand curve: 
$q^*_d = D_d(P^*)$. The dominant firm solution entails a price of $P^*$, with the fringe supplying $S(P^*)$ and the dominant firm supplying the residual of $[D(P^*) - S(P^*)]$ or $q^*$.

An important insight provided by the dominant firm model is in showing how the existence of a competitive fringe restrains the dominant firm’s pricing behavior. Suppose the fringe was absent, so that the dominant firm was instead a monopolist. In that case its demand would equal market demand. As shown in figure 6.4, its marginal revenue curve is $MR$, so that its profit-maximizing output is $q^m$. Price would be $P^m$, which exceeds $P^*$. The existence of a competitive fringe results in the dominant firm charging a lower price.

The dominant firm sets a lower price because its demand is weaker as a result of the fringe. Furthermore, it takes into account how the fringe will respond to its price. Knowing that fringe supply is increasing in its price, the dominant firm sets a lower price in order to reduce...
fringe supply. By the same logic, one can show that if the fringe were to become larger (that is, if its supply curve shifted out), the dominant firm would set a price even lower than $P^*$. Modeling AT&T as the dominant firm in the long-distance telephone market (with MCI, Sprint, and other providers as fringe firms), a recent study used the dominant firm model to estimate the market power of AT&T during the mid-1980s and early 1990s.\footnote{Simran K. Kahai, David L. Kaserman, and John W. Mayo, “Is the ‘Dominant Firm’ Dominant? An Empirical Analysis of AT&T’s Market Power,” \textit{Journal of Law and Economics} 39 (October 1996): 499–517.} Generally speaking, the market power of a firm is the ability to raise price without suffering a significant decline in demand. This is typically measured by the price elasticity of the firm’s demand curve, which is the percentage change in the firm’s demand divided by the percentage change in the firm’s price. A more elastic demand curve means that a firm’s demand is more responsive to its price so that, for example, its demand declines more in response to a given rise in price. So as to avoid the low level of demand associated with high prices, a firm with a relatively price-elastic demand curve will tend to set price relatively close to marginal cost.

Let $\eta_{\text{ATT}}$ denote the absolute value of the price elasticity of AT&T’s demand curve. In estimating this elasticity, the authors of the study first showed that

$$
\eta_{\text{ATT}} = \frac{\eta + (1 - s_{\text{ATT}}) \varepsilon_f}{s_{\text{ATT}}}
$$

where $s_{\text{ATT}}$ is AT&T’s market share, $\eta$ is the absolute value of the price elasticity of the market demand curve, and $\varepsilon_f$ is the price elasticity of the fringe firms’ supply curve. There are several properties of this relationship worth noting. First, a more elastic market demand curve (as represented by a higher value for $\eta$) results in a more elastic demand curve for AT&T. Second, a more elastic fringe supply curve (as represented by a higher value for $\varepsilon_f$) results in AT&T’s demand curve being more elastic. This second property is important. When AT&T raises its rates, whether its demand falls a lot or a little partially depends on whether consumers are willing to go to other providers of long-distance service (which depends on how substitutable are the services of AT&T and these other providers) and whether these other providers are able and willing to meet a significant increase in the demand for their services. If they are indeed able and willing and if consumers find their services to be a good substitute, then $\varepsilon_f$ should be high, so that the fringe supply is quite responsive to AT&T’s rates. In that case, AT&T can expect a significant fall in demand from raising its rates. This is reflected in the equation for elasticity in that a higher value for $\epsilon_f$ results in a higher value for $\eta_{\text{ATT}}$. A third property is that the closer AT&T is to being a monopolist (in other words, the closer $s_{\text{ATT}}$ is to a value of one), the less important is the fringe supply elasticity and the more important is the market demand elasticity in determining the elasticity of AT&T’s demand curve. If it
is a monopolist \((s_{\text{ATT}} = 1)\), then \(\eta_{\text{ATT}} = \eta\), as the market demand curve is AT&T’s demand curve.

Using data from 1984–1993, this study estimated \(\eta\) to be 0.49, which means that a 1 percent increase in AT&T’s long-distance telephone rates results in a fall in market demand of around one-half of 1 percent (recall that \(\eta\) is the *absolute* value of the elasticity). Consumer’s demand for long-distance telephone service is then relatively unresponsive to the price of service. The estimate of \(\varepsilon_f\) is 4.38, which means that a 1 percent increase in AT&T’s rates brings forth a 4.38 percent increase in the supply of other providers of long-distance telephone service. Fringe firms are then quite able and willing to increase supply in response to higher rates set by AT&T. With AT&T having about 60 percent of the market \((s_{\text{ATT}} = 0.6)\) and using the formula

\[
\eta_{\text{ATT}} = \frac{\eta + (1 - s_{\text{ATT}})\varepsilon_f}{s_{\text{ATT}}}
\]

the price elasticity of AT&T’s demand curve is then estimated to be

\[
\frac{0.49 + (1 - 0.6)4.38}{0.6} = 3.74.
\]

That is, AT&T could expect its demand to drop by 3.74 percent in response to a rise in its rates of 1 percent. This result is very price-elastic and suggests that, in spite of its large market share, AT&T does not have much market power. If it tried to set relatively high rates, it would suffer an appreciable drop in the demand for its service. With this estimate of its demand elasticity, AT&T’s price-cost margin is estimated to be around 0.27.\(^{30}\) By comparison, various studies find the average price-cost margin in U.S. industries to lie between 0.3 and 0.6. Relative to other industries, it appears that AT&T has below-average market power.

**Dynamic Analysis: Limit Pricing**

One of the more interesting features of dominant firm industries is the manner in which they evolve. In the case of U.S. Steel, its dominance deteriorated over time. It had 65 percent of the market in 1901, but its market share was down to 42 percent by 1925 and just 24 percent by 1967. In contrast, Alcoa maintained a clearly dominant position in the aluminum industry from 1893 to 1940. Similarly, IBM has been able to maintain its dominance in mainframe

30. A profit-maximizing AT&T will set its price-cost margin so that \(\frac{P_{\text{ATT}} - MC_{\text{ATT}}}{P_{\text{ATT}}} = \frac{1}{\eta_{\text{ATT}}}\), where \(P_{\text{ATT}}\) is its price and \(MC_{\text{ATT}}\) is its marginal cost. For a proof of this relationship, see note 7 in chapter 5. The estimate of its price-cost margin is then \(\frac{1}{3.74} = 0.27\).
computers, as has AT&T in long-distance telecommunications. Why is it that some firms remain dominant and others do not?

To explore this and related issues, dynamic versions of the dominant firm model have been developed. The pioneering work in this area was by Darius Gaskins, though we will follow the contribution of Kenneth Judd and Bruce Petersen. The new factor that the dynamic model brings in is the way the size of the fringe changes over time. Of central importance is how the growth of the fringe depends on the pricing behavior of the dominant firm.

Rather than start with a supply function for the fringe, let us begin by specifying the cost function for a fringe firm and derive its supply that is consistent with the maximization of a fringe firm’s profits. Assume that to produce one unit of output, a fringe firm must spend \( c_f \) on variable inputs and use one unit of capacity. If we denote \( x(t) \) as the total capacity of the fringe at time \( t \), the maximum amount of output that the fringe can produce at \( t \) is then \( x(t) \).

We will assume that the fringe’s initial capacity is relatively low and that \( c_f \geq c_d \), so that the dominant firm is at least as efficient as the fringe firms.

Given the dominant firm prices at \( P(t) \) at time \( t \), fringe profit at \( t \) from producing \( q \) units is \( [P(t) - c_f]q \). When \( P(t) > c_f \), fringe profit is higher the more it produces. In that case, each fringe firm maximizes its profits by producing at capacity. Thus, fringe supply is \( x(t) \) when \( P(t) > c_f \). When \( P(t) < c_f \), then a fringe firm prefers to produce zero. It follows that the fringe supply function is

\[
S(P(t)) = \begin{cases} 
  x(t) & \text{if } P(t) \geq c_f \\
  0 & \text{if } P(t) < c_f.
\end{cases}
\]

In order to be able to expand future production above existing capacity, fringe firms must invest in new capacity. Capacity is assumed to live forever and cost \( Z \) per unit. The key assumption in this analysis is how fringe firms finance investment in capacity. It is assumed that investment is financed solely from retained earnings. Retained earnings are that part of profit that is not given out as dividends. Suppose \( P(t) \geq c_f \). Using the fringe supply function, fringe profit at \( t \) is then \( [P(t) - c_f]x(t) \). Let \( u(t) \) denote a fringe firm’s retention ratio. It is the proportion of current profit retained by the firm. Assume that all retained earnings are invested in capacity. The growth in the fringe’s capacity at \( t \), denoted \( \Delta x(t) \), is then

\[
\Delta x(t) = [P(t) - c_f]x(t)u(t) \frac{1}{Z} \quad \text{if } P(t) \geq c_f.
\]  

In words, the fringe has profit of \( [P(t) - c_f]x(t) \) at \( t \), so that its retained earnings are \( [P(t) - c_f]x(t)u(t) \). If a unit of capacity costs \( SZ \), then \( S \) buys \((1/Z)\) units of capacity. Because it is

spending \([P(t) - c_1]x(t)u(t)\) on capacity and $1 buys \((1/Z)\) units of capacity then the fringe adds capacity equal to \([P(t) - c_1]x(t)u(t)(1/Z)\), as stated in (6.2). In addition to choosing how much to produce, each fringe firm decides on what retention ratio to set at each point in time. It acts so as to maximize the value of the firm, that is, the sum of future discounted dividends.

There are several important properties to note about the fringe capacity expansion equation 6.2. If the dominant firm sets a higher price, fringe profit expands, which means that the fringe has more earnings for investment. If the retention ratio remains the same, the fringe will expand at a higher rate. Furthermore, the bigger the fringe (as measured by its current capacity), the larger is its growth, as there is more profit that can be used for investment.

Before moving on, we should briefly discuss the assumption that investment must be done out of retained earnings. Although there are many reasons why capital market imperfections might make external finance either unavailable or too expensive, suffice it to say that, empirically, internal finance is important for small firms. In 1970–1979, corporations with less than $5 million in assets retained, on average, over 80 percent of their earnings. Even for corporations with assets between $5 million and $25 million, the average retention ratio was 75 percent.\(^{32}\) It is common for new firms to retain all of their earnings and give out no dividends.

The dominant firm is going to choose a price for each \(t\) so as to maximize its sum of discounted profits. As shown in the static analysis, the optimal price depends on how much the fringe can supply today. In addition, however, it depends on the growth of the fringe as specified in (6.2). The higher today’s price, the higher is fringe profit and, holding the retention ratio fixed, the more the fringe invests in capacity expansion. Because the dominant firm’s demand is lower the greater fringe capacity becomes, higher growth of the fringe means that the dominant firm’s future demand curve will be weaker. Hence the dominant firm’s future demand will shift in more, the higher it prices today. The effect of the dominant firm’s current price on its future demand will play a central role.

En route to describing the optimal price path for the dominant firm, let us consider two extreme strategies. One pricing strategy is for the dominant firm always to set price so as to maximize current profit. Let us call this myopic pricing. Typically, setting such a high price will cause the fringe to invest in capacity and expand. Because \(x(t)\) is increasing over time, the dominant firm’s demand curve will be shifting in. Weaker demand results in the myopic price falling over time, along with the dominant firm’s profit. Figure 6.5 shows the time path of price and fringe capacity; figure 6.6 shows the time path of the dominant firm’s profit. Depending on the relative unit cost of the dominant firm and the fringe, the dominant firm’s market share could go to zero.

Figure 6.5
(a) Price Paths for a Dominant Firm. (b) Time Path of the Size of the Fringe

Figure 6.6
Profit Paths for a Dominant Firm
A firm whose behavior was consistent with a myopic pricing strategy was the Reynolds International Pen Corporation.33 Reynolds was one of the inventors of the ballpoint pen. Starting in 1945, it sold its ballpoint pen for between $12 and $20, while unit cost was only 80 cents. In response to this high price, a hundred competitors rushed into the market. By 1948, Reynolds’s market share was zero! However, it made off with considerable profits over that three-year period.

The polar opposite case is for the dominant firm to set price so as to prevent all fringe expansion. The price at which fringe expansion is zero is called the limit price. We will denote it by $\bar{P}$. Clearly, pricing at $c_f$ will achieve the goal of zero fringe expansion, as fringe firms have no earnings to finance investment. However, the limit price is actually higher than $c_f$. Though a price exceeding $c_f$ does allow positive fringe profit, fringe firms find it optimal not to invest in capacity if they expect a price close to $c_f$ into the future. The reason is as follows.

To produce one more unit forever requires an initial investment of $Z$ (the cost of the additional unit of capacity) and a per period cost of $c_f$. The present value of an initial cost of $Z$ and $c_f$ forever is $Z + (c_f/r)$, where $r$ is the discount rate. If the fringe expects to get a price of $P$ forever, its discounted revenue is $P/r$. It will choose to invest in another unit of capacity if and only if

$$\frac{P}{r} > Z + \frac{c_f}{r} \Rightarrow P > rZ + c_f.$$

Thus the limit price is $rZ + c_f$, which exceeds $c_f$.

In comparing myopic pricing and limit pricing in figure 6.6, note that myopic pricing gives higher profits upfront (that is, before $t^*$) and that limit pricing gives higher profits in the future (that is, after $t^*$). A dominant firm then prefers myopic pricing to limit pricing when its discount rate is high, as it values future profit much less than current profit. In contrast, if the dominant firm is very patient (that is, its discount rate is low), it would prefer limit pricing to myopic pricing, as the former yields a higher level of profit in the long run.

In fact, the price path that maximizes the dominant firm’s sum of discounted profits entails neither myopic pricing nor limit pricing but rather something in between. The optimal price path is shown in figure 6.5(a). Note that it starts out above the limit price but below that which maximizes current profit. Price falls over time until it reaches the limit price at time $t^0$ and then remains there forever. It follows that the fringe expands up to $t^0$, inasmuch as price exceeds the limit price, and stops growing after time $t^0$.

The dominant firm’s price path is determined by a crucial dynamic trade-off. Setting a price closer to that which maximizes current profit raises current profit but reduces future profit,

34. There are a number of assumptions made to derive this result, including that the market demand function be linear.
as it causes the fringe to expand at a faster rate. This expansion causes the dominant firm’s future demand curve to shift in at a faster rate. In order to slow down fringe expansion, the dominant firm prefers to charge a price below that which maximizes current profit.

An important property of the dominant firm’s pricing path is that price falls over time. This result occurs because the dominant firm’s demand curve is shifting in because of the expansion of the fringe. Weaker demand results in a lower price. However, a second factor is also at play. From (6.2), note that the rate of fringe expansion is higher the bigger the fringe becomes, because its earnings are bigger. As a result, initially fringe expansion is relatively small because of its small size. Hence a high price early on is less costly because the fringe will expand slower than it would if that same price were set later when it is bigger.

### Strategic Competition

When a firm acts to improve its future position in the market, it is said to engage in strategic competition. Among its many forms, this behavior can take the form of raising the cost of entry so as to reduce the number of future entrants, or investing in cost-reducing capital so as to achieve a cost advantage vis-à-vis one’s competitors. In the remainder of this chapter, we will review a variety of forms of strategic competition.

The analysis that we are about to pursue is quite distinct from what we have done thus far. In chapter 5 and in the section on market structure in this chapter, we examined how market structure affects firm behavior. In contrast, strategic competition is concerned with the reverse causality—firm behavior affecting market structure. More specifically, an incumbent firm influences its share of the market and the number and relative capabilities of its competitors. This issue was analyzed in the preceding section for the case where a dominant firm faces a competitive fringe. We showed that a dominant firm reduces price over time in order to reduce the growth of fringe firms. Although the dominant-firm model does consider how firm conduct affects future market structure, the analysis of this section differs in an important way. We assumed that the dominant firm’s competitors were small price-taking firms. We now want to allow for all firms to be large in the sense that their output decisions affect the market price. All firms are assumed to be strategic.

The most extreme form of strategic competition involves reducing the number of one’s competitors. This can entail driving some existing competitors out of the industry or deterring prospective firms from entering. **Predatory pricing** is a pricing strategy designed to promote the exit of other firms. It will be reviewed in chapter 9. The branch of strategic competition concerned with preventing entry is referred to as **strategic entry deterrence.** It will be the central focus of our ensuing analysis. Because strategic entry deterrence is a form of
monopolization practice, the theories that we will review here are relevant to the antitrust analysis performed in chapter 9.\textsuperscript{35}

**Limit Pricing**

There are two central questions underlying the literature on strategic entry deterrence. First, how can incumbent firms affect a potential entrant’s decision to enter? Second, if they can influence that decision, how does this ability affect the behavior of incumbent firms? We can also ask, even if entry does not occur, does the threat of entry induce incumbent firms to act more competitively?

The earliest work exploring strategic entry deterrence was that of limit pricing. Limit pricing is the use of price by established firms to deter entry. A form of limit pricing was analyzed when we showed how a dominant firm could use price to prevent expansion by competitive fringe firms. Here we shall assume that a new firm is not small and could be as large as established firms.

**Bain-Sylos Model**

The earliest model of limit pricing is due to Paolo Sylos-Labini and that industrial-organization legend, Joe Bain.\textsuperscript{36} (For a picture of Joe Bain, see the June 1983 issue of the *American Economic Review*.) A central assumption in this model is what is known as the Bain-Sylos postulate:

The entrant believes that, in response to entry, each incumbent firm will continue to produce at its pre-entry output rate.

This can be illustrated using figure 6.7. Assume that there is a single incumbent firm (or, alternatively, a perfectly colluding cartel) and that it is producing \( q^0 \) and selling it at \( P^0 \) prior to entry. Given the Bain-Sylos postulate, this means that the output of the entrant will simply add to \( q^0 \), causing price to fall. In short, the line segment \( AB \) is the residual demand facing the potential entrant (where the entrant’s origin is at \( q^0 \)). For convenience, we can shift the residual demand curve \( AB \) leftward to \( P^0B' \), thereby making the origin for the entrant along the vertical axis.


\textsuperscript{36} This model was developed independently by Bain (1956) and Paolo Sylos-Labini, *Oligopoly and Technological Progress* (Cambridge, Mass.: Harvard University Press, 1962).
Notice that the incumbent firm can manipulate the residual demand curve by its choice of its preentry output. For example, in figure 6.7 higher output than $q^0$ would imply a lower residual demand curve (that is, the dashed curve $CD$). This means that the incumbent firm could choose its output so that the residual demand curve facing the potential entrant would make entry just unprofitable.

This situation is shown in figure 6.8. The long-run average cost curve for a typical firm in this industry is $AC(q)$. This average cost function holds for both incumbent firms and new firms, so that there is no barrier to entry in terms of an absolute cost advantage. As shown, average cost declines until $\hat{q}$, minimum efficient scale, is reached and then becomes constant.

The key question is, Can the incumbent firm create residual demand curve for the entrant such that entry is unprofitable? The answer is yes. The residual demand curve $PB'$ that is just tangent to $AC(q)$ is one for which there is no output for a new firm that gives it positive profit. Working backward from the residual demand curve, an incumbent firm output of $\bar{q}$ is necessary to generate the residual demand curve. The price associated with $\bar{q}$, denoted $\bar{P}$, is the limit price. It is the maximum price that deters entry. We then find that if the incumbent firm prices at $\bar{P}$, it deters entry and earns above-normal profit of $[\bar{P} - AC(\bar{q})]\bar{q}$. 

Figure 6.7
Residual Demand Curve under the Bain-Sylos Postulate
Critique of the Bain-Sylos Postulate

The key assumption in this analysis is that a potential entrant expects the incumbent firm to respond to entry by maintaining its output at its preentry rate. Industrial organization economists generally consider the Bain-Sylos postulate to be a “bad” assumption. First, it is generally thought to be undesirable to assume something about how an agent behaves other than that he acts to maximize his own well-being. In this light, the Bain-Sylos postulate has been criticized in that it assumes how the incumbent will respond to entry (or how the potential entrant believes the incumbent will respond to entry, which is just as bad). Rather than make such an assumption, we want to derive how a profit-maximizing incumbent firm will respond to entry.

This more methodological criticism aside, the Bain-Sylos postulate is a bad assumption because an incumbent firm will typically choose not to behave in the manner that is assumed. In response to entry, an incumbent firm will typically want to reduce its output below its preentry level. Recall from the Cournot model that a firm’s optimal output rate is lower the more its competitors produce. Because entry entails more output being provided by one’s competitors, the profit-maximizing output of the incumbent firm should be lower after entry, not the same. In the Bain-Sylos model, entry is deterred only because the potential entrant believes the incumbent firm’s threat to maintain its output. Yet the threat is not credible!
There is an even more basic point here. In fact, the entry decision is wholly independent of the incumbent firms’ preentry output. To see this crucial point, consider a three-stage model in which the incumbent firms choose their outputs in stage 1, the potential entrant decides whether or not to enter in stage 2, and the active firms (which include the potential entrant if it entered) simultaneously choose output in stage 3 (that is, engage in Cournot competition). The profitability of entry depends on a new firm’s profit at the Cournot solution in stage 3 as well as the cost of entry. The key question to ask is, How does the incumbent firm’s output in stage 1 affect the Cournot solution achieved in stage 3? If the incumbent firm’s preentry output is to affect the entry decision, it must affect the cost of entry and/or the postentry equilibrium. If the postentry demand and cost functions are independent of past output decisions, then the postentry equilibrium will be independent of the incumbent firms’ preentry output. Hence the entry decision is independent of preentry output.\footnote{This point was originally made in James W. Friedman, “On Entry Preventing Behavior,” in Steven J. Brams et al., eds., \textit{Applied Game Theory} (Vienna: Physica-Verlag, 1979).}

### Strategic Theories of Limit Pricing

For an incumbent firm to affect entry decisions, its preentry behavior must affect the profitability of entry. What might cause this effect to occur? The presumption is that if entry occurs, all active firms will achieve some oligopoly solution. As we found in chapter 5, the determinants of an oligopoly solution include the market demand curve, firms’ cost functions, and the number of firms (as well as, perhaps, firms’ discount rates). It follows that for incumbent firms to influence the profitability of entry, they must affect the postentry demand function or their cost functions or a new firm’s cost function. A central goal in the literature on strategic competition is to identify and explore the intertemporal linkage between incumbent firms’ preentry behavior and the postentry structure in terms of cost and demand functions.

The way we will proceed in our discussion is as follows. In this section, we will describe some of the ways in which incumbent firms’ preentry output or price decisions can affect the postentry demand or cost functions and thereby affect the profitability of entry. Having established a linkage between preentry decisions and the postentry outcome, the next step would be to examine how an incumbent firm might exploit this linkage. This will be postponed until the following section, where we investigate the use of capacity to deter entry. Though capacity is the entry-deterring instrument, the method of analysis and general insight applies as well to when preentry price or output is the instrument used by incumbent firms.

In our discussion, let us assume there is just one incumbent firm and one potential entrant. In trying to find ways in which the incumbent firm’s preentry output can affect the postentry equilibrium, one needs to think of reasons why past output would affect current demand or cost functions. One source of this intertemporal linkage is \textit{adjustment costs}. In many manu-
farming processes, costs are incurred in changing the rate of production. To increase output, a firm may need to bring new equipment on line. Installation of this equipment can require shutting down the production process, which is costly in terms of lost output. To reduce output, a firm may need to lay off workers, which is also costly. Adjustment costs are those costs incurred from changing the firm’s rate of production.38

An example of a cost function with adjustment costs is

$$C(q') = 100 + 20q' + \frac{1}{2}(q' - q'^{-1})^2$$

where \(q'\) is the period \(t\) output and \(q'^{-1}\) is output from the previous period. The cost to adjusting output is measured by $$\frac{1}{2}(q' - q'^{-1})^2$$. Notice that it is minimized when \(q' = q'^{-1}\), so that there is no change in output. It is greater the bigger the change in output.

When the incumbent firm incurs a cost to adjusting its production rate, we want to argue that its preentry output will affect the profitability of entry. The more a firm produces today, the higher is its profit-maximizing output in the future. Because of the cost to adjusting its output, a firm will tend to produce an output close to its past output. Thus, if the postentry equilibrium is the Cournot solution, then the incumbent firm will produce at a higher rate after entry, the more it produced in the preentry period. As shown in figure 6.9, increasing its preentry output shifts out the incumbent firm’s postentry best reply function. A rise in its preentry output then shifts the postentry equilibrium from point \(A\) to \(B\). Because the incumbent firm produces more at \(B\), postentry profit for the new firm is less. An incumbent firm may then be able to deter entry by producing at a sufficiently high rate prior to the entry decision.

By the preceding analysis, one can motivate the Bain-Sylos postulate by the assumption that the incumbent firm faces infinitely large adjustment costs, so that it would be too costly to change its output in response to entry. This is obviously a very extreme assumption, though it does make the Bain-Sylos postulate an assumption on the structure of the model rather than an assumption on behavior.

What this analysis has shown is how the existence of adjustment costs can create a linkage between an incumbent firm’s preentry output and the postentry equilibrium. Given that there is a way in which an incumbent firm can affect the profitability of entry, the next question is whether it is optimal to use it so as to deter entry. It is possible that it could take a very high preentry output in order to make entry unprofitable. In that case, an incumbent firm might prefer to produce a lower output and let entry occur. A detailed treatment of the decision to deter entry is provided in the next section, in which we explore capacity as the entry-deterring instrument.

38. For an analysis of entry deterrence when there are adjustment costs, see M. Therese Flaherty, “Dynamic Limit Pricing, Barriers to Entry and Rational Firms,” Journal of Economic Theory 23 (October 1980): 160–82. This was one of the earliest pieces in the recent line of work on entry deterrence.
Let us briefly mention several other ways in which the incumbent firm’s preentry output or price can affect the profitability of entry. Some production processes have a learning curve; that is, the more experience a firm has with the process, the more ways it finds to lower cost. One reason is that intricate labor operations, such as in aircraft assembly and the manufacture of computer components, can become more efficient as workers gain experience. Using cumulative past output as a measure of experience, an example of a cost function with a learning curve effect is

\[
C(q') = 100 + \left[ 20 + \frac{1}{1 + Y'} \right] q'
\]

where \( Y' \) is the sum of past outputs: \( Y' = q'^{-1} + q'^{-2} + \ldots \). Note that the more a firm produced in the past, the lower is its marginal cost today. Hence the higher an incumbent firm’s preentry output becomes, the lower is its marginal cost after entry, which means the more it will produce in the postentry equilibrium. Hence, entry deterrence could occur by setting a high
preentry output, which would lower the incumbent firm’s marginal cost and give it a cost advantage vis-à-vis a new firm.\textsuperscript{39}

Earlier in this chapter, we mentioned that consumers are hesitant to switch brands of a good because there are certain costs associated with doing so. To switch banks you must close out your account, which entails a certain amount of time and effort. If the quality of an untried brand is uncertain, then a consumer will have to incur costs associated with learning about the new brand. Such costs are avoided by sticking with the brand one is currently using, as personal experience has already relieved this uncertainty. Such costs are referred to as \textit{switching costs}.

The demand for a new firm’s product comes from consumers who currently buy from existing firms and consumers who are currently not in the market. If there are switching costs, a new firm is at a disadvantage when it comes to competing for the first type of consumer. It will have to offer such consumers a price discount to induce them to switch because of the costs associated with doing so. No such price discount has to be offered to consumers who are not currently buying. Incumbent firms can make entry less profitable by increasing the number of consumers who have attached themselves to an existing brand. This is most easily achieved by offering a low preentry price. A new firm would then have to offer a large price discount in order to get much of any demand for its product. The price discount required may be so large as to make entry unprofitable.\textsuperscript{40}

In the preceding discussion we considered how the incumbent firm’s preentry output can affect the postentry demand function or its postentry cost function. Given this linkage, the incumbent firm may be able to make the postentry equilibrium sufficiently unattractive so as to deter entry. Alternatively, if an entrant is uncertain about what demand and cost will be after entry, the incumbent firm may be able to deter entry by influencing a potential entrant’s \textit{beliefs} over postentry demand or cost functions rather than the actual demand or cost functions themselves.

To consider this possibility, suppose that the potential entrant is uncertain of the incumbent firm’s marginal cost. Assume that entry is profitable when the incumbent firm’s marginal cost is comparable to or higher than that of the potential entrant. If instead the incumbent firm has a considerable cost advantage, entry is assumed to be unprofitable. Because an incumbent firm’s marginal cost affects its profit-maximizing output, one would expect its preentry output to provide information about its cost. A potential entrant would tend to infer from a high preentry output (or a low preentry price) that the incumbent firm has low


marginal cost and, therefore, entry would be unprofitable. Of course, the incumbent firm might then produce a lot even if its marginal cost is high, in order to try and mislead the potential entrant into believing that it has low cost. Hence, the incentive to signal that you have low cost results in the incumbent firm producing at a higher rate than it would if there were no threat of entry. Whether the incumbent firm is successful in deterring entry depends on the particular model. Nevertheless, this signaling phenomenon does generate a role for potential competition in keeping an incumbent firm’s price below the monopoly level even if entry never takes place.41

In summary, there are two important conclusions from the limit-pricing literature. First, there are a number of different ways in which preentry output or price can affect the postentry equilibrium and thereby influence the decision to enter. Second, even if entry is deterred, the threat of entry will generally induce incumbent firms to produce at a higher rate.

Investment in Cost-Reducing Capital

In a pioneering paper, Avinash Dixit provided some fundamental insight that has fueled much of the research into strategic entry deterrence.42 Because his paper is representative of the type of analysis and insight in models of strategic competition, we will consider it in some depth.

Let us begin with a description of the technology available to firms. To operate at all in this industry, a firm must incur some fixed cost, which we denote $K$. In order to produce one unit, a firm needs one unit of capacity, which costs $r$, and variable inputs that cost $w$. If a firm currently has a capacity stock of $x$, its cost function is then

$$C(q) = \begin{cases} 
K + rx + wq & \text{if } q \leq x \\
K + (w + r)q & \text{if } q \leq x.
\end{cases}$$

Given preexisting capacity stock of $x$, a firm has a fixed (and sunk) cost of $K + rx$. To produce $q$ when it does not exceed capacity requires only variable inputs that cost $wq$. However, if output is in excess of capacity, one must buy $(q - x)$ additional units of capacity. This costs $r(q - x)$, and therefore total cost is $K + wq + rx + r(q - x) = K + (w + r)q$.

There are three stages to the game and two firms—one incumbent firm and one potential entrant. Firms have homogeneous products. Initially, neither firm has any capacity.

Stage 1: The incumbent firm invests in capacity, denoted $x$.

Stage 2: The potential entrant observes the incumbent firm’s capacity and decides whether or not to enter. Entry costs $K > 0$.


**Stage 3:** Active firms simultaneously choose how much to invest in capacity and how much to produce. The incumbent firm carries $x$ from stage 1.

If there was entry, there would be two active firms in stage 3. Otherwise there is just one—the incumbent firm.

To derive an equilibrium for this game, let us begin with the thought experiment of what would happen in stage 3 if entry occurred and how this would depend on the incumbent firm’s initial investment in capacity. The presumption is that each firm chooses a quantity so as to maximize its profit, which means a Nash equilibrium. The point we want to argue is that, generally, the higher is the initial capacity of the incumbent firm, the higher is the incumbent firm’s postentry Nash equilibrium quantity and the lower is the new firm’s postentry Nash equilibrium quantity. In figure 6.10 is the marginal cost curve that the incumbent firm faces in stage 3 given an initial capacity of $x_1$. Marginal cost is $w$ if the incumbent firm produces below its initial capacity and jumps to $w + r$ if it produces above $x_1$, as it must add capacity, which costs $r$ per unit. If the incumbent firm’s initial capacity is instead $x_1^0$, then its marginal cost is lower for all quantities between $x_1$ and $x_1^0$. Generally, when a firm’s marginal cost is lower, it desires to produce more. As a result, in the postentry game, the incumbent firm finds it optimal to produce at a higher rate when its initial capacity is $x_1^0$ than when it is $x_1$. Because this higher quantity means that the new firm can expect a lower market price for any quantity that it would produce, it chooses to produce a lower amount. In other words, higher initial

---

**Figure 6.10**

Incumbent Firm’s Marginal Cost Curve
capacity for the incumbent firm credibly commits it to producing more in stage 3 because its marginal cost is lower. This commitment induces a new firm to produce less. We can further conclude that a new firm’s postentry profit is then lower, the greater is the initial capacity investment (in stage 1) of the incumbent firm, because any firm’s profit is lower when its rival produces more.

What we have described is a linkage between the incumbent firm’s preentry capacity investment and the postentry profit for a new firm. Through this linkage, the incumbent firm may have an instrument for deterring entry. By having sufficiently great capacity, the incumbent firm may be able to depress postentry profit for a new firm to the point that entry would be unprofitable.

To explore whether the incumbent firm will choose to deter entry, we consider a simplified version of the Dixit game as shown in Figure 6.11. To read this diagram, start from the top and read downward. At the start, the incumbent firm has two choices: low capacity (Low) or high capacity (High). Thus, we are restricting \( x_1 \) to being one of two values. One can think of building a small plant or a large plant. After it makes its investment decision, the potential entrant decides whether to enter. If entry does not occur, the incumbent firm earns the profit associated with being a monopolist. This is specified to be 15 when the incumbent chooses Low in the capacity stage and 12 when it chooses High. Without the threat of entry, it would then choose to invest a low amount in capacity.

**Figure 6.11**
Modified Dixit Game
If entry does take place, then firms make simultaneous quantity decisions where each firm can decide to produce at a low rate \((q_L)\) or at a high rate \((q_H)\). A matrix lists the profits for the two firms from the four possible quantity pairs. The first number in a cell of a matrix is the profit for the incumbent firm, whereas the second number is the profit for the new firm (recall that \(K\) is the cost of entry). For example, if the incumbent firm chooses \(x_1 = \text{Low}\), entry occurs, and both firms produce \(q_H\), then the profits of the incumbent firm and the new firm are 5 and 5\(-K\), respectively. Implicit in these profit numbers is that the incumbent firm prefers to produce at a high rate when it invests heavily in capacity (as its marginal cost is low) and prefers to produce at a low rate when it does not invest heavily in capacity (as its marginal cost is high for a high rate of production).

If \(x_1 = \text{Low}\) and entry occurs, the postentry (Nash) equilibrium has both firms producing \(q_L\). Given that the new firm produces \(q_L\), the incumbent firm earns 8 by producing \(q_L\) and 7 by producing \(q_H\), so it prefers \(q_L\). Given that the incumbent firm produces \(q_L\), the new firm earns 8\(-K\) by producing \(q_L\) and 7\(-K\) by producing \(q_H\), so it prefers \(q_L\). You should prove to yourself that the other three possible outcomes, \((q_L, q_H)\), \((q_H, q_L)\), and \((q_H, q_H)\), are not Nash equilibria. If instead \(x_1 = \text{High}\), then the postentry equilibrium has the incumbent firm producing \(q_H\) and earning 7, and the new firm producing \(q_L\) and earning 6\(-K\).

How will the potential entrant behave? Well, it depends on the cost of entry and the incumbent firm’s capacity. If \(x_1 = \text{Low}\), then the postentry equilibrium profit for a new firm is 8\(-K\). Hence, if \(x_1 = \text{Low}\), entry is profitable (and will occur) if and only if \(K < 8\) (note that the potential entrant always earns 0 from not entering). If \(x_1 = \text{High}\), then postentry profit is 6\(-K\) so that it enters if and only if \(K < 6\).

The final step in deriving equilibrium behavior for this three-stage game is to determine the optimal behavior of the incumbent firm at the capacity stage. There are three cases to consider. Case 1 is when the cost of entry is high: \(K > 8\). Because entry is unprofitable regardless of the incumbent firm’s capacity, the incumbent firm will choose the same capacity as if there were no threat of entry: \(x_1 = \text{Low}\). In Case 2, the cost of entry is intermediate: 6 < \(K < 8\). Under Case 2, the incumbent firm deters entry when \(x_1 = \text{High}\) and induces entry when \(x_1 = \text{Low}\). It then earns 8 from \(x_1 = \text{Low}\) and 12 from \(x_1 = \text{High}\). Hence it optimally chooses to set \(x_1 = \text{High}\) and deters entry. Note that it invests more in capacity than if there were no threat of entry. Finally, Case 3 occurs when entry is relatively inexpensive: \(K < 6\). Entry occurs regardless of the incumbent’s capacity. The incumbent then earns 8 from \(x_1 = \text{Low}\) and 7 from \(x_1 = \text{High}\), so that it invests a low amount in capacity. Thus the incumbent firm chooses the same capacity as a monopolist would who had no threat of entry. In a more general model, the incumbent firm would strategically raise its capacity in anticipation of entry in order to reduce the new firm’s output.

There are several important lessons to be learned from the preceding analysis. Since capacity is durable and lasts into the postentry period, an incumbent firm’s capacity investment affects its future cost function and thereby affects the postentry equilibrium. Because of its
durability, capital is a natural instrument for strategically deterring entry and, more generally, for improving one’s position in the market. A second lesson is that even if there is a strategy that effectively deters entry, an incumbent firm may not choose to use it. In the preceding game, when $6 < K < 8$, entry was strategically deterred by the incumbent investing a lot in capacity. However, we could make high capacity relatively more expensive so that the incumbent firm’s profit when it chooses $x_1 = \text{High}$ is less than its profit by having low capacity and allowing entry.

Raising Rivals’ Costs

We have thus far discussed a number of ways in which an incumbent firm can improve its future position in the market by giving itself a cost advantage over its competitors. This could involve, for example, a learning curve or cost-reducing investment. In this section, we consider a strategy in which an incumbent firm gives itself a cost advantage by raising its rivals’ costs rather than lowering its own cost.

We will consider one strategy, of which there are several, for raising the cost of a rival. Suppose that firm 1 has a relatively capital-intensive production process while its competitor, firm 2, has a relatively labor-intensive production process. Further suppose that the industry is unionized and the union bargaining process works in the following manner. The union first bargains with firm 1 concerning its wage. On agreement with firm 1, the union demands that firm 2 pay the same wage. The (credible) threat of a strike by the union should be able to induce firm 2 to accept the wage that resulted from negotiations between the union and firm 1. Bargaining between the United Auto Workers and the domestic auto manufacturers takes place in this manner.

From a strategic perspective, it may actually be in firm 1’s best interests to agree to a relatively high wage with the union. Though a higher wage raises its cost, it raises firm 2’s cost even more as firm 2 uses more labor in producing its product. Agreeing to a high wage then gives firm 1 a cost advantage over firm 2, even though firm 1’s cost has increased! The rise in firm 2’s marginal cost will cause it to reduce its output, which will raise firm 1’s revenue. As long as this revenue increase is bigger than the increase in firm 1’s cost, firm 1 will have raised its profits by agreeing to a higher wage.

Preemption and Brand Proliferation

Consider a market where there are two possible products: $X$ and $Y$. Assume that these products are imperfect substitutes and that all firms have access to the technology to produce

either one. Initially, the industry is characterized as follows. There is a lone incumbent firm that produces only $X$. Suppose that demand for $Y$ is sufficiently weak so that it is unprofitable for the incumbent firm to offer $Y$, nor is it profitable for a new firm to come in and offer $Y$. Furthermore, it is assumed that a potential entrant would find it unprofitable to enter and produce $X$, as it would put it in direct competition with the incumbent firm.

Now suppose that there is an unanticipated permanent increase in the demand for $Y$. Under this new demand structure, let $\Pi(A, B)$ denote the profit to a firm if its product line is $A$ and its competitor’s product line is $B$. There are four possible product lines: product $X$, product $Y$, products $X$ and $Y$ (denoted $XY$), and no products (denoted $N$). For example, $\Pi(XY, Y)$ is the profit of a firm offering both $X$ and $Y$ when its competitor offers just $Y$. To produce product $Y$, a firm has to incur a fixed cost of $F_Y$.

Assume that if there was no threat of entry, the incumbent firm would choose not to offer $Y$: 

$$\Pi(X, N) > \Pi(XY, N) - F_Y.$$  \hspace{1cm} (6.3)

The left-hand side is the profit from a monopolist offering only $X$ and the right-hand side is the profit from it expanding to offering $Y$ as well. Equation 6.3 will hold if the increase in demand for $Y$ is sufficiently small relative to the fixed cost $F_Y$. We will further assume that entry by a new firm with product $Y$ is profitable if the incumbent firm offers only $X$ and is unprofitable if it offers both products:

$$\Pi(Y, X) - F_Y > 0 > \Pi(Y, XY) - F_Y.$$  \hspace{1cm} (6.4)

Here, $\Pi(Y, X) - F_Y$ is the profit to a new firm from exclusively offering $Y$ and $\Pi(Y, XY) - F_Y$ is the profit to a new firm from offering $Y$ if the incumbent firm offers both $X$ and $Y$. The condition $0 > \Pi(Y, XY) - F$ is natural if competition is sufficiently intense when two firms offer the same product.

Note that (6.3) and (6.4) are quite consistent. Because products $X$ and $Y$ are substitutes, if product $Y$ is offered, the demand for product $X$ will fall. Although introduction of $Y$ generates positive profit, it reduces profit from the sale of $X$. Of course, a new firm that offers only $Y$ does not care about the reduced profits on $X$. In contrast, when the incumbent firm considers offering $Y$, it cares about total profits. Thus it values the introduction of $Y$ by the profits earned from $Y$ less the reduction in profits earned from $X$.

In considering how the incumbent firm will behave in response to the rise in demand for $Y$, let us assume that it could put $Y$ on the market before a new firm could. This assumption is actually not important for the story we are telling though it is a natural one. For example, the incumbent firm is apt to learn about the rise in demand for $Y$ before anyone else. We know that if the incumbent firm does not offer $Y$, then entry will occur. In that case, the incumbent firm’s profit is $\Pi(X, Y)$. If instead it was to introduce $Y$, entry would be unprofitable [as $\Pi(Y, XY) - F_Y < 0$] and thus would not occur. The incumbent firm’s profit from offering $Y$ is then
Thus the incumbent firm will find it optimal to preempt entry by introducing $Y$ if and only if

$$\Pi(\text{XY}, N) - F_Y > \Pi(X, Y).$$

(6.5)

Rearranging (6.5), one derives:

$$\Pi(\text{XY}, N) - \Pi(X, Y) > F_Y.$$  

(6.6)

Recall from (6.4) that it was assumed that $\Pi(Y, X) > F_Y$. Thus, if the left-hand side of (6.6) is at least as great as $\Pi(Y, X)$ then, because $\Pi(Y, X) > F_Y$, it must exceed $F_Y$. It follows that (6.6) holds if the following is true:

$$\Pi(\text{XY}, N) - \Pi(X, Y) \geq \Pi(Y, X).$$

(6.7)

Rearranging the expression in (6.7) gives us

$$\Pi(\text{XY}, N) \geq \Pi(X, Y) + \Pi(Y, X).$$

(6.8)

Thus, if (6.8) holds, then the incumbent firm will introduce product $Y$ and thereby deter a new firm from doing so.

The left-hand side of (6.8) is profit earned by a single firm offering both products $X$ and $Y$. The right-hand side is industry profit from one firm offering product $X$ and a second firm offering product $Y$. Because there are two competing firms in the market, generally competition will drive prices down below the level that a two-product monopolist would set (unless the two firms were to perfectly collude). Assuming that there is no cost disadvantage from offering both products, (6.8) must then be true. One two-product firm earns more than two single-product firms competing against one another.

The point here is simple and quite intuitive. If a new firm enters and offers $Y$, there will be two firms competing in the market—one with product $X$ and one with product $Y$. This competition results in prices being driven below that which would be set by a single firm selling both products. In contrast, the incumbent firm can coordinate the pricing of products $X$ and $Y$ if it offers $Y$. Hence the value of introducing $Y$ to the incumbent firm is greater than that to a new firm, so that the incumbent firm will introduce product $Y$ before a new firm can.

Although we gave the incumbent firm first shot at offering product $Y$, this is not important to the story. Suppose that, instead of assuming that demand for $Y$ suddenly increases, we assume it rises slowly over time in an anticipated manner. At some point in time, call it $t^*$, entry by a new firm with product $Y$ yields a normal return (that is, the sum of discounted profits equals zero). Entry before $t^*$ results in a below-normal return while entry after $t^*$ results in an above-normal return (assuming no one else has introduced $Y$). By the reasoning given previously, the incumbent firm earns positive profit from offering $Y$ at $t^*$ (more than a
new firm would earn) because it can coordinate the pricing of $X$ and $Y$. Hence the incumbent firm must earn positive profit by introducing product $Y$ just before $t^*$. In contrast, a new firm would not offer it anytime before $t^*$, as it would be unprofitable. We then find that the incumbent firm will preempt entry by introducing product $Y$ before a new firm would find it profitable to do so.

The strategy outlined here is referred to as brand proliferation. Incumbent firms offer new brands in order to fill niches in the market that would have provided room for profitable entry. Note that, in the absence of the threat of entry, product $Y$ would not have been introduced (see equation 6.3). Hence, its introduction is solely for the purpose of deterring entry. As discussed in chapter 9, the leading manufacturers of ready-to-eat breakfast cereal were accused by the Federal Trade Commission of using an entry-deterrence strategy of brand proliferation, though the case was eventually dropped.

Generally, society is better off having a new firm supply product $Y$ than having the incumbent firm do so. The reason is that entry results in lower prices and higher consumer surplus. It is important to emphasize, however, that it may or may not be socially optimal for product $Y$ to be supplied at all. Although the introduction of $Y$ increases product variety, the resulting rise in consumer welfare may be inadequate to compensate for the fixed cost $F_Y$. This conclusion could be true even if entry by a new firm with product $Y$ is profitable.

The reasoning as to why competition may not result in the socially optimal product variety is as follows. A new firm will enter with product $Y$ when it results in positive profits. Part of these profits come from taking part of the incumbent firm’s demand. This part of a new firm’s profits does not add to social welfare but is simply a transfer from the incumbent firm to the entrant. Recall that the change in social welfare from entry equals the sum of the change in industry profits (which equals the sum of new firm’s profits and the change in profits of existing firms) and the change in consumer surplus. Although the change in consumer surplus is typically positive and a new firm’s profits are positive (otherwise there would be no entry), industry profits are usually reduced.

It is possible for industry profits to fall more than consumer surplus rises so that entry actually reduces social welfare. As an example, consider the case when products $X$ and $Y$ are close substitutes, so that the introduction of product $Y$ would not greatly improve product variety. If firms anticipate colluding in price, then entry will not improve product variety very much, nor will price be much lower. The cost of entry could overwhelm these gains so as to reduce industry profits and social welfare (even though consumer surplus is higher). Competition may not result in the socially optimal product variety.

Let us offer a few final remarks. First, the strategy of preemption has also been shown to apply to the construction of new plants and to the patenting of new innovations. Second, an important assumption in the foregoing analysis is that it is very costly for the incumbent firm to take a product off the market. In particular, it is assumed that if the incumbent firm introduced $Y$, then, if a new firm came in and sold $Y$, it would not be optimal for the incumbent firm to take $Y$ off the market. If, in fact, product exit costs are not high, it would actually be profitable for the incumbent firm to take $Y$ off the market but it would not be profitable for a new firm to do so. Thus, if exit costs are low, the preemption strategy will not work.\footnote{See Kenneth L. Judd, “Credible Spatial Preemption,” \textit{Rand Journal of Economics} 16 (Summer 1985): 153–66.}

**Summary**

In chapters 5 and 6 we analyzed the feedback relationship between market structure and firm conduct. Initially, market structure was taken as exogenous, and we examined how concentration and entry conditions influence the way firms behave. Our analysis then turned to exploring how firm behavior can affect structure. Under the title of strategic competition, we found there to be many ways in which firms can influence their future market share and the number of firms. Examples include creating a cost advantage over one’s existing competitors by raising their costs or preemting future entry through one’s product decisions. Although for many years industrial organization economists treated entry conditions as being exogenous to firms, in fact entry conditions are partially determined by the behavior of established firms.

**Questions and Problems**

1. The HHI can be argued to be a better measure of industrial concentration than the concentration ratio. Comment.
2. Suppose that an industry has ten firms with market shares of the following percentages: 25, 15, 12, 10, 10, 8, 7, 5, 5, and 3.
   a. Derive the four-firm concentration ratio.
   b. Derive the HHI.
   c. Derive the effect of a merger between the fifth and sixth largest firms on the HHI.
   d. Suppose the government decides that it wants this industry to have a four-firm concentration ratio of 40 percent or less. How might this goal be achieved?
3. In 1972 a U.S. senator proposed the Industrial Reorganization Act. Among other things, this act required that firms in an industry be broken up if (a) their after-tax return on stockholders’ equity exceeded 15 percent for five consecutive years, or (b) the four-firm concentration ratio exceeded 50 percent. Discuss whether or not you think this piece of legislation should have been enacted. (By the way, it was not.)

4. How would you define a barrier to entry?

5. Sunk cost has been said to be a barrier to entry. Explain the mechanism for this generalization to be true. State which definition you are using.

6. Suppose that the demand for long-distance telephone service is \( D(P) = 50 - 2P \), where \( P \) is price. Prior to recent deregulation, this market was monopolized by AT&T. Assume that AT&T’s cost function is \( C(q) = 100 + 5q \).
   
   a. If AT&T had been an unregulated monopolist, derive the price that it would have charged.
   
   b. Derive the price that a regulatory agency would set if it was interested in maximizing consumer welfare subject to AT&T earning at least normal profits.

Since deregulation, AT&T has continued to be the dominant firm. Suppose AT&T’s competitors are small price-taking firms that can be represented by the supply function \( S(P) = 2P - 20 \).

   c. Using the static dominant firm model, derive the price that AT&T would charge. Derive AT&T’s market share.

   d. How much is AT&T willing to pay to be an unregulated monopolist?

   Suppose that we extend this model to a multiperiod setting, and assume that the fringe finances growth through retaining earnings.

   e. Will AT&T’s current price be higher or lower than that derived in part c?

7. Ace Washtub Company is currently the sole producer of washtubs. Its cost function is \( C(q) = 49 + 2q \), and the market demand function is \( D(P) = 100 - P \). There is a large pool of potential entrants, each of which has the same cost function as Ace. Assume the Bain-Sylos postulate. Let the incumbent firm’s output be denoted \( q' \).

   a. Derive the residual demand function for a new firm.

   b. Given that the incumbent firm is currently producing \( q' \), if a potential entrant was to enter, how much would it produce?

   c. Find the limit price. Hint: Find the output for Ace such that the slope of a new firm’s average cost curve equals the slope of a new firm’s residual demand curve.

   Suppose, instead of assuming the Bain-Sylos postulate, that we assume active firms expect to achieve a Cournot solution.

   d. Does entry depend on \( q' \)? Explain.

   e. Will there be entry?

   Consider the Dixit capacity investment model when the inverse market demand curve is \( P(Q) = 100 - Q \), \( w = 10 \), \( r = 30 \), and \( K = 156.25 \).

   a. Derive a new firm’s postentry best reply function.

   b. Given initial capacity of \( x \), derive the incumbent firm’s postentry best reply function.
c. Given \( x \), derive the postentry equilibrium profit of a new firm.

d. Derive the minimum capacity that makes entry unprofitable.

e. Derive the optimal capacity choice of the incumbent firm.

9. From 1979 to 1987, Kellogg introduced the following brands of cereal: Graham Crackers; Most; Honey & Nut Corn Flakes; Raisins, Rice & Rye; Banana Frosted Flakes; Apple Frosted Mini Wheats; Nutri-Grain; Fruity Marshmallow Krispies; Strawberry Krispies; Crispix; Cracklin’ Oat Bran; C-3PO; Apple Raisin Crisp; Fruitful Bran; OJ’s; Raisin Squares; Just Right/Nugget & Flake; All Bran with Extra Fiber; Apple Cinnamon Squares; All Bran Fruit & Almonds; Strawberry Squares; Pro Grain; Muselix; Nutri-Grain Nuggets; and Nutrific.

a. Should the Antitrust Division of the Justice Department be concerned?

b. Five other companies introduced a total of fifty-one new brands over that same period. Does this fact change your answer in part a?

10. The demand for product \( X \) is \( P_X = 10 - 2X - Y \), where \( Y \) is the quantity of a substitute product that currently is not being produced. The marginal cost of \( X \) is a constant equal to $1. Entry is completely barred and a monopolist, “Incumbent,” produces \( X \).

a. Find Incumbent’s price, quantity, and profit.

b. Incumbent wishes to investigate the possibility of introducing \( Y \), which is also protected from entry by other firms. The demand for \( Y \) is \( P_Y = 10 - 2Y - X \) and it also has a constant marginal cost of $1. However, there is a fixed cost of introducing \( Y \) of $4. Find the values of \( X, Y, P_X, P_Y \), and profit for Incumbent. Will Incumbent introduce \( Y \)?

c. Would it be in society’s interests to have Incumbent introduce \( Y \)?

d. Now assume that entry is no longer barred. For simplicity, assume that if a new firm, “Entrant,” decides to introduce \( Y \), then Entrant and Incumbent will collude perfectly and settle at the joint profit maximum. Of course, given the demands and costs as assumed previously, the prices and quantities found in part b will apply. Will Entrant have an incentive to introduce \( Y \)?

e. Still assuming that entry is not barred, will Incumbent have an incentive to preempt the entry by Entrant and offer \( Y \) first? (It is assumed that if Incumbent offers \( Y \), then a second seller of \( Y \) will have negative profits.)

f. If the fixed cost of introducing \( Y \) is now taken to be $6, answer parts d and e. Is it in society’s interests to have \( Y \) introduced?

g. Society’s calculation of benefits for comparison with the $6 introduction cost consists of three parts: the increase in consumer surplus due to \( Y \), the increase in producer surplus from \( Y \), and the loss in producer surplus from \( X \). On the other hand, Entrant compares only one of these parts with the $6. Is it ever possible for these two “decision rules” to give the same answer? That is, is it possible for social and private decision rules to coincide such that private decisions are socially “correct”? Explain.

11. Assume the same demand and cost curves as in problem 10, except now take the fixed introduction cost of \( Y \) to depend on bidding by Incumbent and Entrant. That is, assume that a third party owns a patent on \( Y \) and has requested bids for the right to produce \( Y \).
a. Referring to your answers in parts d and e in problem 10, what would be the maximum amounts each would be willing to pay for the patent?

b. Assume now that a Cournot solution holds if Incumbent sells $X$ and Entrant sells $Y$. Find the equilibrium prices, quantities, and profits.

c. Under the Cournot solution scenario, find the maximum amounts that Incumbent and Entrant would pay for the patent so as to become the sole producer of $Y$.

d. Explain the intuition underlying your answer to part c and why it differs in a qualitative way from the answer to part a.
In chapter 5 we examined price-fixing agreements among firms and the evolution of antitrust law toward such conspiracies. In this chapter we continue our study of how cooperation among rivals can harm competition, but through the process of merger. A horizontal merger is defined as one in which rivals in the same market form one company. An example is the merger of the two aircraft manufacturers, Boeing and McDonnell-Douglas. Of course, not all horizontal mergers harm competition, though the potential to harm is clear whenever the number of competitors is reduced. Mergers, unlike price-fixing cartels, involve integration of the firm’s facilities, which raises the possibility of socially beneficial economies of combined operations. This difference explains the fact that price fixing is a per se offense, while mergers are considered under the rule of reason.

A vertical merger is one between two firms with potential or actual buyer-seller relationships. An example is the acquisition of Detroit Steel Corporation by Cleveland-Cliffs Iron (a supplier of iron ore). All other mergers—which are neither horizontal nor vertical—are classified as a conglomerate merger, which is further subdivided into three classes by the Federal Trade Commission (FTC). A product extension merger occurs with the merger of firms that sell noncompeting products but use related marketing channels or production processes. The Pepsico acquisition of Pizza Hut is an example. A market extension merger is the joining of two firms selling the same product but in separate geographic markets. An example is the 1994 acquisition by Walmart of Woolco Canada, which represented Walmart’s entry into the Canadian retail sector. Another example is the 1996 merger of Bell Atlantic and Nynex (renamed Verizon), where each offered local telephone service in nonoverlapping regions. Finally, there is the “pure” category of conglomerate mergers between firms with no obvious relationships of any kind. The merger between R. J. Reynolds (tobacco) and Burmah Oil and Gas is an example.

While the mechanism that permits horizontal mergers to potentially harm competition is clear, this is not so for conglomerate and vertical mergers. Perhaps the most obvious way for conglomerate mergers to harm competition is through agreements to remove potential competitors. This was the government’s claim in the Procter & Gamble—Clorox merger. Procter & Gamble (a giant detergent manufacturer) was alleged to have been eliminated as a potential entrant into the bleach market when it acquired Clorox. In this respect, horizontal and conglomerate mergers can be similar in their potential threats to competition, that is, the elimination of rivals (actual or potential).

The threats to competition from vertical mergers are less obvious and can be viewed as unilateral actions that potentially inflict harm on rivals. Note, for example, that the merger of an iron-ore supplier and a steel manufacturer does not change the number of competitors in either market. One popular complaint by judges has been that such mergers “foreclose” markets to rivals. Simply put, rival iron ore suppliers are harmed by no longer having the acquired steel manufacturer as a possible buyer, and this loss is thought to harm competition. Because vertical mergers are perhaps best viewed as an exclusionary activity, we will
postpone discussion of vertical mergers until the next chapter, where we also consider other vertical relationships.

The next section briefly describes historical trends in merger activity and changes in the applicable antitrust laws. The remaining sections of this chapter will examine the reasons for mergers and provide detailed analyses of horizontal and conglomerate mergers.

**Antitrust Laws and Merger Trends**

There has been an interesting interdependence between antitrust law and the trend of mergers in the United States. To begin, the United States has experienced five major merger “waves.” The first wave, occurring roughly during 1890–1904, has been described as “merger for monopoly”: “The conversion of approximately 71 important oligopolistic or near-competitive industries into near-monopolies by merger between 1890 and 1904 left an imprint on the structure of the American economy that fifty years have not erased.”

Figure 7.1 shows the acquisition volume as a fraction of the economy (as measured by Gross Domestic Product) for the first wave.

Perhaps the most famous of these mergers occurred in the steel industry. During the 1880s, over two hundred iron and steel makers were merged into twenty larger firms. In 1901, J. P. Morgan then engineered a merger among twelve of these larger firms to form United States Steel Corporation, which then had about 65 percent of the market. The result was a sharp rise in prices and a handsome $62.5 million share of these monopoly rewards for Mr. Morgan. Other well-known firms created in this period through mergers include General Electric, American Can, DuPont, Eastman Kodak, Pittsburgh Plate Glass, American Tobacco, and International Paper.

Although the Sherman Antitrust Act was passed in 1890, it was not enforced vigorously until the early 1900s. The first decision to have a chilling effect on the merger wave was the Northern Securities decision in 1904. Of course, the Sherman Act did not contain specific antimerger provisions; Section 1 concerns combinations in restraint of trade, and Section 2 deals with monopolization. In Northern Securities, the government’s attack was based on both sections, and the attempt to combine two railroads (Northern Pacific and Great Northern) was found to be illegal. Also in 1911, two famous monopolies, or trusts as they were called, were found guilty and subsequently broken up (Standard Oil and American Tobacco).

Because the Sherman Act applied to mergers only when the merging firms were on the verge of attaining substantial monopoly power, the Clayton Act was passed (in part) in 1914 to remedy this limitation. Section 7 reads:

That no corporation engaged in commerce shall acquire, directly or indirectly, the whole or any part of the stock or other share capital of another corporation engaged also in commerce where the effect of such acquisition may be to substantially lessen competition between [the two firms] or to restrain such commerce in any section or community or tend to create a monopoly of any line of commerce.

Unfortunately, the reference to stock acquisitions left a large loophole. By purchasing a competitor’s assets, mergers could escape the reach of the Clayton Act.

The second merger wave took place over the period 1916–1929. Though mergers to monopoly were now discouraged by the antitrust laws, “mergers to oligopoly” became fashionable. For example, Bethlehem Steel was formed as the second largest steel manufacturer by combining several smaller companies. The Great Depression ended this second wave.

Next came the “conglomerate merger” wave, which began after World War II and peaked around 1968. It, too, differed from its predecessors. The reason for this difference was the passage of the Celler-Kefauver Act of 1950 and the strict judicial interpretations of that
legislation. As we will describe later, horizontal and vertical mergers involving relatively small market shares were found to be illegal.

The Celler-Kefauver Act was passed in response to a rising concern by Congress over the beginnings of the third merger wave. An influential FTC report suggested that unless something was done, “the giant corporations will ultimately take over the country.”

While the report was criticized by many economists, it and a government defeat in a steel merger case led to the Celler-Kefauver Act. This act amended Section 7 of the Clayton Act to read:

That no corporation engaged in commerce shall acquire, directly or indirectly, the whole or any part of the stock or other share capital and no corporation subject to the jurisdiction of the Federal Trade Commission shall acquire the whole or any part of the assets of another corporation engaged also in commerce, where in any line of commerce in any section of the country, the effect of such acquisition may be substantially to lessen competition, or to tend to create a monopoly.

The principal change was to plug the asset acquisition loophole of the original law.

The 1980s witnessed a fourth merger wave during which the total value of acquisitions rose from around $50 billion in 1983 to over $200 billion by 1988. Some of the larger acquisitions included Philip Morris’s purchase of Kraft for $13.44 billion, Eastman Kodak’s purchase of Sterling Drug for $5.1 billion, and Campeau Corporation’s acquisition of Federated Department Stores for $6.51 billion. And in 1989 the closing of the largest acquisition up to that time took place—the leveraged buyout (LBO) of RJR-Nabisco by Kohlberg, Kravis, Roberts & Co. (KKR)—for $25 billion.

LBOs became very popular in the 1980s. An investor group headed by either the company’s top managers or buyout specialists (like KKR) would put up about 10 percent of the bid price in cash. They would then borrow against the company’s assets and raise, say, 60 percent in secured bank loans and 30 percent or so from “junk bonds.” (Junk bonds are very risky bonds that must promise a high return because of the risk.) The investor group then buys all the outstanding stock of the company, taking the company private.

Normally, the new owners begin selling off parts of the company to reduce its debt. The new owners also try to cut costs and lay off employees to increase profitability and to ensure that the huge new debt can be serviced. Eventually, the investor group hopes to reap large profits by taking the streamlined company public again.

The fifth merger wave is of very recent vintage, having occurred in the 1990s. In terms of the volume of acquisitions as a fraction of the economy, figure 7.1 shows that this wave exceeds even the volume of the first merger wave. One of the most spectacular of these mergers was between Time Warner, itself the product of a past megamerger, with America Online. At the time of its approval by the FTC, this merger of giants had a market capitalization in excess of $100 billion.

The merger wave of the 1990s is different from the LBOs of the 1980s. Many mergers took place in industries in the midst of deregulation, including electric power, telecommunications, and banking and financial services. Because of structural changes and enhanced competition, some mergers served to gain entry to new markets. In pharmaceuticals, mergers between direct competitors may have achieved economies of scale and scope. Downsizing and consolidation are other factors that were particularly relevant in the defense and health care industries.

Reasons for Mergers

There are many diverse reasons for mergers. It is not possible to examine all of these here; however, it is important to realize that not all mergers take place for anticompetitive reasons or to enhance productive efficiency. Of course, given the focus of this chapter on antitrust policy, it follows that these motives are our primary interest. We shall examine several representative explanations for mergers in the following subsections.

Monopoly

As we discussed earlier, the first merger wave around the turn of the twentieth century consisted of numerous “mergers for monopoly.” The classic example was J. P. Morgan’s series of mergers leading to the formation of United States Steel (now USX). It is inconceivable that such overt attempts to monopolize a market would be attempted in today’s tough antitrust environment. Nevertheless, mergers leading to lesser degrees of market power are still a serious concern.

Economies

There may be economies from combining two firms that can lead to greater profitability. These cost savings can take many forms. Two broad categories are pecuniary and real economies. Pecuniary economies are monetary savings from buying goods or services more cheaply. For example, the larger size resulting from a merger may give the combined firm bargaining strength relative to its suppliers. On the other hand, real economies represent true resource savings because of increased specialization or scale economies or sharing capital in the joint production of some goods. Clearly, real economies are socially beneficial and should be encouraged; pecuniary economies merely reflect redistributions of income between buyers and sellers.

An example of real economies in a horizontal merger is that involving three English antifriction bearing manufacturers. These three firms were able to revamp production

assignments among their plants so as to eliminate duplication and lengthen runs. These changes led to a 40 percent increase in output per employee within three years.

Of course, production economies are not the only possibilities in mergers. Marketing can be improved in certain cases through the pooling of sales forces, the use of common advertising campaigns, and so on. Economies in finance, research and development, and other areas are also possible.

**Reducing Management Inefficiencies**

Takeovers of one firm by another can lead to savings by replacing an inefficient management with a more efficient one. The idea here is based on the problem of separation of management and control in modern corporations. In brief, there is a conflict in the objectives of the owners (shareholders) and the management: shareholders seek maximum profits, while management, though concerned with profits, has other interests. Managers care about their salaries, job security, power in controlling corporate resources, size of office and staff, and so on. A simple graphical depiction of these interactions, which economists call the principal-agent relationship, is given in figure 7.2.

The principal (or owners) hires an agent (management) to run the business and earn maximum profits for the principal. Owners are uncertain about the profit possibilities that result from managerial decisions. In figure 7.2 this “profit possibility frontier” is shown as a relation between profit and output \( q \). For simplicity, think of \( q \) as a proxy variable that represents some of the variables that managers care about in addition to profit—for example, the size of the firm in and of itself. In contrast to management, the owners do not have good information on this frontier. They must try to induce the management to pick the point on it that the owners prefer. If owners could “see” the frontier, they would simply order the management to choose output \( q^* \) or be replaced. (Of course, \( q^* \) corresponds to maximum profit.)

Management has a utility function that represents its preferences for profit and \( q \). The indifference curve \( U \) shows this trade-off and indicates that management’s choice is to give up some potential profit (\( AC \) for agency cost) for a higher output \( q' \). Management picks point A at the tangency between the profit frontier and the indifference curve, since this maximizes utility. The owners would prefer point \( P \). In short, the agency cost \( AC \) is what the principal must give up because it has to contract with the agent to manage the firm, and the principal’s lack of information about the profit frontier makes it impossible simply to impose that the agent pick point \( P \). The principal, of course, tries to use various incentive devices—profit-sharing plans, stock options, and the like—to induce the agent to operate closer to \( P \), but the basic conflict in objectives means that \( AC \) will never be zero.
Michael Jensen has argued that this conflict between managers and owners played an important role in explaining the large number of acquisitions in the 1980s. He noted that it was particularly important in the oil industry. For various reasons, the industry was characterized by large cash flows in excess of the funds required for investment projects that promised positive net present values. The industry needed to reduce investments in exploration and development because excess capacity meant that these investments would have negative returns. Management, though, would see reduced exploration and development investment, and paying out the excess cash flows as dividends, as undesirable. It would tend to reduce the size of the firm, leading to reductions in the company’s exploration and development divisions and could be seen as a symptom of a dying company. Management may then avoid reducing investment even though doing so is what owners would prefer.

Horizontal Mergers

As noted earlier, a horizontal merger provides the clearest example of possible anticompetitive effects. The reason, of course, is that any such merger reduces the number of competitors and therefore raises the possibility of creating market power. However, since mergers result in the integration of the firms’ productive facilities, there is also the possibility of achieving socially beneficial cost savings.

In chapter 5 we sketched a model developed by Oliver Williamson that compares the social benefits and costs of antitrust activity. We shall reconsider that model here in more depth and examine various qualifications.

Benefits and Costs

In figure 7.3 the horizontal line $AC_0$ represents the level of average costs of two firms before combining, and $AC_1$ shows the average costs after merger. Before merger, the degree of competition is sufficient to force price down to $AC_0$. After merger, costs fall and market power is created, leading to a price increase from $P_0$ to $P_1$.

The merger results in a deadweight loss in consumers’ surplus equal to the shaded triangle $A_1$. However, there is a gain to society because of the cost savings, given by the shaded rectangle $A_2$. It represents the cost savings in producing output $q_1$ at the lower average cost.

The main result of this analysis is that a relatively small percentage cost reduction will offset a relatively large price increase, thereby making society indifferent to the merger. For example, if a merger is expected to increase price by 20 percent, only a 2.4 percent cost reduction is required to equate areas $A_1$ and $A_2$ in figure 7.3. (These particular numbers also assume


6. The calculations use $A_1 = \frac{1}{2} (\Delta p)(\Delta q)$ and $A_2 = (\Delta AC)q_1$. Substituting for $\Delta q$ from the definition of price elasticity, $\eta$, we get

$$A_1 = \frac{1}{2} (\Delta p)(\eta q_0 \Delta p) \frac{1}{p_0}.$$ 

Equating $A_1$ and $A_2$ yields

$$(\Delta AC)q_1 = \frac{1}{2} \frac{(\Delta p)(\eta q_0 \Delta p)}{p_0}.$$ 

Because $AC_0 = p_0$, divide the left side by $(AC_0)q_1$ and the right side by $p_0q_1$. The result is

$$\frac{\Delta AC}{AC_0} = \frac{1}{2} \eta \left( \frac{q_0}{q_1} \left( \frac{\Delta p}{p_0} \right) \right)^2.$$
a unitary elasticity of demand.) Table 7.1 presents the cost reductions required for alternative assumptions about price increases and demand elasticities.

There is a simple logic as to why the welfare effect of cost reductions tends to swamp those of price rises. First, note that the area of triangle $A_1$ is $\frac{1}{2}(q_0 - q_1)(P_1 - P_0)$ and that of rectangle $A_2$ is $q_1(AC_0 - AC_1)$. Obviously, if the reduction in average cost is sufficiently large relative to the rise in price, then $A_2$ is bigger than $A_1$, so the merger raises welfare; similarly, if the rise in price is sufficiently large relative to the reduction in average cost, then the merger lowers welfare. More interesting is to analyze the welfare gain when the change in cost is

Finally, assuming a constant elasticity demand curve,

$$\frac{q_0}{q_1} = \left(\frac{1}{1 + \Delta p/p_0}\right)^{-\eta}.$$  

So for $\eta = 1$, $\Delta p/p_0 = 0.2$, we get $q_0/q_1 = 1.2$, and therefore $\Delta AC/AC_0 = 0.024$. 

---

Figure 7.3
Social Benefits ($A_2$) and Costs ($A_1$) of a Horizontal Merger—Perfect Competition in Premerger Stage
comparable in size to the change in price. To consider such a case, let us suppose that they are proportional to each other, so that

\[ AC_0 - AC_1 = a(P_1 - P_0) \]

where \( a > 0 \) and we are supposing it is not too different from one. As the welfare change is the gain due to cost savings less the forgone surplus due to reduced supply, it equals

\[ aq_1(P_1 - P_0) - \frac{1}{2}(q_0 - q_1)(P_1 - P_0), \]

where we have substituted \( aq_1(P_1 - P_0) \) for \( AC_0 - AC_1 \). Collecting terms, the change in welfare is

\[ (P_1 - P_0) \left[ \left( a + \frac{1}{2} \right) q_1 - \frac{1}{2} q_0 \right]. \]

If the change in cost is comparable in size to the change in price and the rise in price is not too large (or demand is not too elastic), so that \( q_1 \) is not too much smaller than \( q_0 \), it follows that

\[ \left( a + \frac{1}{2} \right) q_1 - \frac{1}{2} q_0 > 0. \]

Hence, a merger for which the decrease in cost is comparable in size to the increase in price will always raise welfare.

If one thinks about a merger that changes price (and quantity) by a small amount and changes average cost by a small amount, then the gain from the merger is a small number, \( AC_0 - AC_1 \), multiplied by a non-small number, \( q_1 \), while the loss from the merger is less than half of two small numbers, \( (q_0 - q_1) \) and \( (P_1 - P_0) \), multiplied together. When both sides of a triangle are small and only one side of a rectangle is small, the area of the rectangle will

### Table 7.1

Percentage Cost Reduction Sufficient to Offset Percentage Price Increases for Selected Values of \( \eta \) (Elasticity of Demand)

<table>
<thead>
<tr>
<th>( \frac{\Delta P}{P} )</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>( \frac{1}{2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.43</td>
<td>0.28</td>
<td>0.13</td>
<td>0.06</td>
</tr>
<tr>
<td>10</td>
<td>2.00</td>
<td>1.21</td>
<td>0.55</td>
<td>0.26</td>
</tr>
<tr>
<td>20</td>
<td>10.37</td>
<td>5.76</td>
<td>2.40</td>
<td>1.10</td>
</tr>
</tbody>
</table>

*Source:* Computed by authors using formula in note 6.
exceed that of the triangle. Therefore, it takes a bigger rise in price to offset a fall in unit cost.

Unfortunately, this powerful conclusion is weakened if we presume the industry is not competitive prior to the merger. Consider figure 7.4, which is identical to figure 7.3 except that the premerger price, still denoted $P_0$, now exceeds the premerger cost. The loss from reduced supply is no longer a triangle but rather trapezoid $B_1$, while the gain from cost savings remains a rectangle, now denoted $B_2$. First note that the forgone surplus from reduced supply of $q_0 - q_1$ is larger because the premerger quantity was below the competitive level. Next note that a trapezoid is the sum of a rectangle and a triangle, and since the rectangle portion of the trapezoid cannot be assured of being smaller than $B_2$ when the change in price and cost are small, the argument associated with figure 7.3 doesn’t apply. In sum, when the premerger market is less than perfectly competitive, then the reduction in cost necessary to offset a rise in price isn’t quite as small as suggested by table 7.1.

One final qualification is that the analysis has focused exclusively on how a merger influences the firms’ party to the merger and has ignored its effect on the decisions of other firms.
in the market. For example, if the merged firm ends up pricing higher, this should induce the other firms to also price higher, as the rise in a competitor’s price serves to shift out a firm’s demand function.7 The rivals’ response to the merger then serves to augment the welfare losses from enhanced market power.

**The Staples–Office Depot Case**

At this point, it is useful to sketch a recent case in which the FTC obtained a preliminary injunction in a federal district court to block a proposed merger between two of the leading office superstore chains. The economics of the case nicely illustrate the benefits and costs of the Williamson model. The two office superstore chains were Staples and Office Depot, and the judge issued the injunction in June 1997 after a brief hearing in which both sides presented economic evidence.8

The relevant market was judged to be office superstore (OSS) chains, which consisted of the two merger partners plus OfficeMax. However, this judgment was strongly contested by the defendants, who argued that non-OSS firms such as Wal-Mart, Kmart, and Target should be included in the relevant market. The defendants argued that well over 80 percent of office supplies are purchased through these non-OSS outlets. However, the FTC maintained that OSS firms were different from non-OSS firms because they carried a broad range of office supplies—up to 7,500 items, compared to no more than 2,400 items carried by, say, Wal-Mart. This large selection permits one-stop shopping, which may be very appealing to small business customers. The FTC also argued that the defendants’ own documents defined OSS firms as “the competition.”

Next, the FTC produced both documents obtained from the firms and econometric studies showing that prices were higher in markets with fewer superstores. Referring to table 7.2, markets that had both Staples and Office Depot charged 11.6 percent lower prices than markets with only Staples. And adding Office Depot to when Staples and OfficeMax were already present lowered prices by 4.9 percent. Furthermore, direct competition among Staples and Office Depot was increasing over time. Among those markets for which Staples had a store, it was in competition with Office Depot in 46 percent of them in 1995 but in 76 percent by 2000.

Sophisticated econometric studies by both sides were submitted that attempted to control for factors other than just the existence of the other superstores, for example, determinants of cost and demand. The results were that the FTC found a 7.3 percent price increase while

---

7. For more details, see figure 5.5 and the accompanying text.
the defendants found a much smaller increase of 2.4 percent, *absent the effect of any cost savings*.

Hence, the anticompetitive effect of a positive price increase was supported by both sides, although the magnitudes were quite different. However, the parties then attempted to take into account cost-saving efficiencies, and the disagreement was considerable. The FTC argued that 43 percent of the savings claimed by the defendants would be achievable without the merger and should therefore be excluded. For various other reasons, additional savings estimates were regarded as unreliable by the FTC. Finally, the FTC argued that only those cost savings that would be passed on to consumers should be counted. Their econometric evidence indicated that this figure was only 15 percent.

The upshot was that the savings estimate of the FTC was 1.4 percent of sales. After taking into account the pass-through rate and cost savings, the FTC found that prices would increase by 7.1 percent. That is, the net price increase would be 7.3 percent less—$(0.15 \times 1.4)$ percent, or 7.1 percent. The defendants found a net price decrease of 2.2 percent. The judge was persuaded by the FTC that the merger would be anticompetitive and issued the injunction to bar the merger. It was unclear from his opinion how much weight he gave to the sophisticated econometric analysis as compared to the raw documents containing management studies and forecasts.

**Analysis of Market Power Effect**

It is worth exploring in some detail how a merger may raise price. Without altering the way in which firms compete, a merger can raise price by enhancing each firm’s market power. A merger can also alter the character of competition by inducing tacit collusion or, worse yet, leading to the formation of a price-fixing cartel. Let us consider each of these mechanisms by which a merger may lead to higher prices.

To consider the enhancement of market power while maintaining competition, let us utilize the Cournot model of chapter 5. There it was stated that, in equilibrium, firms produce so that the price-cost margin is negatively related to the number of firms, $n$, and the (absolute value of the) demand elasticity, $\eta$, according to the following formula:
\[ \frac{P - MC}{P} = \frac{1}{n\eta}. \]

A merger, by reducing the number of firms, raises the extent to which price exceeds cost (assuming the merger does not affect the cost functions). To understand the intuition, a firm that considers supplying \( \Delta q \) units more (where \( \Delta q \) is small) will earn profit of approximately \((P - MC)\Delta q\) on each of those additional units but will lose profit equal to \(q\Delta P\) on its preexisting supply of \(q\) because its higher output reduces the price by \(\Delta P\) that those units fetch. The former effect induces a firm to produce more while the latter effect induces it to produce less, and a firm’s optimal quantity is one that balances those two forces. Merger affects this calculus because the size of \(q\Delta P\) depends on how much a firm is producing in equilibrium. With a merger resulting in fewer firms, each firm produces more, so that \(q\) is higher, which makes firms more concerned with maintaining a high price. While firms expand their supply in response to the merger, it is constrained by this effect, so that total industry supply is lower and thus price is higher.

Now let us turn to the possibility that a merger causes firms to shift from competing to colluding. In the Cournot model, collusion means firms coordinate their production with the intent of raising the price-cost margin above \(\frac{1}{n\eta}\). The difficulty in doing so is that it requires each firm to constrain its quantity below that which maximizes current profit. Each firm is then tempted to cheat on the agreement by producing more. To deter such cheating, firms have an understanding that such cheating will be punished by a firm’s rivals responding aggressively and thereby lowering the future profit stream of the deviator. This may take the form of a temporary price war or simply the discontinuance of collusion so that the price-cost margin reverts back to the lower level of \(\frac{1}{n\eta}\). The stability of collusion then rests on the reduction in future profits from cheating exceeding the rise in current profit, so that each firm prefers not to cheat. If that condition holds, then collusion can be sustained and otherwise it cannot.

Merger can influence that condition through its effect on the number of firms. The short-run gain from cheating is typically lower when there are fewer firms. With fewer firms, each firm’s market share at the collusive outcome is higher which means that the gain in market share by cheating is not as great. Thus, merger reduces the gain to cheating. However, merger can also reduce the loss to cheating. Suppose the threatened punishment is reversion to competition so that a firm that anticipates cheating can expect a future price-cost margin of \(\frac{1}{n\eta}\). Since this price-cost margin is higher when \(n\) is smaller, merger weakens the punishment from cheating. In netting out these two counterbalancing effects, it has generally been found that fewer firms reduce the gain to cheating more than it reduces the loss to cheating—
collusion is easier! A merger can then lead to a substantial increase in price if it makes collusion stable where before it was unstable.\(^9\)

**Analysis of Cost Savings Effect**

Having explored how the market power effect operates, we next turn our attention to mechanisms by which mergers create cost savings. Because greater cost savings are more likely to result in the merger being approved by the authorities, the merger participants have an incentive to overstate cost savings, and unfortunately, authenticating such claims is tricky. Though the exact source of cost savings is apt to be idiosyncratic to each merger, there is one general type we can address, which is the more efficient allocation of industry supply. Specifically, a merged firm can coordinate output across production units so as to reduce total cost.

To see how the argument works, suppose a merger leaves technology and input prices unchanged, so cost functions are unchanged. Firms 1 and 2 propose merging where their marginal cost curves are as depicted in figure 7.5; note that firm 2 is more efficient than firm 1. Prior to the merger, firm 1 is producing \(q_1\) units with marginal cost \(MC_1(q_1)\) and firm 2 is producing \(q_2\) units with marginal cost \(MC_2(q_2)\). By virtue of having a lower marginal cost curve, firm 2 produces more.

To keep our analysis of cost savings untainted by market power effects, suppose the merged firm would keep its total supply unchanged at \(Q' = q_1 + q_2\), and think about it allocating \(Q'\) across its two production units (previously, firms 1 and 2) so as to minimize total cost. The claim we want to make is that, as long as the cost-minimizing solution has both production units being active, it must result in their marginal costs being equalized. To prove this claim, let us suppose marginal costs are not equal and then show that supply can be reallocated so as to reduce total cost. Let production unit 1’s quantity be denoted \(q_0\), which implies production unit 2 produces \(Q' - q_0\), and suppose their marginal costs are unequal and, furthermore, \(MC_1(q_0) > MC_2(Q' - q_0)\). Now let’s move a single unit of supply from production unit 1 to unit 2. This reduces the cost of unit 1 by \(MC_1(q_0)\) and raises the cost of unit 2 by \(MC_2(Q' - q_0)\), so that total cost declines by \(MC_1(q_0) - MC_2(Q' - q_0)\), and thus we have lowered total cost! Whenever marginal costs are unequal, there is always a way in which to rearrange supply so as to lower total cost. Thus, starting from quantities \(q_1'\) and \(q_2'\), the merged firm can continuously shift supply from unit 1 to unit 2, lowering cost as it does so, until the cost-minimizing quantities of \(q_\bar{\bar}\) and \(Q' - q_\bar{\bar}\) are reached, where, as shown in figure 7.5, marginal costs are equalized.\(^10\)

\(^9\) A more detailed analysis of collusion is provided in chapter 5.

\(^{10}\) This argument for a cost-minimizing solution to have equalization of marginal costs presumes that both quantities are positive, for if one unit’s quantity is zero and its marginal cost is higher, it is not possible to further shift supply away from it and to the lower cost unit.
In light of the preceding analysis, it is interesting to ask whether a more efficient allocation of supply can lower price even with a market power effect at play. Unfortunately, it cannot. If the merger is profitable (that is, the profit of the merged firm exceeds the premerger profits of the firms party to the merger) and the cost functions are unchanged (as presumed in the previous discussion), then price must rise, even if cost is lower.\footnote{This is proven in Joseph Farrell and Carl Shapiro, “Horizontal Mergers: An Equilibrium Analysis,” \textit{American Economic Review} 80 (March 1990): 107–26.}

As a final remark, can we infer anything about the welfare effect of a proposed merger if competitors are challenging it? Remaining within the Williamson model, all that a competitor cares about is whether the merged firm will produce more or less compared to the combined premerger output of the firms engaged in the merger. If merger causes the merged firm’s output to fall (rise), then the competitors are better (worse) off, since there is more (less) demand left for them to supply. So suppose the merged firm produces less (because cost

\begin{figure}[h]
\centering
\begin{tikzpicture}
\begin{axis}[
view={0}{90},
axis x line=bottom,
axis y line=left,
axis line style={-},
axis x discontinuity=crunch,
axis x line style={->},
axis y line style={->},
axis x line style={very thick},
axis y line style={very thick},
xtick={0,1,2},
xticklabels={$q$, $q'$, $q''$},
ytick={0,1,2},
yticklabels={$0$, $q'$, $q''$},
]
\addplot coordinates{(0,0) (1,1) (2,2)};
\addplot coordinates{(0,0) (1,1) (2,2)};
\end{axis}
\end{tikzpicture}
\caption{Analysis of Cost Savings from a Merger}
\end{figure}
savings are small or nonexistent). The stronger demand faced by competitors induces them to supply more, but, on net, total industry supply can be shown to fall, so that price is higher. In that case, competitors will not challenge the merger, because their profit is higher. However, consumer surplus is lower—as price is higher—and the merger is likely to reduce welfare. Analogously, if the merged firm produces more (due to sufficiently large cost savings), competitors will respond by contracting their supply and, on net, price is lower. Competitors are clearly worse off and may challenge the merger, but the merger is welfare-enhancing as price is lower and there are cost savings. Thus, if competitors challenge a merger, then the merger must improve welfare and should be approved!

An important caveat to the above policy statement is that if the framework is richer than the Williamson model, it is possible that a merger can reduce competitors’ profits and reduce welfare. This can occur when the merged firm engages in anticompetitive practices such as those that raise competitors’ costs. If competitors have higher costs, then their profits are likely to be lower, and it can also result in price being higher, so that consumers are worse off as well. Such an argument was made in connection with the 1985 acquisition of Pan American World Airway’s Pacific Division by United Airlines. Northwest Airlines challenged the merger, and the purported device by which the merger would raise its cost was the proprietary reservation system of United used by travel agents.

Effects of Airline Mergers

We conclude this section with a short discussion of an empirical study of the effect on air fares of fourteen airline mergers that took place in 1985–1988. The mergers included Delta and Western, American and Air Cal, U.S. Air and Piedmont, and Eastern and Texas, among others.

As we will discuss in detail in chapter 17, airlines were regulated as to fares and entry and exit until the passage of the Airline Deregulation Act of 1978. By 1985, the airlines had been through a period of a great deal of new entry, followed by price competition and bankruptcies. There were many mergers in the late 1980s as a movement toward consolidation took place.

The focus of the study was airline routes, each of which was considered to be a separate market. Hundreds of routes were affected by each merger, and, on average, there were 196 unaffected routes for each route affected by a merger. The empirical strategy is to compare airfare changes on routes affected by a merger to airfare changes on unaffected routes. The

12. A more detailed discussion of raising rivals’ costs is provided in chapter 8.
unaffected routes were used as a control group to capture industry-wide factors such as changes in fuel costs and seasonal variations in demand that influence airfares. That is, suppose that fares on a merger-affected route (say, New York–Chicago) rose by 5 percent, but that fares on an unaffected route of similar distance (say, Chicago–Philadelphia) also rose by 5 percent. One could conclude that the merger had no effect on fares and that the fare increase was due, perhaps, to higher fuel costs.

The authors recognized that any change in airfares of merging firms reflects the joint effect of increased efficiency, which would tend to lower fares, and increased market power, which would increase fares. Hence, they argued that the direction of the fare change, if any, would indicate which of these effects dominates.

A major finding was that, over the period from merger talks through merger completion, the merging firms increased airfares by an average of 9.44 percent relative to other routes unaffected by the merger. Hence, they concluded that market power effects dominated efficiency gains. More striking is that rival firms—that is, nonmerging firms serving the merger-affected routes—increased their fares by an average of 12.17 percent!

Cases

Before we describe the leading horizontal merger cases, it is useful to emphasize that the preceding benefit-cost analysis is the approach many economists advocate; it is not the way courts necessarily evaluate the legality of mergers. As we shall see, courts have not followed the Williamson method.

**Brown Shoe Case of 1962**

As discussed earlier, Section 7 of the Clayton Act was amended in 1950 to plug the “asset loophole.” The Supreme Court’s first ruling under the new Section 7 came in the *Brown Shoe* case of 1962.¹⁵ This case involved the merger of Brown, the fourth largest manufacturer of shoes in the United States, with about 4 percent of the market, and G. R. Kinney Company, the twelfth largest, with a 0.5 percent share. Both companies were also engaged in shoe retailing. Brown had 2.1 percent of the national retail market and Kinney had 1.6 percent. Though the case involved both horizontal and vertical dimensions, we shall deal only with the horizontal aspects here.

The first step the Court took was to discuss the definition of the relevant retail shoe market.¹⁶ That is, should total retail shoes sold in the United States be the market, or should men’s shoes and women’s shoes in large cities be separate relevant markets? It clearly matters,


¹⁶. The horizontal dimensions of the shoe-manufacturing market were not at issue before the Supreme Court. The district court found that the merger of Brown’s and Kinney’s manufacturing facilities was economically too insignificant to be illegal, and the government did not appeal the lower court’s decision.
since the two firms had a combined share of only about 4 percent of the national market, while they had 57 percent of the market for women’s shoes in Dodge City, Kansas. According to the Court,

The “area of effective competition” must be determined by reference to a product market (the “line of commerce”) and a geographic market (the “section of the country”).

The outer boundaries of a product market are determined by the reasonable interchangeability of use or the cross-elasticity of demand between the product itself and substitutes for it. However, within this broad market, well-defined submarkets may exist which, in themselves, constitute product markets for antitrust purposes. The boundaries of such a submarket may be determined by examining such practical indicia as industry or public recognition of the submarket as a separate economic entity, the product’s peculiar characteristics and uses, unique production facilities, distinct customers, distinct prices, sensitivity to price changes, and specialized vendors.

Applying these considerations to the present case, we conclude that the record supports the District Court’s findings that the relevant lines of commerce are men’s, women’s, and children’s shoes. These product lines are recognized by the public; each line is manufactured in separate plants; each has characteristics peculiar to itself rendering it generally noncompetitive with the others; and each is, of course, directed toward a distinct class of customers.

Next, the Court turned to the geographic dimensions of the market:

The criteria to be used in determining the appropriate geographic market are essentially similar to those used to determine the relevant product market. Congress prescribed a pragmatic, factual approach to the definition of the relevant market and not a formal, legalistic one. The geographic market selected must, therefore, both “correspond to the commercial realities” of the industry and be economically significant. Thus, although the geographic market in some instances may encompass the entire Nation, under other circumstances it may be as small as a single metropolitan area.

The Court then found that the relevant geographic markets were “every city with a population exceeding 10,000 and its immediate contiguous surrounding territory in which both Brown and Kinney sold shoes at retail through stores they either owned or controlled.”

The Court found some high combined market shares. For example, the maximum was the 57 percent in Dodge City noted earlier. However, this was atypical. The most important statistic seemed to be that in 118 separate cities, the combined share of one of the relevant product lines exceeded 5 percent.

If a merger achieving 5 percent control were now approved, we might be required to approve future merger efforts by Brown’s competitors . . . . The oligopoly Congress sought to avoid would then be furthered and it would be difficult to dissolve the combinations previously approved.

Thus the Court gave a lot of weight to its reading of the intent of Congress. “What Congress saw as the rising tide of economic concentration . . . [and provided authority] for arresting mergers at a time when the trend to a lessening of competition . . . was still in its incipiency.”
It is useful to contrast the Court’s opinion in *Brown Shoe* with the Williamson trade-off analysis described earlier. The small market shares found hardly seem likely to give rise to monopolistic price increases. However, the Court’s view was to stop a trend toward increased concentration in its incipiency. It noted, for example, that there was a 10 percent decrease in the number of shoe manufacturers between 1947 and 1954.

What about cost savings? In *Brown Shoe* the Court recognized that integrated operations could create efficiencies; however, such efficiencies were not regarded to be as important as maintaining a “decentralized” industry.

The retail outlets of integrated companies, by eliminating wholesalers and by increasing the volume of purchases from the manufacturing division of the enterprise, can market their own brands at prices below those of competing independent retailers. Of course, some of the results of large integrated or chain operations are beneficial to consumers. But we cannot fail to recognize Congress’ desire to promote competition through the protection of viable, small, locally owned businesses. Congress appreciated that occasional higher costs and prices might result from the maintenance of fragmented industries and markets. It resolved these competing considerations in favor of decentralization. We must give effect to that decision.

In a 1967 decision in a conglomerate merger case, the Court made an even stronger statement concerning cost savings: “Possible economies cannot be used as a defense to illegality.”  Various interpretations are possible. One is that the measurement problem is so great that cost savings cannot be used practically as a defense in a judicial process. An alternate interpretation, which almost all economists find distasteful, is that cost savings are harmful because they lead to the failure of small, inefficient retailers. That this latter interpretation is not unreasonable was illustrated by the attempt by United Technologies in 1978 to take over Carrier Corporation. Carrier tried to avert being taken over by claiming “that the merger would make Carrier a more technologically progressive, potent competitor.”

A third interpretation of the *Brown Shoe* decision is that antitrust has multiple objectives with economic efficiency being just one. Thus the objective of maintaining many small retailers must be balanced (somehow) against the higher costs to consumers. One antitrust expert, Robert Bork, has disputed the view that Congress intended multiple goals for antitrust:

In Brown Shoe, in fact, the Supreme Court went so far as to attribute to Congress a decision to prefer the interests of small, locally owned businesses to the interests of consumers. But to put the matter bluntly, there simply was no such congressional decision either in the legislative history or in the text of the statute. . . . The Warren Court was enforcing its own social preferences, not Congress.

Cases of the 1960s and 1970s

Two merger cases were decided by the Supreme Court in 1964 that we shall briefly describe. One involved a merger between a tin can manufacturer and a glass bottle maker.20 In this case the Court found the relevant market to be metal cans and glass bottles combined. The two companies, Continental Can and Hazel-Atlas Glass, had 22 percent and 3 percent shares, respectively, and the Court found these to be too high.

The second 1964 case was a merger between Alcoa, a producer of aluminum electrical conductor cable, and Rome Cable, a specialist in copper cable.21 However, Rome did produce a small amount of aluminum cable. In this case, selecting from many possible choices for the relevant market, the Court chose a market that included only aluminum cable, and Alcoa’s 27.8 percent and Rome’s 1.3 percent were found to be illegal.

In the first case, tin cans and glass bottles were considered substitutes, and in the second case, aluminum cable and copper cable were not considered substitutes. According to one expert these two decisions are logically inconsistent. Also, the Alcoa-Rome and Continental-Hazel-Atlas decisions exhibit a different sort of consistency: the consistent willingness of the courts to accept market definitions that resolve inherent doubts on the side of preventing mergers with possible anti-competitive effects.22

The Von’s decision in 1966 has been especially significant as a precedent.23 Von’s was the third largest grocery chain in the Los Angeles area in 1960 when it acquired the sixth-ranked chain, Shopping Bag Food Stores. The combined firm had only 7.5 percent of the market and was second to Safeway Stores.

Despite the low share of 7.5 percent, the Court found the merger to be illegal. The emphasis in the decision was on the trend toward fewer owners in single grocery stores in Los Angeles, as they had declined from 5,365 in 1950 to 3,818 in 1961. Also, from 1953 to 1962 the number of chains with two or more grocery stores increased from 96 to 150. According to Justice Black, “The basic purpose of the 1950 Celler-Kefauver bill was to prevent economic concentration in the American economy by keeping a large number of small competitors in business.”

In a 1974 case the Supreme Court handed the Justice Department its first defeat on a market definition issue.24 The issue was whether coal production or uncommitted coal reserves were the appropriate variable for calculating market shares. With its composition changed by the
addition of new conservative judges, the Court found that market shares based on coal reserves were not significant enough to find the merger unlawful.

**Hart-Scott-Rodino Act Enforcement**

In the last few decades there have been no Supreme Court cases concerning horizontal mergers, although there continue to be lower court cases. One reason stems from the passage of the Hart-Scott-Rodino Act in 1976, which has led to an important alternative to court cases. That act provides for the prior notification of large proposed mergers to both the FTC and the Antitrust Division of the Department of Justice. One of the agencies then assumes responsibility for reviewing each merger for possible anticompetitive effects before it can actually occur.

In many cases, rather than going to court, some type of settlement is negotiated between the firms and the agency involved. These negotiations involve meetings between lawyers, expert economists, and other officials from all parties. The intent of the agreement is to alter the merger to eliminate anticompetitive features. In some cases the agreement requires the agency to monitor the firms for several years to insure that the outcome of the merger is acceptable. For example, the FTC approved a 1994 merger between Eli Lilly, a pharmaceutical manufacturer, and PCS, a pharmacy-benefits manager, but it imposed complex rules and reporting requirements so that PCS benefit plans would not favor Lilly products. In 1997 the FTC launched a reinvestigation of the merged firms to evaluate their performance. In some ways, these negotiated settlements have shifted a large fraction of the antitrust process from courtrooms to a form of regulation.

For another example, consider the testimony of Joel Klein, head of the Antitrust Division, regarding a 1997 merger between Raytheon and Hughes Aircraft. The merger was permitted to go forward, but only under some important conditions.

We insisted on a broad range of remedies, including divestiture of two defense electronics businesses, to preserve competition in sophisticated technology for U.S. weapons systems. For another weapons system for which both Hughes and Raytheon were competing for development and production at the time of the merger, a new antitank missile for the Army, we required Raytheon to establish procedures to prevent the two competing teams of employees from sharing information with each other, thereby preserving the independence of these teams in competition. And, as an additional remedy, Raytheon agreed to set firm prices with the Air Force on certain air-to-air missiles for which Raytheon and Hughes had been giving competing bids before the merger, saving the Air Force $180 million over the next four years.²⁵

The current attitude of the courts toward horizontal mergers has been summarized as follows:

Under the current case law, market shares at the time of merger are only the beginning point. They are used to establish a presumption of illegality. The precise level at which that presumption is triggered, however, has changed remarkably over the years. As recently as the late 1960s, aggregate market shares of under 10 percent in a relatively unconcentrated Los Angeles grocery store market were sufficient to support a finding of unlawfulness. Since that time, the triggering point for illegality has risen steadily. In the 1970s, acquisitions that resulted in aggregate shares of approximately 10–20 percent could be held unlawful. More recently, the courts may well approve aggregate concentration in excess of 25 percent.26

U.S. Department of Justice Merger Guidelines

The U.S. Department of Justice (DOJ) and the Federal Trade Commission (FTC) were notified of 4,926 prospective mergers in 1999–2000, but only eighty them were perceived as posing competitive problems. Out of those eighty, thirty-six were settled with some agreement between the firms and the agency in charge and thirty-eight were abandoned or restructured by the firms’ party to the merger. Only six were actually challenged in court by the DOJ or FTC, with three ultimately being abandoned, one ending with a consent decree, and two blocked by the court.27

As those numbers reveal, only a small fraction of mergers are perceived as troublesome from an antitrust perspective. In order to provide guidance to firms as to what types of mergers are likely to be challenged, the DOJ issued merger guidelines in 1968. The emphasis was on the market shares of the firms involved in the proposed merger and when they would be large enough so as to create concern about market power. The guidelines were substantially revised in 1982 which, among many operational changes, gave emphasis to the role of entry conditions; that is, whether entry or the threat of entry would constrain the merged firm from utilizing market power. A substantive revision of these guidelines occurred in 1992, at which time they were co-issued by the FTC.28

A major contribution of the guidelines is in defining the relevant antitrust market. As we noted in the discussion of the cases earlier, delimitation of the relevant market is vital in determining the legality of a merger. The DOJ defines the market conceptually as follows:

A market is defined as a product or group of products and a geographic area in which it is sold such that a hypothetical profit-maximizing firm, not subject to price regulation, that was the only present or future producer or seller of those products in that area likely would impose at least a “small but significant and nontransitory” increase in price, assuming the terms of sale of all other products are


27. These statistics are from Russell W. Damtoft “Merger Control in the United States of America” (paper presented at the International Conference on Enhancing Competition Policy, Brazil, May 2001). Damtoft served in the International Antitrust Division of the FTC.

held constant. A relevant market is a group of products and a geographic area that is no bigger than necessary to satisfy this test.

This has come to be known as the SSNIP test, which is an acronym for “small but significant and nontransitory increase in price.” Operationally, the test is taken to be a 5 percent increase in price lasting for one year, but it serves only as a benchmark. An example should be instructive. For simplicity, assume that the product, gravel, has no close substitutes. Hence we can focus on the geographic dimension of the market. Furthermore, assume that a merger between two gravel suppliers located in Centerville is being examined by the DOJ. At the present time, the market in Centerville is competitive and the price is $100. Furthermore, assume that the cost per unit by suppliers everywhere is $100, and that it costs 25 cents to haul a ton of gravel one mile. The issue is to determine the geographic limits of the market. Should it stop at the city limits of Centerville? Given that it is costly to haul gravel from suppliers outside the city limits, how many miles should the market extend beyond the city limits?

In figure 7.6 the horizontal axis shows Centerville located along a “highway” that extends to the east and west, where the numbers indicate the number of miles from Centerville’s city

![Figure 7.6](image-url)  
Geographic Market Definition
limits. The height of the vertical line at Centerville represents the $100 competitive price at that location. The lines sloping upward to the east and west have vertical heights equal to $100 plus the miles away from Centerville multiplied by 25 cents per mile. Hence, at a distance of 20 miles to the east, the height of the sloping line is $100 + (0.25)(20) = $105. This can be interpreted as the delivered price that a supplier 20 miles to the east could sell gravel for in Centerville.

The guidelines provide the answer to the market definition problem. The market should include all suppliers who would need to be part of a hypothetical cartel such that the price in Centerville could be raised by, say, 5 percent, to $105, on a nontransitory basis. If it costs 25 cents per mile to transport a ton of gravel, then a supplier 20 miles away could sell in Centerville at $105. Hence, all suppliers within 20 miles of Centerville should be part of the market. Notice that if the price increase is taken to be 10 percent, then the market boundaries should extend out to 40 miles, implying a market with more suppliers.

The example makes the important point that one must decide on the percentage price increase before the market boundaries can be determined. In short, market power is a continuous variable—there is no magical way to determine a market without incorporating some standard. In fact, there were arguments at the DOJ as to whether a 5 percent or a 10 percent increase should be specified. Although 5 percent was specified, the guidelines point out that it is not “an inflexible standard that will be used regardless of the circumstances of a given case.” A higher price increase of course means a broader market and thereby permits more mergers to slip by unchallenged. (Two particular merging firms will have lower market shares in a 40-mile market than in a 20-mile market and, as we shall explain, the guidelines are more likely to endorse mergers that involve lower market shares.) A lower price increase means possibly prohibiting relatively harmless mergers or those that may promote efficiency.

In applying the guidelines, the DOJ uses the HHI concentration index, which was explained in chapter 6. The HHI is used to determine which mergers qualify as being safe from challenge because they are unlikely to have adverse competitive effects. Those that do not fall into these “safe harbors” are then analyzed further with respect to entry conditions, efficiency considerations, and other relevant factors so as to determine if they will be challenged.

There are three categories of market concentration: Unconcentrated (HHI below 1,000), Moderately Concentrated (HHI between 1,000 and 1,800), and Highly Concentrated (HHI above 1,800). For example, if the Census Industries in table 6.2 were relevant markets, “Cereal Breakfast Foods,” with an HHI of 2,774, would be Highly Concentrated, while “Audio and Video Equipment,” with an HHI of 415, would be Unconcentrated.

The safe harbors are shown as the shaded region in figure 7.7. That is, the DOJ will consider not only the postmerger market concentration but also the increase in concentration...
resulting from the merger. Hence, all mergers with postmerger HHI values of 1,000 or less are safe. Mergers that produce an increase in HHI of less than 100 points in Moderately Concentrated markets postmerger are also safe. As an example, consider a market shared equally by eight firms, which would then have an HHI of 1,250. If two firms merged, the HHI would increase by 312.5. Hence, in this case the merger would be viewed as likely to have adverse competitive effects, and the DOJ would study other factors to determine whether a challenge should be made.

Finally, mergers that produce an increase in HHI of 50 points or less in Highly Concentrated markets (1,800 or above) are safe. As an example, consider a market shared equally by five firms, which would then have an HHI of 2,000. If two firms merged, the HHI would increase by 800, and this would fall outside of the safe harbor. Hence, the DOJ would presume that the merger is likely to create or enhance market power or facilitate its exercise. Unless further analysis showed that entry would be easy or that important efficiencies would be created, the merger would most likely be challenged.

As an application of this procedure, in 1986 the FTC challenged a proposed merger between Coca-Cola and Dr. Pepper. They argued that it would increase the HHI for the carbonated soft drink industry by 341 points to a level of 2,646. This clearly violated the guidelines, and the FTC was successful in halting the merger.\(^\text{29}\)

Conglomerate Mergers

Conglomerate mergers involve firms that are not sellers in the same market nor do they stand in a buyer-seller relationship. Our earlier example of a “pure” conglomerate merger was the merger between the cigarette manufacturer R. J. Reynolds and Burmah Oil and Gas. Two other categories of conglomerate mergers discussed were product extension (PepsiCo and Pizza Hut) and market extension (Walmart and Woolco Canada). These latter two categories are more likely to be challenged by the antitrust authorities. The reason is a concern for reducing potential competition. We will consider the potential competition theory shortly. First, however, we turn to some of the efficiency-enhancing characteristics of conglomerate mergers.

Potential Benefits

One conglomerate of historical importance is International Telephone and Telegraph (ITT). In 1960, ITT was a large manufacturer of telecommunication equipment and an operator of telephone systems. It embarked on a diversification program in that year largely through conglomerate mergers. Along the way, it acquired such diverse firms as Hartford Fire Insurance, Continental Baking, Sheraton Hotels, Avis Rent-a-Car, Canteen (vending machines), and over a hundred more. By 1980, ITT had become the thirteenth largest industrial firm in the United States. Although ITT is an extreme example of conglomerate merger growth, it certainly raises the issue of whether such firms are beneficial or harmful to the economy.

Of course, conglomerate firms vary widely in their internal organizational structure. In some, the central management staff may be quite knowledgeable about the operating problems of each division. In others, the top management may be concerned only with the profit and loss statements of its components. For this reason, it is hazardous to generalize about the efficiency properties of conglomerates.

Nevertheless, it can be argued that certain conglomerate organizations are superior to the capital market in allocating investment funds. The idea is that the top management of a conglomerate has access to information and possesses the controls to change the operations of its division. Banks and stockholders are much further removed from the internal operations of firms. Also, as Williamson has observed,

The general management and its support staff can perform a further capital market function—assigning cash flows to high yield uses. Thus, cash flows... are exposed to an internal competition. ... Moreover, because the costs of communicating and adapting internally are normally lower than would be incurred in making an investment proposal to the external capital market, it may be practicable to... [employ] a sequential decision process (in which additional financing is conditional on prior
In short, such conglomerates serve effectively as “miniature capital markets.”

A second benefit of conglomerate mergers, especially if other mergers are restricted by antitrust authorities, is the takeover threat. The idea is that managements are constantly being pressured to perform efficiently by the threat of a takeover by another firm. That is, suppose that firm A is run by a slack or incompetent management team making only 80 percent of its potential profits. Firm B will then have an incentive to buy A, fire its management, and boost A’s profits. Of course, the stockholders of firm A also have an incentive to fire the inefficient management; however, the costs of organizing a sufficiently powerful group of stockholders to carry out this plan may be too high. Hence, conglomerate mergers can have beneficial results in providing incentives for managerial efficiency.

**Anticompetitive Effects and Cases**

Numerous anticompetitive claims have been made against conglomerate mergers. They have been charged with creating the opportunities for reciprocal dealing and predatory pricing, producing politically undesirable giant size, and eliminating potential competition.

Reciprocal dealing refers to the practice of buying from a supplier only on the condition that the supplier buys from you. For example, Consolidated Foods tried to get its suppliers to buy their onion and garlic needs from its newly acquired Gentry division. The competitive effects caused by this practice are controversial; for instance, some argue that reciprocity may inhibit competitive pricing, while others argue that it can actually invigorate competition.

Predatory pricing refers to deliberately pricing below cost to drive out rivals, and raising the price to an excessive level after their exit. This tactic, as well as reciprocal dealing, is not confined to conglomerate firms; also, predatory pricing will be treated in depth in chapter 9. Furthermore, the merger guidelines cite the elimination of potential competition as their only concern regarding conglomerate mergers. For these reasons, we will restrict further discussion here to potential competition.

First, we should clarify the meaning of a potential competitor. That is to say, how can one distinguish between an actual and a potential competitor? Clearly, if a firm could quickly shift from producing zero widgets to producing and selling positive quantities, arguably it could fit either category. To resolve this definitional problem, the guidelines have offered the following arbitrary rule. An actual competitor is one that “has existing productive and distributive facilities that could easily and economically be used to produce and sell the relevant product within one year in response to a small but significant and nontransitory increase in price.” Hence, actual competitors are included as part of the relevant market. On the other

---

hand potential competitors are those that “must construct significant new productive or distributive facilities in order to produce and sell the relevant product.”

We shall use the Procter & Gamble case\(^{31}\) of 1967 as an example of how the elimination of potential competition can be effected by conglomerate merger. Procter & Gamble, the dominant producer of soaps and detergents in the United States, with 1957 sales of $1.1 billion, acquired Clorox, with sales of only $40 million. Clorox, however, was the leading manufacturer of household liquid bleach, with 49 percent of the national market.

The Supreme Court held that the merger violated Section 7 of the Clayton Act for several reasons. Here, we confine the discussion to the Court’s opinion regarding potential competition.

The Commission also found that the acquisition of Clorox by Procter eliminated Procter as a potential competitor. . . . The evidence clearly shows that Procter was the most likely entrant. . . . Procter was engaged in a vigorous program of diversifying into product lines closely related to its basic products. Liquid bleach was a natural avenue of diversification since it is complementary to Procter’s products, is sold to the same customers through the same channels, and is advertised and merchandised in the same manner. . . . Procter had considered the possibility of independently entering but decided against it because the acquisition of Clorox would enable Procter to capture a more commanding share of the market.

It is clear that the existence of Procter at the edge of the industry exerted considerable influence on the market. First, the market behavior of the liquid bleach industry was influenced by each firm’s predictions of the market behavior of its competitors, actual and potential. Second, the barriers to entry by a firm of Procter’s size and with its advantages were not significant. There is no indication that the barriers were so high that the price Procter would have to charge would be above the price that would maximize the profits of the existing firms. Third, the number of potential entrants was not so large that the elimination of one would be insignificant. Few firms would have the temerity to challenge a firm as solidly entrenched as Clorox.

Thus, by acquiring Clorox, Procter removed itself as a potential competitor. Because Procter was the most likely entrant and perhaps unique in its capability to enter, the Court viewed this merger as removing an important constraint on pricing policies in the bleach market. It is also generally believed by antitrust authorities that other forms of entry by Procter would have been preferable to acquisition of the leading firm. That is, more procompetitive alternatives would have been either new entry or entry by a so-called “toehold” acquisition of a small competitor.

In the merger guidelines, several criteria are given that must be met before a potential competition merger will be challenged:

1. The HHI must exceed 1,800.
2. Entry must be difficult.

3. The eliminated potential competitor must have been one of only three or fewer firms having comparable advantages in entering the market.

4. The acquired firm’s market share must be at least 5 percent.

Although the potential competition principle is certainly correct in theory, there are difficult problems involved in establishing empirically who the potential competitors are and what their respective costs are. Even more difficult to establish is whether a firm forbidden to enter by acquiring the leading firm might then enter by building new capacity or by acquiring a smaller firm.

**Summary**

A merger between competitors represents an obvious and immediate creation of market power. This chapter has focused on how such mergers may impact social welfare and the policy challenges faced in analyzing them. Using the Williamson model, the simplest analysis involves weighing any cost savings from a merger against any price increases emanating from increased market power. This enhanced market power may come from each firm’s quantity decision having a bigger impact on price—which leads them to want to constrain supply more—or may come from increased concentration, resulting in competition being replaced with some form of collusion. Cost savings are a bit harder to identify and are likely to vary with each case. One general source of cost savings is when the premerger firms have production units with different levels of efficiencies, as a merged firm can more efficiently allocate supply than the market.

Antitrust law and policy have evolved considerably. The Sherman Act proved quite limited in the role it created for the federal government to intervene in merger activity. Indeed, it was under the watch of the Sherman Act at the turn of the twentieth century that the first merger wave led to the creation of monopolies or near-monopolies in many important industries. While the Clayton Act of 1914 sought to provide the instruments to constrain merger activity, an important loophole permitted many mergers so that new oligopolies were created through mergers and acquisitions in the 1920s. With the amendment of the Clayton Act in the 1950s, along with a policy sentiment highly suspicious of firms with large market shares, merger policy was very aggressive in the 1960s and 1970s as mergers were challenged that involved market shares below 5 percent. There is a general consensus among economists that such a merger policy was far too interventionist and that concerns about market power were misplaced. A more tempered merger policy emerged in the 1980s and was guided by the 1982 Merger Guidelines of the U.S. Department of Justice.

Though the focus of this chapter has been on horizontal mergers, which are clearly the primary type of merger of antitrust concern, conglomerate mergers can also be relevant when they entail the elimination of a potential entrant. More broadly, merger policy in the last two
decades has given increased emphasis to entry conditions: how the threat of entry can constrain the augmented market power related to a merger, and how merger can be anticompetitive by influencing entry conditions and eliminating potential entrants. We have not yet discussed vertical mergers, which, historically, have been one of the most contentious issues in antitrust policy. That important topic we turn to in the next chapter.

Questions and Problems

1. Assume the following facts concerning the horizontal merger model developed by Williamson and shown in figure 7.3. Demand is \( q = 100 - P \); average cost premerger, \( AC_0 = $50 \); average cost postmerger, \( AC_1 = $44 \); and premerger price, \( p_0 = $50 \). Assume that the postmerger price, \( p_1 = $70 \), results from the market power created from the merger.
   a. Calculate the value of the deadweight loss, area \( A_1 \).
   b. Calculate the value of the cost savings created by the merger, area \( A_2 \).
   c. Should the merger be allowed? What qualifications should be considered?

2. Assume all of the facts in problem 1 except that now take the premerger price, \( p_0 \), to be $52. How does this affect your answers to problem 1?

3. Assume a homogeneous good market for cellular phones. Two firms, 1 and 2, have a combined demand of \( q = 40 - 0.4P \) and all manufacturers of cellular phones have constant marginal and average costs of $50. Initially, the price is $50.
   a. Firms 1 and 2 have decided to merge. They can lower their cost curve from $50 to $48 because of economies of combined operations. They expect that as the market leader they can lead the industry to a new price of $60. Ignore industry-wide effects—i.e., use the above demand curve—and compute social costs and benefits of the merger. On this basis should the merger be approved?
   b. Now recognize that the two firms above were initially a part of a five-firm industry in which each firm acts as if it has a “share-of-the-market” demand curve of 20 percent of the market demand. The market demand is \( Q = 100 - P \) (Note that the combined demand curve referred to in part a is in fact 40 percent of the market demand.) Would firm 3 favor or oppose the merger assuming that the phone price rises to $60 and it operates on its “share-of-the-market” demand curve, \( q = 20 - 0.2P \)?
   c. If social benefits and costs are now computed on an “industry-wide” basis, should the merger be approved?
   d. Now assume that greater cost savings are expected by firms 1 and 2. Their cost curve will shift down to $45 rather than to $48. It is now a real possibility that the new combined firm will decide to cut price to $50 (or just a bit below) and take over the entire market. Find the new firm’s profits under the price increase strategy (of $60) and under the monopolization strategy. Given that the new firm will follow the most profitable strategy, will firm 3 favor or oppose the merger now?
   e. How might information about rival firms’ attitudes toward a merger (or their stock prices) be useful to antitrust enforcement agencies?
4. A criticism of the model in problem 1 is that price is not related to cost through the standard monopoly theory. That is, if the merger creates monopoly power, then the postmerger price is precisely related to the postmerger cost (and the elasticity of demand). Is this a valid criticism?

5. In a merger between Owens Illinois and Brockway Glass in 1987, the premerger HHI was 1,538. The two merging firms had market shares of 22.7 percent and 15 percent of the glass bottle market. Would this merger be safe under the Merger Guidelines?

6. Explain how the Agency would decide whether to include, say, plastic bottles, in the market referred to in problem 5.

7. According to some economists, horizontal mergers may not always be profitable even though they reduce the number of suppliers. For example, assume a three-firm industry in which the firms behave according to the Cournot model. Let market demand be \( Q = 20 - P \). Each firm has a constant average cost of $4. Now assume that a merger reduces the number of firms to two. Calculate the combined profits of the two firms premerger, and then calculate the profit of the combined firm in the postmerger situation—a Cournot duopoly. Is this a reasonable way of modeling the profitability of horizontal mergers? For further background, see M. K. Perry and R. H. Porter, “Oligopoly and the Incentive for Horizontal Merger,” *American Economic Review* (March 1985).

8. In what ways do conglomerate mergers merit the attention of antitrust authorities?
In previous chapters we examined anticompetitive acts that result from cooperation among competitors. Rivals may cooperate to divide markets or fix prices, or may merge in order to acquire market power. In this chapter and the next, we analyze how firms can harm competition by inflicting injury on their rivals, including those who might become a rival through entry.

The focus in this chapter is on vertical relationships between buyers and sellers. For example, by acquiring or merging with some of its customer firms, a competitor can exclude its rivals from selling to those customer firms. This is known as foreclosure and is one of the most contentious issues in antitrust practice. At the same time, there are often efficiency-enhancing benefits from vertical integration. This is why such cases are evaluated using the rule of reason.

Foreclosure need not only occur through a merger but also through various contractual devices such as tying or bundling. For example, Microsoft bundled Internet Explorer with its Windows 98 operating system. Computer manufacturers who wished to install Windows 98 on their computers (as almost all did) were required to also install Microsoft’s browser. This restriction was a major part of the monopolization case launched in 1998 by the Department of Justice. Other vertical restrictions that will be described and analyzed in chapter 8 include exclusive dealing, resale price maintenance, and territorial restraints. These vertical restraints typically come under Section 1 of the Sherman Act, which prohibits restraints of trade.

Chapter 9 focuses on practices that serve to monopolize a market and thereby violate Section 2 of the Sherman Act. The primary ones that concern us are pricing to drive out competitors and refusing to supply valuable inputs to competitors.

Before moving on, it is worth noting that there has been a common pattern in the evolution of economic thought on many of these antitrust issues. As the turning point in this evolution is the Chicago school of antitrust of the 1950 and 1960s, we will refer to the three phases as pre-Chicago, Chicago, and post-Chicago. The pre-Chicago doctrine was based on loose and nonrigorous reasoning. For example, if a monopolist supplying an input to a competitive industry acquired one of those firms and exclusively supplied it, the pre-Chicago doctrine was that this situation was bad, as the monopolist extended its monopoly by foreclosing the market to other downstream competitors. Though such a conclusion may seem natural, the approach was deficient in that it presumed rather than proved that the purported practice raised profit and reduced welfare.

By comparison, the Chicago school approach was to develop an economic model that could tell us when a particular practice is profitable and, when it is profitable, what its welfare implications are. This line of analysis generally showed that many of these purportedly “bad” practices are not profitable if their sole purpose is to reduce competition. Given that firms do actually use these practices, the Chicago school then offered some efficiency-enhancing reasons for them (such as reducing cost or price discrimination).
With the tool of game theory, the post-Chicago school has developed since the 1980s. It has shown that the conclusions of the Chicago school are true but only in certain cases. Under more general conditions, these practices—such as tying, exclusive dealing, and predatory pricing—can be both profitable to incumbent firms and welfare-reducing.

Vertical Mergers

A vertical merger occurs when firms that previously formed a buyer-seller relationship merge. Transactions that had occurred in the marketplace now take place within the newly defined firm. Examples of such vertical integration are numerous. The petroleum industry consists of many firms that are vertically integrated, from crude oil discovery and production companies to refineries to retail gasoline stations. The Federal Trade Commission (FTC) undertook antitrust actions in the 1960s to block a series of vertical mergers between cement manufacturers and ready-mixed concrete firms. More recently, the 1996 merger of Time Warner and Turner Broadcasting System possessed a vertical dimension in that Time Warner owned cable systems that distributed programming such as that owned by Turner. As described in chapter 12, the electric power industry in some states has experienced vertical disintegration as part of a restructuring of regulation. Utility distribution companies, which previously both produced and distributed electricity, have sold off much of their power-generating capacity to focus on distribution.

In this section we shall examine the possible benefits and costs of such combinations and review the evolution of antitrust law toward vertical mergers. In thinking about the benefits and costs of a vertical merger, it is useful to start with the view articulated by Ronald Coase in a paper he wrote as an undergraduate. He posed the following profound question: What determines the boundaries of a firm?¹ His perspective was that there are transaction costs in performing operations inside the firm as well as in the market, and that organizations would develop so as to minimize those transaction costs. If vertical mergers only involved such costs, then one could reasonably expect that such mergers would be welfare-enhancing in that they would serve to reduce costs. However, as we will see, one must add to the calculus that mergers may influence price by their impact on market power, in which case the welfare calculation is often ex ante ambiguous.²

---

Benefits

Basically, all firms have to decide whether to “make or buy” an input. One antitrust case con-
cerned Ford Motor Company’s decision to acquire the spark plug supplier Autolite so as to
make spark plugs internally as opposed to continuing to buy plugs on the market. Here we
review some of the many welfare-enhancing benefits of vertical integration.

Technological Economies and Transaction Costs

The most common, yet subtle, benefit of vertical integration is that it produces a more effi-
cient organizational form. Such benefits can arise from purely *technological economies*. A
classic example is the integration of ironmaking and steelmaking, where physical proximity
eliminates the need for reheating the iron before it is made into steel. In those instances, a
vertical merger lowers cost and, unless there are some offsetting effects, is sure to enhance
welfare.

A more subtle but no less important benefit comes from lowering *transaction costs*. In
their classic book on organizations, Paul Milgrom and John Roberts begin to describe from
whence such costs arise:

A fundamental observation about the economic world is that people can produce more if they coop-
erate, specializing in their productive activities and then transacting with one another to acquire the actual
goods and services they desire. The problem of organization then arises because when people are
specialized producers who need to trade, their decisions and actions need to be coordinated to achieve
these gains of cooperation, and the people must be motivated to carry out their parts of the cooperative
activity.³

They then identify two broad sources of transaction costs: coordination and motivation. In
the market, *coordination costs* are those incurred in such activities as determining price and
bringing buyers and sellers together. In the Ford Motor Company example, these are costs
that Ford would incur in searching among possible suppliers for the lowest price for spark
plugs, and in negotiating the contracts to spell out exactly what terms of sale would obtain
(product specifications, credit arrangements, delivery dates, and so on). Within a firm, it is
instead the cost of transmitting information, processing the information to determine a plan,
and then communicating that plan. The Internet has clearly reduced market coordination costs,
as reflected in the popularity of eBay in bringing buyers and sellers together.

A second type of transaction cost is referred to as *motivation cost*, or the cost of inducing
people to behave in a manner necessary for trade. Such costs may emanate from asymmetric
information. For example, a firm may not hire another firm to perform some service because
of the inability to measure performance. If monitoring is easier inside the firm, then such

    Hall, 1992), p. 25.
costs may be lower if it is performed internally. A second source of motivation costs is imperfect commitment. By way of example, consider the case of Holland Sweetener, which built a plant to produce Aspartame, a sweetening agent used in, for example, soft drinks.\textsuperscript{4} At the time, Monsanto was the sole producer of Aspartame and accordingly charged a high price. Though Coke and Pepsi might have been willing to pay Holland Sweetener a price high enough to have made the new plant profitable, once it was built and those costs were sunk, Monsanto knew that it could induce Holland Sweetener to sell at a price below its average cost. As it turned out, Coke and Pepsi used the presence of a competitor to force Monsanto into writing new contracts at a much lower price! The problem is that while Coke and Pepsi desired a competitor, they did not commit to compensating that competitor. In many cases, an ex ante contract can solve the problem—Holland Sweetener should have had Coke and Pepsi agree to buy a certain amount at a certain price prior to building the plant—but in some cases such contracts are difficult to write and enforce. If the input is produced within the firm, these motivation costs can be reduced.

Another example of motivation cost is thought to arise in the manufacturer-retailer relationship. A manufacturer may want retailers to engage in promotional and educational activities associated with its product, such as by having ample sales people to explain how the product operates and what its virtues are. Of course, it is costly for a retailer to offer such services. Now, if the retailer gained the benefit from those services, it would have the right incentives in deciding whether or not to provide them. The difficulty is that some retailers may choose to go the low-cost route of not offering these services and, because of its lower cost, charging lower prices. A customer could then go to the high-price/high-service store to learn about the product but buy from the low-price store. If that were to happen, few retailers would offer such services, since they would incur the cost but would not reap the benefits. The problem is that some retailers are \textit{free-riding} on the services of other retailers. The manufacturer then ends up with too little service being provided to potential customers. One solution is to vertically integrate so that the manufacturer owns the retail outlets.

It is important to emphasize that the point is not that vertical integration generally reduces transaction costs. There are always transaction costs, whether the firm buys or builds the input. The issue is instead that, depending on the particular circumstances, the transaction costs could be reduced through having a company build the input, in which case there is an efficiency gain from vertical integration.

\textbf{Double Marginalization}

If an input supplier has market power, then it will charge a price in excess of marginal cost. This causes too little of the input to be used, which entails a welfare loss. If, in addition, the

\textsuperscript{4} This example is from Adam M. Brandenburger and Barry J. Nalebuff, \textit{Co-opetition} (New York: Currency Doubleday, 1996).
downstream firm purchasing the input has market power, then it too will charge a price in excess of its own marginal cost, incurring yet another welfare loss. Thus, the price of the input is marked up twice: by the upstream firm and, in terms of the final product price, by the downstream firm. This is known as \textit{double marginalization}.

A vertical merger can reduce the incentives for double marginalization. This is most easily shown for the case of successive monopolies, so that a monopoly exists at both levels of a vertical chain. By merging, only a single integrated firm is created, and it can be shown to be socially preferable to having two monopolies.

Assume that an upstream motor monopolist sells to a downstream boat monopolist. The boat monopolist adds other inputs to the motor and produces the final product, a boat, at a constant conversion cost of $C$ per unit, or $100$ per unit. Each boat requires exactly one motor and $C$ dollars worth of other inputs. We can therefore use the symbol $Q$ to refer to both the quantity of motors and the quantity of boats. The boat monopolist is assumed to have no monopsony power—that is, it accepts whatever price is set for the motor as fixed. (The opposite assumption—that the boat monopolist \textit{does} have monopsony power—is the case of bilateral monopoly.)

Figure 8.1 shows the final demand curve for boats $DD$ and the derived demand for motors $D'D'$. Finding the derived demand for motors is actually the heart of the analysis. The approach is to use the equilibrium condition for the downstream boat monopolist. That is, it maximizes profit by equating marginal revenue and marginal cost. Or,

\[ MR = P_m + C \quad \text{... marginal revenue equals marginal cost,} \]

so,

\[ P_m = MR - C \quad \text{... the derived demands for motors} \]

Note that the boat monopolist’s marginal cost is its conversion cost $C$ plus the price of a motor, $P_m$. Hence, the derived demand for motors is just marginal revenue minus the conversion cost $C$. The curve $D'D'$ in figure 8.1 is obtained by finding the marginal revenue curve corresponding to $DD$ (the line $DM$) and shifting it downward by the constant $C$ (that is, by $100$).

Of course, the boat monopolist’s input demand $D'D'$ is also the motor monopolist’s product demand curve. The motor monopolist therefore maximizes profit by setting its marginal revenue curve, $D'M'$, equal to the marginal cost of motors $MC$ (constant at $100$). The profit-maximizing quantity is 140 motors, shown in figure 8.1. The price can be found on the $D'D'$ curve as $P_m = 400$. The boat monopolist now equates its marginal revenue for boats $DM$ to its marginal cost (the horizontal line at $500$, labeled $P_m + C$). They intersect at $Q = 140$, yielding a boat price of $650$.

If the two monopolists merged, the integrated firm would maximize profit by setting marginal revenue $DM$ equal to marginal cost $MC'$ ($MC$ of motors + $C$, or $200$). The
profit-maximizing quantity becomes 300 and the final price is $500. Hence, merger leads to a lower price ($500 versus $650) and a larger quantity (300 versus 140) as compared to successive monopoly. It is also true that total profit is larger. The profit gain is measured by the shaded triangular area in figure 8.1 because profit equals the area beneath the marginal revenue curve less the area beneath the marginal cost curve. That is, the area under marginal revenue $DM$ is total revenue and the area under marginal cost $MC'$ is total cost. Expansion of output from 140 to 300 as a result of the vertical integration therefore adds the triangular area to profit. Both the firms and consumers gain from the merger!

Suppose we were to consider the more realistic case of oligopolies, rather than monopolies, in the upstream and downstream industries. Double marginalization is still a problem, and furthermore, it will be reduced by the consummation of a vertical merger. However, other effects can also arise that are anticompetitive; that is, they act to raise the final product price. Though these effects enhance the profitability of the merger, they also reduce consumer surplus, which puts into question the desirability of the merger from society’s perspective. The next section explores such anticompetitive implications of vertical integration.
Two comments are in order. First, there is another efficiency gain related to reducing double marginalization. Suppose that the downstream firm uses multiple inputs and all but one of them is supplied competitively. For reasons given above, the input for which its suppliers have market power will set price above cost. Not only will this result in too little of the final product being supplied, it can also result in an inefficient input mix as the downstream firm substitutes away from the high-priced, imperfectly competitively supplied input. Vertical integration can eliminate that inefficiency as well. We will look at that in greater depth later in the section titled Extension of Monopoly: Variable Proportions.

A second comment is that even if one shows that a vertical merger is profitable and welfare-improving, that does not necessarily imply that the merger will or ought to occur. The reason is that firms may be able to achieve the same efficiency benefits through other means. In the preceding example, let us suppose the upstream firm offered the downstream firm an alternative pricing arrangement: The downstream firm pays a fixed fee, denoted $F$, and a per unit price, denoted $P'$. Thus, if it buys $Q$ motors, then it pays the manufacturer of motors a total of $F + P'Q$. Now consider the motor monopolist setting $P' = 100$, which is the marginal cost of a motor. The boat monopolist would then set the appropriate monopoly price of 500 and earn the monopoly profit of 90,000. Of course, the upstream monopolist earns zero, but we haven’t gotten to the clever part yet. It sets $F = 90,000$. Since the boat monopolist earns zero profit, it is willing to accept this pricing arrangement (given that the alternative is to not produce and earn zero anyway) and the motor monopolist walks away with the entire monopoly profit of 90,000. Double marginalization is eliminated without merger!

What makes this work is that the upstream firm charges a price equal to marginal cost—which results in the quantity that a single monopolist would produce—and then extracts all of the profit from the downstream firm through the fixed fee. A two-part tariff then eliminates double marginalization just as well as does a vertical merger.

In light of this analysis, it will be important to keep in mind whether there are other means by which firms can achieve the desired efficiency benefits (or the desired anticompetitive effects) and whether these means would be more or less attractive than vertical integration.

**Anticompetitive Effects**

Historically, the most common complaint raised in legal proceedings against vertical merger is that it promotes *market foreclosure*. As the Supreme Court stated in its famous *Brown Shoe* decision: “the diminution of the vigor of competition which may stem from a vertical arrangement results primarily from a foreclosure of a share of the market otherwise open to competitors.” As an example, the acquisition of ready-mixed concrete firms by cement

5. This pricing scheme is known as a two-part tariff and is investigated in greater detail in chapter 11 in association with the optimal pricing of a regulated monopolist.

suppliers was said to foreclose the market for cement to nonintegrated cement suppliers. The idea was that if some of the demand or supply was taken out of the market through vertical integration, the market was made less competitive.

Though the welfare-reducing foreclosure effects of vertical mergers were part of antitrust doctrine for many years, it was not predicated on a careful economic analysis. Eventually, the Chicago school of antitrust economics showed that the foreclosure argument was flawed and, in fact, there were no anticompetitive effects. This classic argument will be reviewed in the next section on extending monopoly power. Only decades later in the 1980s did antitrust policy become more lenient with regard to vertical mergers.

While the Chicago school analysis remains correct, time has shown that their conclusions are relevant only under some special circumstances. Game-theoretic analysis in the last two decades has identified situations in which foreclosure can have anticompetitive effects. In this section we will review these arguments toward understanding when we should be concerned with vertical mergers. One general point is that a necessary condition for a vertical merger to have anticompetitive effects is that there is market power in one or both of the upstream and downstream markets.

**Extension of Monopoly**

Let us consider what one might think is the most extreme case of market foreclosure. Suppose a monopolist supplier sells to a perfectly competitive industry. The monopolist extends its monopoly downstream by acquiring one (or more) of the downstream firms and does not provide its input to other downstream firms. The downstream industry is now a monopoly. By the original foreclosure argument, this arrangement is anticompetitive, because a monopoly has been extended from one industry to a second one.

As the Chicago school showed for the case of a fixed-proportions production technology, however, that is not the case, and, in fact, the final product price is unaffected by vertical integration! This argument is reviewed next. We then consider the case of variable-proportions production technology and show that the effects of the merger are not as unambiguous as either the original foreclosure argument or the Chicago school analysis suggests, and, in fact, the merger could result in either a higher or lower final product price.

**Fixed Proportions**

Fixed-proportions production simply means that each unit of output requires a fixed proportion of the various inputs. Some economists think that this assumption is applicable to situations where a manufacturer sells to retailers; that is, retailers combine the manufacturer’s goods with other inputs in fixed proportions. Earlier, we made this assumption in our discussion of double marginalization. There, each boat required exactly one motor plus $C$ dollars’ worth of all other inputs.
We can use that same example here by assuming that boat supply is perfectly competitive and motors are monopolized. Hence we want to examine the consequences of vertical monopolization of boats by the motor supplier.

Figure 8.2 illustrates the case. First, consider the premerger situation. The final demand for boats is $DD$. Subtracting the fixed conversion cost per unit, $C$, we obtain the derived demand for motors, $D'M'$. The motor monopolist equates its marginal revenue, $D'M'$, with its marginal cost of motors, $MC$. It therefore charges a price of $400 and sells 300 motors. The competitive boat industry has a horizontal supply schedule of $400 plus the $100 conversion cost, or $500, and therefore sells 300 motorboats for $500 each. The motor monopolist earns a profit of $400 less $100, multiplied by 300 units, or $90,000.

7. The competitive industry is in equilibrium when $P_b = C + P_m$, that is, when the price of boats $P$ equals marginal cost (which is just the sum of the fixed conversion cost $C$ and the price of motors $P_m$). So, rewriting this condition as $P_b - C = P_m$, we have the derived demand curve, $P_b - C$. 
Now, assume the motor monopolist monopolizes downstream. The marginal revenue corresponding to the final demand for boats is $DM$. The marginal cost $MC'$ of the combined operation is $200$, or the sum of the marginal cost of motors ($100$) and the conversion cost ($100$). Equating $DM$ and $MC'$ gives an output of $300$ boats and a price of $500$. The integrated firm has a profit of $500$ less $200$, multiplied by $300$ units, or $90,000$.

The result is that the monopolist gains nothing by monopolizing downstream. Profit is the same pre- and postmerger. The motor monopolist is able to extract all of the potential profit by choosing its price of motors. From society’s viewpoint there is absolutely no difference. Presumably, in cases where fixed proportions production obtains, vertical monopolization must have some motivation other than increased monopoly profits.

**Variable Proportions**

We now relax the strong assumption of fixed-proportions production. As an example, assume that shoe manufacturing is a competitive industry that is characterized by variable-proportions production. For simplicity, assume that only two inputs are required: shoemaking capital equipment, $K$, and labor, $L$. That is, a given number of shoes, $Q$, can be produced with alternative quantities of $K$ and $L$. Figure 8.3 shows these alternate production possibilities as the isoquant $Q = Q^*$. (An isoquant is the locus of the various combinations of $K$ and $L$ that can be used to produce a particular output level.)

Next, Sam’s Shoe Machinery is assumed to have a monopoly over $K$. As before, we want to examine the incentives for, and efficiency consequences of, a vertical acquisition by Sam’s of the shoe manufacturing industry. Can Sam’s extract all of the profit by an appropriate choice of the price of $K$, as was true in the fixed-proportions case? And will the price of shoes be affected by vertical acquisition?

In anticipation of the results of the analysis, a key difference between the two cases is the following: as $p_K$, the price of $K$, is increased by Sam’s, the shoe manufacturing industry will substitute away from $K$ and use a more labor-intensive input mix. In the fixed-proportions case, this was not possible. No matter what price for motors was charged, more of the other inputs could not substitute for a motor. Every motorboat required one motor.

Let labor be priced at its true opportunity cost (that is, labor is supplied competitively) and take the marginal cost of $K$ to be $MC_k$. The slope of isocost line $NN$ in figure 8.3 is the ratio of $MC_k$ to the labor price. Hence, point $F$ represents the least-cost input mix from society’s viewpoint for producing $Q^*$ shoes. Because $p_K > MC_k$ by assumption, the actual isocost line facing the shoe industry premerger has a steeper slope, such as $PP$. Hence the shoe industry picks input mix $E$, which minimizes its expenditures on inputs. Because the industry’s payments to Sam’s include a monopoly profit, the expenditures on inputs that it minimizes are not equivalent to true resource costs.

The true resource costs at $E$ are higher than at $F$ by the vertical distance $MN$ (measured in units of $L$). In other words, setting a monopoly price on $K$ causes inefficient production in
shoe manufacturing: the costs of production are too high because the input mix is incorrect (from society’s viewpoint).

If Sam’s monopolized forward into shoe manufacturing, the production of shoes would shift from E to F because the integrated monopoly would minimize costs using the true opportunity costs of K and L. The cost saving MN would then constitute a profit incentive for the vertical acquisition.

Thus far it would appear that such a merger should be permitted; the costs of production would be lowered. However, there is a further step in our analysis, namely, what price will the integrated monopolist charge for shoes, given the lower real cost of production? Unfortunately, mathematical analysis shows that the price can either rise or fall, although the most likely case is probably an increase.8 And if price rises, we are back in a trade-off situation where the benefits are cost savings and the costs are deadweight losses due to monopoly pricing.

---

In summary, we have shown that in the case of variable-proportions production, vertical monopolization will be profitable. The welfare effects can be either positive or negative, depending on the particular parameters (elasticity of demand, elasticity of substitution in production, and so on). However, we should not lose sight of the fact that the real problem is the horizontal monopoly that was assumed to exist in shoe machinery. Only if antitrust authorities could do nothing about this horizontal problem does the analysis here become relevant.

Commitment and the Restoration of Market Power

Returning to the case of fixed proportions production, let us describe a situation in which the upstream monopolist is unable to extract full monopoly profit from downstream firms. In that case, a vertical merger may raise the merging firms’ profit but reduce welfare.

Suppose the upstream firm produces jet engines and there are two downstream firms who produce commercial jets. To simplify matters, assume that there are five airlines, and each airline’s demand for jets is at most one unit. Their valuations are shown in table 8.1. For example, airline A values a jet at 140, airline B at 100, and so forth.

The cost of producing a jet is 10 for the engine and 20 for all other inputs. If an integrated monopolist produced one jet, its profit is 110, as it receives a price of 140 from airline A and its total cost is 30 (see table 8.1). If it produced two jets, it can sell them for a price of 100 (selling to A and B), so total revenue is 200, and given that its total cost is 60, it earns profit of 140. As one can see from table 8.1, the production of two jets is the integrated monopolist’s profit-maximizing solution because it yields the maximum profit of 140.

Now consider the situation faced by an upstream manufacturer of jet engines selling to the two downstream manufacturers of jets, which we will denote X and Y. One proposal is that it approaches each jet manufacturer and offers to sell it one engine at a price of 80. First note that if firm X expects firm Y to buy one engine and thereby produce one jet, firm X is indeed willing to pay 80 for an engine. Firm X expects there to be two jets on the market, so they will sell for 100 each. The profit of firm X is then 100 minus the cost of the engine, which is 80, minus the cost of other inputs, which is 20; its profit is zero, so it is willing to pay 80 (but no more). The engine manufacturer would then make a profit of 140 as it sells two engines at a per unit profit of 70 (the price of 80 less cost of 10).

But let us look more closely at exactly how these deals are consummated. Suppose the engine manufacturer first makes a deal with firm X to buy one engine at a price of 80. It then approaches firm Y and the engine manufacturer shows firm Y the contract it signed with firm X to deliver one engine. While the upstream firm could earn profit of 70 from selling one engine to firm Y at a price of 80, consider it offering two engines each at a price of 50. This

Vertical Mergers and Vertical Restraints

Table 8.1
Consumer Valuations and Profit of an Integrated Monopolist

<table>
<thead>
<tr>
<th>Airline</th>
<th>Valuation</th>
<th>Total Revenue</th>
<th>Total Cost</th>
<th>Total Profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>140</td>
<td>140</td>
<td>30</td>
<td>110</td>
</tr>
<tr>
<td>B</td>
<td>100</td>
<td>200</td>
<td>60</td>
<td>140</td>
</tr>
<tr>
<td>C</td>
<td>70</td>
<td>210</td>
<td>90</td>
<td>120</td>
</tr>
<tr>
<td>D</td>
<td>55</td>
<td>220</td>
<td>120</td>
<td>100</td>
</tr>
<tr>
<td>E</td>
<td>40</td>
<td>200</td>
<td>150</td>
<td>50</td>
</tr>
</tbody>
</table>

will yield additional profit of 80, which would give it higher profit of 150, as it earns 70 from firm X and 80 from firm Y. Furthermore, firm Y is willing to pay 50 and buy two engines, as it knows it could sell each of them for a price of 70 (as jets are sold to airlines A, B, and C) and its cost per unit is 70 (50 for the engine and 20 for other inputs).

So, where is the problem? Note that there will be three jets on the market and each will sell for 70. But firm X paid 80 for an engine, which means it is incurring a loss of 30. Firm X is then no longer willing to pay 80 for an engine. In other words, if firm X anticipates that the engine manufacturer will sign such a deal with firm Y, firm X will not sign the original deal to buy one engine at a price of 80. Hence, this cannot be a market equilibrium, as firm X is not maximizing its profit.

A market equilibrium is a pair of deals between the upstream firm and the two downstream firms such that all parties correctly anticipate what happens and each is acting to maximize their profit. Let us show that the equilibrium has the engine manufacturer sell a total of four engines—two to each jet producer—at a price of 35. Suppose it offers such a deal to firm X and firm Y accepts. The engine manufacturer then goes to firm Y. Firm Y will be willing to pay at most 50 for one engine as, in that case, there will be three jets produced (two by firm X and one by firm Y) and each will fetch a price of 70. This yields profit of 40 for the engine manufacturer from firm Y. If firm Y buys two engines and produces two jets, each jet will have a price of 55, so that the most Y is willing to pay for an engine is 35. This yields a profit of 50 for the engine manufacturer from selling to firm Y (as it earns per unit profit of 25 on each of two engines). One can show that it earns profit of only 30 from selling three engines to firm Y. Thus, given that it sells two engines to firm X, the engine manufacturer maximizes the profit it generates from Y by selling two engines at a price of 35 each. Given the symmetry of the situation, if it is optimal for firm Y to buy two engines at a price of 35 given firm X buys two engines at a price of 35, then firm X’s purchase is optimal given it correctly anticipates what Y will do. This describes an equilibrium.

Unfortunately, the upstream monopolist ends up with profit of only 100, as it sells four engines at a price of 35. This is lower than the profit received by an integrated monopolist. What is the source of the problem and what is its solution? The problem is the inability of the upstream monopolist to credibly commit to the deals it makes with the downstream firms.
It would like to commit to selling only one engine to Y so that X is willing to pay 70 for one engine. But X knows that, after locking X into buying one engine, the upstream firm can earn more profit by selling two engines to firm Y. The usual reason for restricting output—that it lowers the price received on other units—is partially stifled because X is locked into paying a price of 70, and this is true whether Y buys one or two units. It is the lack of commitment that prevents the upstream monopolist from extracting full monopoly profit.

A solution is to vertically integrate so that a firm produces both engines and jets. In that case, it will want to produce two engines, each at a cost of 30, and sell them for 100 each. It does not want to sell any engines to the remaining jet producer, for that would only serve to lower total profit. For example, if it sold one engine, then the other jet producer is willing to pay 50, because the price it will receive for a jet is 70 and its other costs are 20. Now, the integrated firm earns a total profit of 130, 50 from selling an engine and 80 from producing and selling two jets at a per unit profit of 40. This is lower than not selling to the other jet manufacturer.

In this situation, vertical integration doesn’t really extend monopoly power but rather restores it. Monopoly power is lost because of the lack of commitment on the part of the upstream monopolist. A vertical merger achieves that commitment. Regardless, the merger is anticompetitive because it results in a higher final product price. Another device that restores commitment is an exclusive dealing contract whereby the upstream firm agrees to sell to only one downstream firm, say firm X. Then firm X would be willing to buy two engines at a price of 80, as it knows that the engine manufacturer is prohibited from selling any engines to firm Y. We will explore exclusive dealing later.

**Raising Rivals’ Costs**

The preceding analysis identified some anticompetitive effects of vertical mergers. Though shown in the context of an upstream monopoly, they are relevant whenever markets are imperfectly competitive. In this section, we consider an anticompetitive effect that does not arise in the case of a monopoly. When both upstream and downstream markets are oligopolistic, vertical integration can be profitable and raise the final product price by causing downstream competitors to have higher cost. This is an example of an anticompetitive effect known as *raising rivals’ costs*.²

In the context of vertical mergers (and also vertical restraints), two types of raising rivals’ costs have been identified. *Input foreclosure* is when the upstream division of an integrated firm excludes downstream firms from purchasing its input, which results in those firms having higher costs because of having to use inferior inputs or facing higher input prices. *Customer*
**foreclosure** is when upstream suppliers are denied access to selling to the downstream division of an integrated firm. By preventing them having an adequate customer base, input suppliers may experience higher cost or fail to achieve enough variable profit to cover its fixed costs. The latter can result in exit or, in the case of a prospective firm, deter entry. We will explore input foreclosure here, and then examine customer foreclosure when we investigate exclusive dealing.

Consider an industry with two upstream firms, denoted U1 and U2, and two downstream firms, denoted D1 and D2. As depicted in figure 8.4(a), both upstream firms can supply both downstream firms. The upstream firms offer a homogeneous commodity and compete by setting prices. They have a common cost of production of 10 and let the price of U1 be denoted $w_i$, where $i = 1$ or 2. The downstream firms offer differentiated products and require one unit of the upstream commodity to produce one unit of the downstream good. A downstream firm’s unit cost is the sum of its input price, which is $w_i$ if it buys from firm U1, plus the cost of transforming that input into the final product, which is specified to be 15. Let us assume linear demand curves for the two downstream firms (which are the same as specified in chapter 5):

![Diagram](image.png)

**Figure 8.4**
Pre-Vertical Integration (a), Post-Vertical Integration (b)
The downstream firms' profit functions are then:

\[
D_1(p_1, p_2) = 100 - p_1 + 0.5p_2, \tag{8.1}
\]
\[
D_2(p_1, p_2) = 100 - p_2 + 0.5p_1. \tag{8.2}
\]

where \(w_i\) is the price paid for the input by \(D_i\). For example, if \(D_1\) buys from \(U_2\), then \(w_1 = w_2\).

If downstream firms compete in terms of price, it can be shown that the Nash equilibrium prices are:

\[
\hat{p}_1 = 76.67 + .534w_1^1 + 0.133w_2^1 \tag{8.5}
\]
\[
\hat{p}_2 = 76.67 + .534w_2^2 + 0.133w_1^1. \tag{8.6}
\]

Recall the concept of Nash equilibrium from chapter 5. \(\hat{p}_1\) and \(\hat{p}_2\) are Nash equilibrium prices if, given firm \(D_2\) prices at \(\hat{p}_2\), a price of \(\hat{p}_1\) maximizes firm \(D_1\)'s profit and, given firm \(D_1\) prices at \(\hat{p}_1\), a price of \(\hat{p}_2\) maximizes firm \(D_2\)'s profit. As shown in (8.5), \(D_1\)'s equilibrium price is increasing in the input price it pays because its profit-maximizing price is higher when its marginal cost is higher. But the reason that it is increasing in the input price paid by firm \(D_2\) is that \(D_2\)'s price is increasing in its marginal cost and \(D_1\)'s optimal price is higher when its rival prices higher. This is due to products being substitutes, so that a firm's demand is stronger when the competitor sets a higher price and, under normal assumptions, stronger demand implies a higher profit-maximizing price.

These prices can be derived in the following manner. Starting with firm 1, find the price that maximizes its profit as expressed in (8.3). This is found by equating marginal profit to zero,

\[
\frac{\partial \pi_1(p_1, p_2)}{\partial p_1} = 115 - 2p_1 + 0.5p_2 + w_1 = 0,
\]

and then solving it for the firm's price, \(p_1 = 57.5 + 0.25p_2 + 0.5w_1\). This is firm \(D_1\)'s best reply function, as it prescribes the profit-maximizing price given its rival's price (and the input price). In the same manner, one can do this for the other firm, \(p_2 = 57.5 + 0.25p_1 + 0.5w_2\). We then need to find prices that satisfy these two equations simultaneously:

\[
p_1 = 57.5 + 0.25p_2 + 0.5w_1
\]
\[
p_2 = 57.5 + 0.25p_1 + 0.5w_2.
\]

Substituting the second equation into the first,

\[
p_1 = 57.5 + 0.25(57.5 + 0.25p_1 + 0.5w_2) + 0.5w_1,
\]

and solving for \(p_1\) gives us (8.5). The same procedure will produce (8.6).
Let us now examine upstream competition in the absence of a vertical merger. Since the upstream firms offer homogeneous products, they will compete vigorously for the demand of the downstream firms. In fact, competition drives price all the way down to their marginal cost of 10.\(^{12}\) As upstream goods are identical, a downstream firm will buy all of the input from the firm with the lowest price. If both upstream firms were to price above cost, say at 15, and equally share demand, one of them could undercut the price a little bit, say to 14.99, and experience a doubling of its demand. Such a move is clearly profitable. Equilibrium occurs when price is 10, as then no firm wants to lower price, as doing so results in losses, and there is no reason to raise price, as profit remains zero since demand falls to zero. The equilibrium has upstream prices of 10, \(w_1 = 10 = w_2\), and downstream prices of 83.34 using (8.5)–(8.6).

Now consider the prospect of firms U1 and D1 merging. As shown in figure 8.4(b), firms U1 and D1 form a firm and the lone supplier of D2 is U2. The integrated firm will produce the input internally at cost 10, so its cost is the same as without the merger. At first glance, the merger has not benefited these firms, but let us examine the situation faced by the unintegrated firms—upstream firm U2 and downstream firm D2. Here we will make a key assumption that will be discussed later: Assume the integrated firm does not sell its input to D2. The question is, What price will firm U2 set for its input? Where premerger it was in vigorous competition with the other input supplier, it is now in a monopoly position. It will then charge a price above 10, knowing that, while firm D2 will demand less than if the input price is 10, it will still demand a positive amount. Indeed, one can show that the profit-maximizing price is \(w_2 = 72.45\).\(^{13}\) This means that D2 ends up with a (much!) higher marginal cost.

Now that D2’s cost is higher, we know by (8.5)–(8.6) that both firms’ prices are higher.\(^{14}\) The unintegrated downstream firm prices higher because it faces a higher price for the input. The integrated firm prices higher downstream because its rival, firm D2, prices higher. Furthermore, vertical integration has raised the profit of the two merger partners because it has induced a rival to price higher which is always good. This is the raising rivals’ costs effect.

\(^{12}\) This is the Bertrand price game, which is analyzed in the appendix to chapter 11.

\(^{13}\) To derive this price, one must first derive the demand for the input supplied by U2. Substituting the downstream equilibrium prices into (8.2), one derives U2’s demand curve:

\[
100 - (76.67 + 0.534w^2 + 0.133 \times 10) + 0.5(76.67 + 0.534 \times 10 + 0.133w^2) = 63 - 0.467w^2.
\]

U2 then chooses its input price to maximize

\[
(63 - 0.467w_2)(w_2 - 10).
\]

Taking the first derivative, setting it equal to zero, and solving for the input price yields a price of 72.45.

\(^{14}\) The downstream prices are now 91.63 and 116.69 for the integrated firm and the unintegrated firm, respectively.
and is an anticompetitive implication of the vertical merger. Finally, since firms’ prices are higher and the total cost of production remains the same, social welfare is lower due to the merger.

Before one jumps to any general conclusions, there are several important caveats. First, our simple model did not have double marginalization as, prior to the merger, the input is priced at marginal cost. This property is special to the assumption that upstream firms have homogeneous products and thereby price at cost. If products are not perfectly identical (or if firms compete in quantities rather than prices), upstream firms will price above cost. In that case, double marginalization is present in the premerger scenario, and furthermore, vertical integration will reduce it. That implication is a welfare-enhancing feature to the merger. More generally, one can expect two counteracting forces with a vertical merger: raising rivals’ cost effect reduces welfare and reduced double marginalization raises welfare. How they net out depends on the particular situation.

Second, contrary to the assumption made, the integrated firm would actually find it optimal to supply the input to the unintegrated downstream firm. For example, if firm U2 prices at 72.45, then the integrated firm would prefer to undercut at a price of 72.44. Doing so will only marginally lower the price of firm D2, and the integrated firm will make positive profit on the input sales. However, there are other models of competition for which a raising rivals’ cost effect is present and the integrated firm does not want to sell to the unintegrated downstream firms. And in those models the input price and the final product prices can both be higher. So, while our model is subject to this criticism, other (more complicated) models are not.15

The third caveat is that the unintegrated firms might respond by integrating themselves. Indeed, such is optimal in our model. Once again, in a more complex model, one can find circumstances in which it is not optimal to respond with vertical integration, so that our conclusions persist.16 However, the assumptions required for that to be true are rather strong.

Finally, it is worth mentioning a recent analysis showing that a dominant downstream firm that competes with a collection of competitive firms can profitably backward integrate into a competitive upstream industry with potentially negative welfare effects.17 A motivating


example is the steel industry at the start of the twentieth century. U.S. Steel was a dominant firm with a market share exceeding 60 percent of the steel market. It acquired an essential input, iron ore, to the point that it owned 40 percent of reserves. Though there are efficiency benefits from the integration—as the dominant firm is assumed to have lower cost, so that shifting some of the downstream supply to it reduces total industry cost—it also serves to raise the input price faced by its downstream competitors because of the raising rivals’ cost effect. Though the final product price is always higher with integration, and thus consumers are worse off, the welfare effect can be positive or negative.

In summary, we have suggested that harmful effects from vertical integration are unlikely to occur unless there is preexisting market power at one level or both. While this seems to indicate that the real problem is horizontal market power that should be attacked directly, perhaps this approach is not always possible. Hence we shall pursue the analysis from a somewhat different angle. Suppose we recognize that market power at one level can be extended to another level. Does the creation of a “second” monopoly have any harmful consequences for economic efficiency? If it does not, then the merger may have been effected for other reasons—for example, socially beneficial transaction cost savings. In this case, economic efficiency might be better served by favoring such mergers.

**Antitrust Law and Policy**

The history of antitrust policy with respect to vertical mergers is succinctly stated by former FTC Commissioner Robert Pitofsky:

Prior to the late 1970s, the government challenged a number of vertical mergers, and would usually succeed even with relatively small levels of foreclosure. In cases like *Brown Shoe Co. v. United States* and *Ford Motor Co. v. United States*, the courts looked critically even at small amounts of foreclosure and were generally unsympathetic to the efficiencies that could arise from vertical consolidation. . . . the courts’ analysis of vertical mergers began to change in the late 1970s. Some lower courts were less than sympathetic to vertical merger challenges, even when the market shares were significant. In 1982, the Justice Department revised its Merger Guidelines which significantly liberalized the treatment of vertical mergers. Not surprisingly, there were almost no challenges of vertical mergers during the 1980s. More recently, the antitrust agencies have begun to look critically at vertical mergers and joint ventures and have challenged several vertical mergers. This is in part due to more careful economic analysis of the effects of vertical mergers.¹⁹

Let us now flesh out these various stages to vertical merger policy.

---


Historical Development

The *Brown Shoe* case in 1962, discussed earlier under horizontal merger cases, also had important vertical dimensions. The Court held that the relevant market was the entire United States and noted with concern a trend toward increasing vertical integration:

Since the diminution of the vigor of competition, which may stem from a vertical arrangement, results primarily from a foreclosure of a share of the market otherwise open to competitors, an important consideration in determining whether the effect of a vertical arrangement [is illegal] is the size of the share of the market foreclosed.

In *Brown Shoe*, the size of the market foreclosed was on the order of 1 percent! That is, Brown (primarily a shoe manufacturer) could be expected to force Kinney (primarily a shoe retailer) to take only a small volume of Brown shoes, to the exclusion of other shoe manufacturers. However, as in the horizontal part of the case, the Court gave great weight to the trend and to halting the process in its incipiency.

In the Ford Motor Company decision in 1972, the Supreme Court held that Ford’s acquisition from Electric Autolite Company of the name Autolite and associated spark-plug manufacturing assets was illegal. One basis for the decision was that the merger resulted in “the foreclosure of Ford as a purchaser of about ten percent of total industry output.”

By contrast, antitrust policy became extremely lenient in the 1980s. Under the Reagan and Bush administrations, the Department of Justice (DOJ) prevented only one vertical merger, the proposed merging of the cable programs Showtime and The Movie Channel. This change in policy was also reflected in the DOJ merger guidelines. While the 1968 guidelines stated that vertical mergers between a supplier with a 10 percent share and a buyer with a 6 percent share might be challenged, the 1982 guidelines had no similar statements about foreclosure percentages.

The guidelines do caution that vertical mergers will be challenged where they have anticompetitive horizontal effects. In particular, if a vertical merger creates barriers to entry, facilitates collusion, or enhances the ability to evade rate regulation, it may be challenged. In particular, it is noted in Section 4 of the 1984 guidelines that vertical mergers may facilitate collusion in an already highly concentrated manufacturing industry. Two scenarios are described:

A high level of vertical integration by upstream firms into the associated retail market may facilitate collusion in the upstream market by making it easier to monitor price. Retail prices are generally more visible than prices in upstream markets, and vertical mergers may increase the level of vertical integration to the point at which the monitoring effect becomes significant.

The elimination by vertical merger of a particularly disruptive buyer in a downstream market may facilitate collusion in the upstream market. If upstream firms view sales to a particular buyer as sufficiently important, they may deviate from the terms of a collusive agreement in an effort to secure that business, thereby disrupting the operation of the agreement. The merger of such a buyer with an upstream firm may eliminate that rivalry, making it easier for the upstream firms to collude effectively.

A second concern raised by the guidelines is that a vertical merger, by resulting in foreclosure, might make entry more difficult by requiring a prospective firm to enter both the upstream and downstream markets. This "two-level" entry may be riskier, which would serve to discourage entry.

**Time Warner and Turner**

An example of the more assertive antitrust policy in recent years is the FTC's response to a proposed merger involving three giants of the cable television industry: Time Warner, Turner Broadcasting System, and TCI. The FTC ultimately allowed the merger of Time Warner and Turner, though with some restrictions, but only permitted TCI to have a passive interest in the newly formed company.

The vertical dimension to this case is that cable programming (the upstream industry) is an input into cable service (the downstream industry). At the time, TCI was the largest cable service provider, with about 27 percent of all U.S. cable television households, with Time Warner coming in second at around 17 percent. Furthermore, as shown in table 8.2, Time Warner owned cable networks such as HBO and Cinemax. Turner was solely focused on programming and offered such channels as CNN and TNT.

In September 1995, Time Warner agreed to purchase Turner and, due to its shareholding in Turner, TCI would come to own 7.5 percent of Time Warner. The proposed merger would result in the new Time Warner controlling more than 40 percent of programming assets, while Time Warner and TCI would serve about 44 percent of cable subscribers.

The FTC was concerned with two forms of foreclosure in the programming market:

...the complaint alleged that post-acquisition, Time Warner and TCI would have the power to: (1) foreclose unaffiliated programming from their cable systems to protect their programming assets; and (2) disadvantage competing cable distribution systems, by denying programming, or providing programming only at discriminatory (i.e., disadvantageous) prices.²²

In approving the merger, the FTC imposed several conditions in response to these anticompetitive concerns. First, Time Warner could not package the most desirable channels, like HBO, with lesser channels, for the reason that doing so would limit cable capacity for other programming companies. Second, certain reporting requirements were imposed so as to make it easier for the FTC to learn about possible exclusionary activities. Third, the FTC eliminated a long-term agreement requiring TCI to carry Turner programming at preferential

prices. The fear was that it would foreclose non-Time Warner programming on TCI cable systems.

A recent analysis shows that there was justification for the FTC to have been concerned about foreclosure, though, after netting out the efficiency benefits and the anticompetitive costs, consumers are likely to be better off from vertical integration.23 The study finds that cable systems that own premium programming provide, on average, fewer programs—one less premium channel and one to two fewer basic channels. By way of example, consider the shopping networks, QVC (which is owned by TCI and another cable operator, Comcast) and Home Shopping Network. Only 6 percent of TCI and Comcast systems carry HSN, in comparison to 28 percent of all cable systems. And while both networks are carried by 9 percent of all cable systems, this is true of only 5 percent when it comes to TCI and Comcast. Estimates further reveal that, compared to other cable systems, TCI and Comcast are 25 percent less likely to carry HSN and 4 percent less likely to carry both QVC and HSN. Similar results are found for premium movie services.

Though there is, then, evidence of foreclosure, we also know that there are efficiency benefits from vertical integration, such as reduced double marginalization. The evidence is consistent with those efficiency benefits dominating, as integrated cable companies tend to have a higher penetration rate; that is, a higher fraction of households covered by the cable system choose to buy their cable service. This appears to be due to integrated cable systems offering more channels with less duplication and more premium services.

Vertical Restraints

We turn now to a set of business practices that often accomplish some of the same objectives as vertical integration, but through contractual means rather than a merging of firms. The various practices that we shall examine are exclusive dealing, tying, resale price maintenance, and territorial restraints. We begin with some concrete examples of each practice.

Exclusive dealing can be illustrated by an agreement between Exxon and an independent service station that the service station would buy all its gasoline and motor oil supplies from Exxon. Exclusive dealing can be viewed as a way of accomplishing vertical integration by contract. In its relations with game developers, Nintendo used exclusionary contracts in the 1980s, when it had the dominant video game platform. In producing a game for Nintendo, a game developer had to agree not to offer a version of that game for any other video game system for a period of two years.

Tying refers to the practice of a supplier agreeing to sell its customer one product (the tying good) only if the customer agrees to purchase all of its requirements for another product (the tied good) from the supplier. A well-known example was IBM’s practice in 1936 of leasing its tabulating machines only on the condition that the customer purchase all of its needs for tabulating cards from IBM.

The remaining two practices are most common to manufacturer-retailer relations. Resale price maintenance (or RPM) means that the supplier constrains the dealer in terms of the price at which it can resell its product. (It is sometimes referred to as vertical price fixing.) Usually, RPM is either a minimum resale price or a maximum resale price. An example of minimum RPM would be if Electronic Arts required retailers to sell its video games for no less than $50. An example of a maximum resale price would be if the New York Times required its home delivery distributors to sell the newspaper for no more than $10 per week. Note that a vertically integrated firm would set its own resale price to final customers, so RPM is a partial substitute for vertical integration.

A territorial restraint is an agreement between a supplier and a dealer that the former will not allow any other dealer to locate within a certain area—thereby preserving an exclusive marketing territory to that dealer. Such agreements are widespread in the automobile industry. An example would be if Ford Motor Company agreed to allow only one Ford dealership
in a city. Again, of course, if Ford Motor were completely integrated into retailing its cars, it would obviously choose where it would locate its retail outlets.

These practices are generally judged under either Section 1 of the Sherman Act or Section 3 of the Clayton Act. Section 3 of the Clayton Act specifically mentions both exclusive dealing and tying, and holds them to be illegal “where the effect . . . may be to substantially lessen competition or tend to create a monopoly.” While exclusive dealing comes under the rule of reason, tying is judged under a “modified” per se rule. The rule of reason is applied to territorial restraints, and maximum RPM though minimum RPM is per se illegal.

**Exclusive Dealing**

Exclusive dealing is a contract between a supplier and a dealer stating that the dealer will buy all of its supplies from that supplier. In effect, exclusive dealing is an alternative way of accomplishing vertical integration—it is “contractual” integration rather than the more permanent ownership integration that we discussed earlier. And just as vertical mergers worry the courts because of possible foreclosure of rivals, exclusive dealing is believed to have similar anticompetitive effects.²⁴

**Economic Analysis**

Just as vertical integration is often the efficient organizational form because it reduces transactions costs, the same can be said in favor of exclusive dealing. Benefits may include lower selling expenses by the supplier and lower search costs by the dealer. Also, the supplier may find it worthwhile to invest in developing the skills of the dealers if it knows that the dealers will be devoting all their efforts to selling the supplier’s products. Another factor is that the supplier may find it worthwhile to promote the products nationally if it knows that the dealers will not substitute a lower-priced nonadvertised brand when consumers flock to their stores.

**Chicago School Theory**

As an exclusive contract between a supplier and a buyer means that all other sellers are foreclosed from selling to that buyer, there is the anticompetitive concern of foreclosure; buyers will be harmed by the elimination of other sellers. The Chicago school claims, however, that a buyer would not sign such a contract that commits it to buying from one seller unless doing so made it better off than not signing the contract. And, furthermore, such an acceptable contract to the buyer would actually prove unprofitable to the seller.

To see how this argument works, consider a situation in which there is currently one monopoly supplier with marginal cost \( c' \). Given the demand curve \( D(P) \) in figure 8.5, it will set a price and quantity to equate marginal revenue, \( MR \), and marginal cost. This results in a price of \( P^m \). Now suppose a more efficient entrant comes along with marginal cost \( c'' \) which is less than \( c' \). If there is no exclusive contract signed then the two firms compete and, assuming their products are homogeneous and they compete by setting prices, the equilibrium will be that the new firm prices just below \( c' \) and meets all demand of \( Q^o \). The incumbent seller does not want to match or undercut this price since it will be pricing below cost. Furthermore, if it prices at \( c' \) (which it is content to do), the new firm will not want to set a price higher than \( c' \) as then the original firm will get all of the demand.  

Now consider the incumbent seller offering a contract to the buyer which excludes her from buying from any other firm. If she signs the contract, the buyer knows that the seller

---

25. This is the Bertrand price game when firms have different costs. The case of identical costs is covered in the appendix to chapter 11.
will take advantage of its monopoly position and charge a price of \( P^m \). In that case, the buyer gets surplus equal to triangle \( A \). If she doesn’t sign the contract then she’ll face a price just below \( c' \) from the new firm and receive surplus equal to the sum of \( A, B, \) and \( C \). A buyer must then be paid an amount equal to \( B + C \) in order to get her to sign the contract. Now here comes the rub. The amount that the original seller must pay the buyer to accept exclusivity exceeds monopoly profit, which is measured by rectangle \( B \). Thus, the seller is worse off by an amount equal to the deadweight welfare loss \( C \) from pursuing exclusive dealing. If exclusive dealing is used, the Chicago school argues, it must then be for efficiency reasons which serve to expand the surplus in some manner.

As is typical, the Chicago school analysis is correct but not universally applicable. More recent analysis has found rationales for exclusive dealing predicated not on efficiency but on harming rivals and consumers. A common feature to these models is that the presence of an externality creates the opportunity for some agents, those who are party to the contract, to extract surplus from those excluded from the contract. In the first model, a properly designed exclusionary contract between the incumbent seller and buyers can extract surplus from a more efficient entrant. This can be anticompetitive because it reduces the likelihood of a more efficient firm entering. In the second model, entry is once again deterred but now it is because the incumbent seller signs a contract with only some of the buyers and this transfers surplus from the remaining buyers. Once again, efficient entry is prevented.

**Exclusionary Contracts That Extract Surplus from a More Efficient Entrant**

The first step in this analysis is to expand the set of contracts between the incumbent seller and the buyer.\(^{26}\) An exclusionary contract is now described by a pair \((p, x)\), where \( p \) is the price that a buyer pays for each unit it purchases from the incumbent seller and \( x \) is a penalty it must pay to that seller for each unit the buyer buys from other sellers. The innovation in the contract is that penalty. A pure exclusive dealing contract can be thought of as one in which \( x \) is infinity so that, once signed, a buyer would never buy from any other firm. As we will see, lowering the penalty can induce the buyer to accept the contract and this can reduce welfare.

Suppose the buyer and the incumbent seller agree to a contract \((c', x)\) so it is supplied at cost. With this contract in place, the buyer effectively faces a price of \( p_E + x \) from the entrant when the entrant charges a price of \( p_E \) since the buyer must pay \( x \) to the incumbent seller for each unit she buys from the entrant. Thus, the buyer will only buy from the entrant if \( p_E + x < c' \). Given it enters, the equilibrium would then have the entrant pricing just below \( c' + x \) (as long as \( x \) is sufficiently small so that \( c'' < c' - x \)) and selling \( Q^o \) units.

---

What have the buyer and seller gained from such a contract? Without the contract, the prospective firm enters and sells just below $c'$, in which case the incumbent seller earns zero profit and consumers get surplus of $A + B + C$ (as they pay a price of $c' - x$ but also a penalty of $x$). With the contract, the consumer gets the same surplus but the incumbent seller receives profit of $xQ^o$ from the penalty clause. In sum, they are better off and, in order to break the buyer’s indifference about signing the contract, the incumbent seller could provide an initial payment to the buyer for signing the contract and thereby share the anticipated gain of $xQ^o$.

What the incumbent seller and buyer have done is to extract some of the surplus created by the entry of a more efficient firm. Without the contract, the entrant earns profit of $(c' - c^o)Q^o$, but now it earns only $(c' - c^o - x)Q^o$. In fact, the buyer and incumbent seller could extract all of the entrant’s surplus by setting $x = c' - c^o$. As shown in figure 8.5, the incumbent seller and buyer would then add rectangle $E$ to their original surplus of $A + B + C$.

We have shown how an exclusive dealing contract that allows the buyer an escape clause with a penalty can both enhance the incumbent seller’s profit and be attractive to the buyer, something the Chicago school thought wasn’t possible. But what we have not shown is that it is anticompetitive; entry still occurs, and price falls. Thus far, social welfare, the sum of profit and consumer surplus, has remained the same; it’s just that we have shifted it around among the various parties.

Let us now enrich the model by first assuming there is a fixed cost to entry, $k > 0$. The net profit to an entrant is then $(c' - x - c^o)Q^o - k$. In that case, if $x = c' - c^o$ then the entrant would not enter since its variable profit is zero so it could not cover its cost of entry. But the incumbent seller and the buyer do not want to prevent entry entirely as then the additional surplus is never created and thus there is nothing to extract.

The next modeling change is to suppose that $k$ is random, unknown to the incumbent seller and buyer, but known to the potential entrant when it makes its entry decision. It will enter when $k$ is sufficiently low. For example, when $x = 0$, the entrant will enter when $k ≤ (c' - c^o)Q^o$, as it will charge a price (just below) $c'$. In that case, the probability of entry is the probability that $k$ is less than $(c' - c^o)Q^o$. It can be optimal for the incumbent seller and buyer to set $x > 0$ in order to extract some surplus, in spite of it having the undesirable implication of lowering the probability of entry. This results in a welfare loss, since it is now less likely that a more efficient firm will enter. In those instances in which there would have been entry without exclusive dealing but there is not now, price is higher. The incumbent seller and buyer are willing to forgo that surplus in exchange for getting a higher fraction of surplus when there is entry.
Exclusionary Contracts That Extract Surplus from Some Customers

Let us now turn to a second way in which an exclusionary contract is profitable to the incumbent seller, acceptable to buyers, and welfare-reducing. Suppose the situation depicted in figure 8.5 is for a single buyer and there are three identical buyers, whom we will call Manny, Moe, and Jack. Each buyer has demand curve $D(P)$, so that total demand is $3D(P)$. In the absence of exclusionary contracts, the net profit from entry is $(c' - c")3Q^o - k$. Let us make the following assumption:

$$(c' - c")2Q^o - k > 0 > (c' - c")Q^o - k.$$ 

Alternatively stated (and referring back to figure 8.5), $2E$ exceeds $k$ but $k$ exceeds $E$. Thus, if a new firm is able to sell to two or three buyers, variable profit is sufficient to cover the cost of entry. However, if it only gets one buyer, then entry is unprofitable.

Consider the incumbent seller entering into an exclusive dealing arrangement with Manny and Moe whereby they are prohibited from buying from any other firm. In that case, entry is unprofitable, as the entrant has only one buyer to whom to sell. Now let us turn to the situation faced by Manny. By not signing the contract, the most he can receive in surplus is $A + B + C$, which he gets when entry occurs. The least surplus he can get from signing the contract is $A$. Thus, Manny will surely sign the contract if the incumbent seller provides compensation of $B + C$ plus a little bit more. The same is true for Moe. Therefore, the cost to the incumbent seller from locking in Manny and Moe and thereby deterring entry is no more than $2(B + C)$. The value of deterring entry is $3B$ as, regardless of whether or not a buyer has signed an exclusive dealing contract, the incumbent seller charges the monopoly price. The incumbent seller can then offer an exclusive dealing contract that Manny and Moe will accept as long as $3B$ exceeds $2B + 2C$. It is not difficult to find examples where that does indeed occur. We conclude that the incumbent seller can successfully deter entry by a more efficient firm by signing some (but not all) of the buyers to an exclusive dealing contract.

The surplus extraction in this case is not from the entrant as the entrant does not enter and thus there is no surplus to extract. The surplus comes instead from Jack; that is, the buyer left out in the cold without a contract. Jack ends up with surplus of only $A$, while Manny and Moe each receive $A + B + C$. And price remains at $P^m$, with an inefficiently high cost of $c'$. Exclusive dealing lowers surplus by an amount of $3(C + E) - k$; there is the deadweight welfare loss of $3C$ and the increase in total cost of $3E$, but then the entry cost is avoided. This is strictly positive since, by assumption, $3E$ exceeds $k$.

If the incumbent seller is clever, it may be able to play the buyers off against each other so that it gets them to agree to exclusive dealing for almost nothing! Suppose it offers an exclusive dealing contract to Manny and Moe with a small payment. And it tells each of them that if he refuses the deal, it’ll be offered to Jack. If Manny and Moe anticipate the other one signing the deal and anticipate Jack signing the deal if one of them refuses, each will believe that two buyers will end up locked in to buying from the incumbent seller. Thus, regardless of whether or not one of them signs the deal, they expect entry to be deterred, which means surplus of A. If they sign it, they get surplus of A plus the small payment that the incumbent seller is offering. In that case, Manny and Moe should sign the contract. Exclusive dealing occurs, entry is deterred, and the incumbent seller ends up with surplus of almost \(3B\).

In sum, the Chicago school argues that for an incumbent seller to induce all buyers to sign an exclusive dealing agreement, it requires payment exceeding monopoly profit. From this they concluded that exclusive dealing does not work as a strategy to deter efficient entry. What the previous analysis shows is that it doesn’t require signing all buyers to an exclusionary contract to deter entry. And if that is true then the amount of required payment can be less than monopoly profit in which case exclusive dealing can be profitable for the incumbent seller. Because an entrant has to earn enough variable profit to cover its fixed cost (that is, there are scale economies), the trick is for the incumbent firm to sign enough buyers—having enough customer foreclosure—so that a new firm lacks that critical mass to make entry profitable. A closely related mechanism is relevant to understanding how exclusive dealing could have been profitable to Microsoft as an anticompetitive device. In that case, it is network externalities, rather than traditional scale economies, that are relevant. This we will discuss in chapter 9.

Antitrust Law and Policy

Historical Development

The courts have treated exclusive dealing harshly. In a 1922 case, *Standard Fashion Company v. Magrane-Houston Company*, the Court found an exclusive dealing arrangement between a manufacturer of dress patterns and a dry-goods store to be illegal.\(^{29}\) The reason was that it was believed that rival pattern manufacturers were foreclosed from the market. The Supreme Court approved an evaluation of the problem given by the Circuit Court of Appeals:

---

28. For an analysis of how buyer coalitions can ensure that exclusionary contracts are welfare-enhancing, see Robert Innes and Richard J. Sexton, “Strategic Buyers and Exclusionary Contracts,” *American Economic Review* 84 (June 1994): 566–84. When buyers are themselves firms who compete in selling the good, the argument for exclusionary contracts being anticompetitive is also weakened. This is shown in Chiara Fumagalli and Massimo Motta, “Exclusive Dealing and Entry When Buyers Compete” (European University Institute, June 2003), pdf.

The restriction of each merchant to one pattern manufacturer must in hundreds, perhaps in thousands, of small communities amount to giving such single pattern manufacturer a monopoly of the business in such community.

The Circuit Court went on to observe that this could lead to ever higher concentration in the pattern business nationally “so that the plaintiff . . . will shortly have almost, if not quite, all the pattern business.” This analysis is not sound, however, since it ignores the issue of what a pattern manufacturer must give up to obtain an exclusive dealing arrangement in the first place. That is, the dry-goods stores can benefit by tough bargaining with potential pattern suppliers before signing up with a particular one. In short, this theory of monopolizing through foreclosure is no more persuasive than the foreclosure arguments that we have discussed earlier in this chapter.

In a case decided in 1961 the Supreme Court refused to strike down an exclusive dealing arrangement between Tampa Electric Company and Nashville Coal Company. The reason was that the arrangement involved only about 0.77 percent of total coal production and this was insufficient to qualify as a substantial lessening of competition in the relevant market. However, in a 1949 case involving an exclusive dealing arrangement between Standard Oil Company of California and about 6,000 retail service stations, the Court held that 6.7 percent of the market was sufficient for illegality. Hence, whether exclusive dealing is likely to be illegal seems to depend on the market shares involved.

**Visa–MasterCard**

With transactions surpassing $1 trillion annually, the general-purpose charge and credit card market is one of the most significant financial markets in the United States. The four most significant cards are well-known to all readers—Visa, MasterCard, American Express, and Discover. A recent decision in this market affirmed the importance of exclusive dealing as an antitrust issue.

Visa and MasterCard are networks owned by the thousands of member banking institutions. Member banks, such as Citibank, may either be an “issuer”—so they issue Visa and/or MasterCard—or an “acquirer,” or both. When an individual seeks to make payment using, say, Visa, the merchant conveys the transaction to an “acquiring” bank (who is contracted with the merchant), who then relays it to the “issuing” bank (who issued the card to the individual). If the transaction is approved, the issuer retains a fee equal to 1.4 percent of the transaction amount and the acquirer retains 0.6 percent. Thus, a merchant with a $100 transaction actually receives $98.

By comparison, American Express and Discover work quite differently by performing both issuing and acquiring roles. However, they are similarly compensated by merchants, with American Express’s transaction fee typically being 2.73 percent and Discover’s about 1.5 percent.

There were two issues to this case. The first was the “dual governance” feature to Visa and MasterCard in that many banks were member-owners of both networks. The courts did not find that to be a violation of the Sherman Act. The second issue is the one that concerns us and is the exclusionary nature of the agreement made between the Visa–MasterCard networks and the member banks. Visa’s by-law 2.10(3) and MasterCard’s Competitive Programs Policy prohibited member banks from issuing the cards of certain other competitors, including American Express and Discover.

The claim of the plaintiffs was that this was a violation of Section 1 of the Sherman Act. In applying the rule of reason, the U.S. Court of Appeals nicely summarized the procedure: \(^{33}\)

As an initial matter, the government must demonstrate that the defendant conspirators have “market power” in a particular market for goods or services. Next the government must demonstrate that within the relevant market, the defendants’ actions have had substantial adverse effects on competition, such as increases in price, or decreases in output or quality. Once that initial burden is met, the burden of production shifts to the defendants, who must provide a procompetitive justification for the challenged restraint. If the defendants do so, the government must prove either that the challenged restraint is not reasonably necessary to achieve the defendants’ procompetitive justifications, or that those objectives may be achieved in a manner less restrictive of free competition.

To establish market power, it was first noted that this is a highly concentrated industry. In 2001, Visa and MasterCard had 73 percent of all transactions (46 and 27 percent, respectively), with American Express (20 percent) and Discover (6 percent) having most of what remained. Of course, high concentration is typically necessary but not sufficient for there to be market power. Evidence was put forth that in spite of its attempts to get banks to issue its card, American Express had not convinced a single bank to do so because the exclusive dealing agreement meant that the bank would lose its membership in the Visa and MasterCard consortia. By comparison, American Express had been successful in getting banks to issue its card outside of the United States, in count where this exclusionary provision was absent. (In 1996, the European Commission forced Visa and MasterCard to eliminate the exclusionary provision in Europe.) It was felt that the exclusion of these other cards lessened competition.

Though the defendants sought to provide an efficiency rationale for their exclusionary contracts, it proved to be unconvincing. In 1999, U.S. District Judge Barbara Jones found this to be an unreasonable restraint of trade and in violation of the Sherman Act. As part of the judgment, Visa and MasterCard were required to repeal the exclusionary features of their

---

contracts with banks. This decision was affirmed by the U.S. Court of Appeals for the Second Circuit in 2003.

Tying

Tying is the practice of a seller conditioning the purchase of one product on the purchase of another product. Earlier we used the example of IBM requiring its tabulating machine customers to buy its tabulating cards from IBM. Many similar examples have arisen in antitrust cases. Some examples include the tie-in of salt to salt dispensers, ink to duplicating machines, cans to can-closing machines, and staples to stapling machines.

Such examples generally have the characteristic that the customer buys or leases a "machine" and then must purchase the inputs that are used with the machine from the same supplier. The inputs used with the machine will vary with the intensity of use that various customers make of the machine. This variable proportions case is only one type of tying arrangement.

Tying in terms of fixed proportion also occurs. For example, a movie distributor may require a theater owner to take movie B if he wants movie A. This arrangement is generally referred to as block booking. A more recent example is Microsoft requiring that its operating system be purchased with Internet Explorer.

Historically, the courts have viewed tying as a device for extending monopoly over one product, such as duplicating machines, to the tied product, ink. This is known as the "leverage theory" of tying. In a 1912 case, Chief Justice Edward D. White offered the following observations on the danger of tying (the judge was distressed that the majority opinion declared the tie to be legal):

Take a patentee selling a patented engine. He will now have the right by contract to bring under the patent laws all contracts for coal or electrical energy used to afford power to work the machine or even the lubricants employed in its operation. Take a patented carpenter’s plane. The power now exists in the patentee by contract to validly confine a carpenter purchasing one of the planes to the use of lumber sawed from trees grown on the land of a particular person. . . . My mind cannot shake off the dread of the vast extension of such practices which must come from the decision of the court now rendered. Who, I submit, can put a limit upon the extent of monopoly and wrongful restriction which will arise.34

An empirical problem with this argument is that it does not fit the facts of many cases against tying:35

Can it sensibly be accepted that G.S. Suppiger Co. tied salt to its salt-dispensing machinery as part of a scheme to monopolize the American salt market? Did Morgan Envelope tie its toilet paper to its dispenser as part of a grand scheme to monopolize the American bathroom tissue market? Why do we

see again and again . . . cases involving the tying of rivets, staples, windshield wipers, repair parts, varnish, etc., when the tying monopolist’s share of the market for the tied product remains minuscule?

We now know the logic of the original leveraging argument is flawed and for the same reasons that the argument for vertical integration being an extension of monopoly is flawed (assuming fixed proportions production). Suppose a consumer is interested in purchasing a system, such as a stereo, that requires components $A$ and $B$. A consumer is willing to pay up to $500 for this system. Product $A$ is produced by a monopolist at cost 175, while the market for product $B$ is competitive with unit cost of 100. In the absence of tying, product $B$ would be priced at 100 and the monopolist would price product $A$ at 400 so that the total expenditure to the consumer would be exactly at the consumer’s maximum willingness to pay. The monopolist’s profit would be 225. Now consider the monopolist engaging in tying by producing both components and selling systems to consumers. The most the monopolist can charge is 500, and it costs the monopolist 275 to produce it, so the monopolist makes 225; the same as without tying. The point is that there is one monopoly profit and the monopolist can, under certain conditions, extract it without tying.

In fact, tying can actually reduce profit. For suppose the version of component $B$ produced by other firms, let us call it $B^*$, is superior to that produced by the monopolist, so that consumers are willing to pay up to $550 for a system with components $A$ and $B^*$. With competition, the price of $B^*$ is at its cost of 100, in which case the monopolist can charge 450 for product $A$ and earn profit of 275. If it engages in tying, then it can only charge a price of 500 for the system, as it offers the inferior component $B$. Its profit declines to 225. Key here is that its market power for component $A$ allows the monopolist to extract the additional surplus created by the superior complements offered by competitors. That additional surplus vanishes with tying, and with it vanishes some of the monopolist’s profit.

Of course, tying has been used by firms, so if it is not to extend monopoly power, then what is its purpose? To begin, there are clearly efficiency reasons for some physical ties; that is, the products are physically (and not contractually) linked. There are many such examples, including the automobile. It is possible to imagine a car sold as a group of separate products: basic automobile, tires, radio, battery, and so forth. Since consumers are interested in the “package,” transactions costs are reduced by the tie-in.

Another efficiency rationale that has been used by defendants in tying cases is one of quality control—the tied good is necessary for the satisfactory performance of the tying good. IBM argued that it had to tie its cards to its tabulating machines because inferior cards would cause the machines to malfunction, causing a loss of goodwill from its customers. Of course, if this claim is correct, such tying is socially beneficial. The courts have generally not agreed with this argument, however, and have observed that the manufacturer of the tying good could simply state the specifications necessary for the tied goods. It would, of course, be in the interests of the customers to use only the “proper” tied goods.
In *Siegel et al. v. Chicken Delight, Inc.*, a 1971 case decided by a Circuit Court of Appeals, Chicken Delight used such a quality control defense unsuccessfully.\(^{36}\) Chicken Delight licensed several hundred franchisees to operate its stores. It did not charge its franchisees a franchise fee or royalties; rather, it allowed its franchisees to use its trademark and follow its business methods in exchange for purchasing their cooking equipment and certain supplies from Chicken Delight. It was also the case that the prices for these purchases were higher than the prices charged by other suppliers. The court held that Chicken Delight could have achieved the necessary quality control by specification of the appropriate cooking equipment and supplies. It was therefore unnecessary, in the court’s view, for Chicken Delight to require purchases of these items from itself.

There are several nagging questions about the view that simply stating the required specifications is a perfect substitute for the tie-in. One point is that it may be costly to convince buyers of the need for the specifications stated when cheaper alternatives exist. Another point is that Chicken Delight might have a “free-rider” problem. The reputation of Chicken Delight could be damaged if a few franchisees decided to use cheap, low-quality equipment and supplies knowing that customers in general would identify with the regional reputation of Chicken Delight for good quality. That is, the few franchisees using inferior supplies would continue to have customers who relied on the overall quality of all Chicken Delight stores (even though their loyal repeat business might be small). Hence, these franchisees could free-ride off the high quality of the rest, and tying the supplies might be a way of combating this problem.

A successful defense using the quality control argument was made in a 1960 case involving a cable television system supplier, Jerrold Electronics.\(^{37}\) Jerrold sold only on a complete systems basis, including installation, equipment, and maintenance. However, the legality of Jerrold’s tying was restricted to the “early years” of the industry, when the technology was in its infancy. After the technology had been in existence for a number of years, ensuring the availability of competent independent suppliers and service personnel, such tying was no longer legal.

A third argument for tying is that it is a way in which to evade price regulation. For example, when gasoline was under maximum price controls in the 1970s, the excess demand was great. Cars lined up for miles in certain cities to buy gasoline. Because the price could not be increased to clear the market, a gasoline station might tie its gasoline to other products or services to avoid the price ceiling. For example, one station was alleged to have offered gasoline at the controlled price to anyone who purchased a rabbit’s foot for $5!


Perhaps the most convincing efficiency rationale for contractual tying is that it is a form of price discrimination. In that the argument is a bit subtle, we’ll take our time developing it in the next sub-section. Price discrimination is not necessarily anti-competitive and, in fact, generally raises social welfare though perhaps benefiting firms at the cost of consumers. Are there any (valid) anticompetitive theories for tying? Only in the last 15 years have some emerged, and we will review them as well.

**Price Discrimination**

A useful way to think about price discrimination can be shown by example. In figure 8.6 we depict the usual profit-maximizing monopolist equilibrium where the monopolist is permitted to select a single price. The solution is determined by the usual marginal revenue (MR) equals marginal cost (MC) condition, and the price equals \( P^* \). The monopolist’s profit is the area \( A B C P^* \). As we explained in chapter 4, the area under the demand curve equals the total willingness-to-pay by consumers, and the area under the marginal cost curve equals total cost. This implies that total “potential profit” is larger than the actual profit by the amount of the two shaded triangles. (Total potential profit is equal to the area of triangle \( RSC \); it would be the profit under perfect price discrimination.) In other words, it is in the monopolist’s interest to try to extract a larger profit by price discrimination.

In many of the tying cases, the firm practicing tying has either a patent monopoly or some market power over the tying product. Hence it is useful to think in terms of tying as a pricing scheme designed to extract more of the consumers’ surplus—or to appropriate some of the triangular shaded areas in figure 8.6.

A simple block-booking example illustrates the point. Assume that the maximum values to theater owners for two movies are as follows:

<table>
<thead>
<tr>
<th>Maximum Value to Theater Owners</th>
</tr>
</thead>
<tbody>
<tr>
<td>Movie A</td>
</tr>
<tr>
<td>Fox Theater</td>
</tr>
<tr>
<td>York Theater</td>
</tr>
</tbody>
</table>

To obtain the maximum revenue, the movie distributor has several possibilities, although some may be ruled out because of legality or nonfeasibility. First, perfect price discrimination would entail charging separately the maximum value for each movie to each individual:

Perfect Price Discrimination: Revenue = $100 + $70 + $60 + $80 = $310.

38. See George J. Stigler, *The Organization of Industry* (Homewood, Ill.: Richard D. Irwin, 1968), chap. 15, for the original analysis of block booking.
Thus the maximum potential revenue is $310. Charging separate prices may not be possible, though. Assume then that the distributor can charge only one price for each movie. An examination of the values in the table indicates that the best he could do would be to charge $60 for movie A and $70 for movie B. This “normal” pricing outcome would yield:

**Normal Pricing Case:** Revenue = $60 + $60 + $70 + $70 = $260.

There is one further possibility—block booking. Suppose that the distributor offers a bundle of movies A and B for a single price. The bundled price for movies A and B to the Fox Theater could be $170, but this would cause the York Theater to decline the bundle and generate a total revenue of only $170. Hence the best bundled price would be $140, inasmuch as this would keep both theaters as customers:

**Block Booking Case:** Revenue = $140 + $140 = $280.
Of course, the point is that block booking yields higher revenue than normal pricing. This approach does not always work, however. In this case, Fox will pay more for A than York will, and York will pay more for B than Fox. If, for example, Fox will pay more for both movies, block booking gives results identical to normal pricing.

The courts take a harsh view of block booking. In the Loew’s, Inc., case that was decided by the Supreme Court in 1962, six major film distributors of pre-1948 copyrighted films for television exhibition were found guilty of block booking. As one example, Associated Artists Productions negotiated a contract with one television station for $118,800 in which the station had to take a bundle of 99 films. To get *Casablanca*, for example, the station had to take *Tugboat Annie Sails Again*.

According to Justice Goldberg:

This Court has recognized that “tying agreements serve hardly any purpose beyond the suppression of competition.” They are an object of antitrust concern for two reasons—they may force buyers into giving up the purchase of substitutes for the tied product, and they destroy the free access of competing suppliers of the tied product. . . . The standard of illegality is that the seller must have “sufficient economic power with respect to the tying product to appreciably restrain free competition in the market for the tied product. . . .” The requisite economic power is presumed when the tying product is patented or copyrighted.

Hence the Court is concerned that tying can foreclose rivals from the tied market and this result is harmful to competition. Again, our discussion earlier in this chapter on foreclosure applies here as well. The question remains of course as to the appropriate public policy if we accept the view that the fundamental explanation is price discrimination. Price discrimination is known to have ambiguous welfare effects; in some cases price discrimination raises total economic surplus and in others it has a negative effect. In cases involving patents there is an additional consideration: it may increase the appropriability of the returns to innovation and therefore be socially beneficial. (Of course, one must hold the belief that generally innovators are able to appropriate an inadequate fraction of the social returns to innovation.) We will consider the welfare effects of price discrimination further in our next example.

We now turn to a highly simplified illustration of tying of the variable-proportions type. The example is of a monopolist of copying machines who has two potential customers with different preferences for copying services. This difference in preferences is an essential part of the rationale for tying. The general idea is that tying gives the monopolist the ability to tailor his prices to “fit” his customers better than if he could charge only a single price to everyone.


The monopolist has constant costs of producing copying machines of $1,000 per unit. The customers derive no utility from the machines but only from the copying services that they produce in combination with paper. The number of packages of paper can be assumed to measure the quantity of services consumed by the customers. Hence we assume that the two consumers have the demand curves for copying services (paper) as shown in figure 8.7:

Demand by Customer 1: \( q_1 = 100 - p_1 \)

Demand by Customer 2: \( q_2 = 200 - 2p_2 \)

For convenience, we will assume that paper is supplied competitively and at a price of zero (the zero price makes the calculations easier). Hence, consider the monopolist’s problem in confronting the two demand curves in figure 8.6. Ignoring income effects, the areas under the demand curves and above the horizontal axes represent the consumer surpluses. That is, with the price of paper equal to zero, the areas give the surpluses of the consumers from copying services. These are shown as $5,000 and $10,000. So the monopolist could charge $5,000 per machine and sell to both customers, or charge $10,000 and sell to only customer 2. That is, customer 1 would not pay any more than $5,000 for a copying machine, since $5,000 extracts his total surplus. (We assume implicitly that the monopolist cannot charge...
separate prices of $5,000 to customer 1 and $10,000 to customer 2.) The two cases give the following profits:

Profit at Machine Price of $5,000 = 2($5,000 − $1,000) = $8,000

Profit at Machine Price of $10,000 = $10,000 − $1,000 = $9,000.

Hence the monopolist would do better by selling at a price of $10,000 and forcing the first customer out of the market.

Now, assume that the monopolist decides to practice tying. He can buy the paper on the market (at the zero price) and mark it up to sell to his copying machine customers. That is, the monopolist simply says that he will now charge a fixed price $P$ for the machine and a price per unit for paper $p$. All paper must, of course, be purchased from the monopolist even though it is cheaper in the competitive market. This requirement may present enforcement problems for the monopolist, but we will ignore them here. (It is also necessary to ensure that the two customers do not get together and share one machine.)

Figure 8.8 shows the profit-maximizing solution. The monopolist should charge a machine price of $2,812.50 and a paper price of $25. As shown in the figure, the first customer will buy 75 packages of paper at the $25 price. The consumer surplus is then $2,812.50, which is extracted completely as the price of the machine. The second customer buys 150 packs and also pays the $2,812.50 price for the machine. Hence, total profit under tying is

Profit under Tying: $2($2,812.50 − $1,000) + $25(75 + 150) = $9,250.

The first term is the profit from the machine sales and the second term is the profit on paper. The point, of course, is that tying permits the monopolist to extract a higher overall profit. Tying has permitted the monopolist more flexibility: he can lower the machine price, thereby attracting customer 1 into the market, and make up for lowering the machine price by making profits on paper sales.

Notice also that the monopolist is no longer limited to obtaining equal revenues from both customers. Under the solution in figure 8.8, the monopolist gets $4,687.50 from customer 1 and $6,562.50 from customer 2. They pay equal machine prices, but customer 2 buys more

41. The solution is found as follows. To keep both customers in the market, the machine price should equal the consumer surplus (after the price of paper is increased) of customer 1. Hence, express the machine price as the area of the consumer surplus triangle: $\frac{1}{2}(100 − p)(100 − p)$. Then profit can be written as a function of the price of paper only:

$(100 − p)(100 − p) + p(300 − 3p) − 2,000.$

The first and second terms are revenues from machines and paper, respectively, and the last term is the machine costs. The value of $p$ that maximizes profit is $25. Note that the best profit with only one customer is $9,000, and tying would not be employed.
paper because of its higher demand for copying services. Hence the paper plays the role of metering the demand for copying services, where the customer with the higher demand pays more. In principle, the tying of paper would be irrelevant if an actual meter could be used to record usage and a fee could be charged based upon usage. This means that tying here is a form of two-part pricing, where a fixed charge is made plus a price per unit. It is a type of pricing often used by public utilities.

Turning now to the public policy concern of whether tying is socially harmful, we can calculate total economic surplus with and without tying. First, assume the case of no-tying. The monopolist would choose to charge $10,000 for the machine, as already explained. Hence there is a consumer surplus of $10,000 (captured by the monopolist from customer 2—customer 1 stays out of the market), and the monopolist incurs costs of $1,000. The welfare measure is

\[
\text{Total Surplus (No-Tying)} = 10,000 - 1,000 = 9,000.
\]

Allowing tying, the total surplus can be seen easily by referring to figure 8.8. It equals the two consumer surplus triangles ($2,812.50 for customer 1 and $5,625 for customer 2) plus
the two areas representing payments for paper ($1,875 for customer 1 and $3,750 for customer 2) less the costs for two machines ($2,000). The welfare measure is therefore

\[ \text{Total Surplus (Tying)} = \$12,062.50. \]

Hence, for this particular example, tying leads to a higher total surplus. However, this is not a general finding. As is often true, price discrimination can either increase welfare or decrease it, depending on the particular situation. For the interested student, an example that leads to the opposite result is simply to change customer 2’s demand to \( q_2 = 130 - p_2 \). In this new situation, total surplus is \$11,450 from no-tying and only \$11,225 for tying. This conclusion is true despite the fact that tying is more profitable for the monopolist.\(^{42}\) A key difference between the two situations is that in the latter case the consumers have more similar demand curves, and the no-tying solution keeps both customers in the market. Note that if no-tying keeps both consumers in the market, tying must cause total surplus to fall because consumers go from a paper price equal to marginal cost to a price above marginal.

**Modern Theories of Leveraging**

In recent years, economists have developed new theories—some clearly motivated by the Microsoft case—which show how tying can be anticompetitive. One argument supposes the tied product, in addition to being valued as a complement to the monopolized product, is intrinsically valued by some consumers. For example, consider the original Sony Walkman, which had a monopoly on portable tape players. Headphones are used with a Walkman, but a consumer could buy headphones to use with other stereo equipment. There are then two markets: the systems market, whereby a consumer buys a Walkman and headphone together, and the stand-alone market, in which a consumer buys only the headphones. A second way in which tying can be anticompetitive is when a firm supplying the tied good may also enter the monopolist’s primary market. In these cases, tying may be profitable either because it allows a monopolist to extend its market power to the tied good or because tying serves to protect its monopoly market from competition.\(^ {43}\)

**Tying That Extends a Firm’s Monopoly Power**

Consider the following simple model with three products: \( A, B_1, \) and \( B_2.\)\(^ {44}\) Product \( A \) is of value only in conjunction with \( B_1 \) or \( B_2. \) Thus, \( A \) and \( B_1 \) are complements as are \( A \) and

\(^{42}\) In this case, the tying solution is to charge \$3,612.50 for the machine and \$15 for paper. The profit is \$8,225 as compared with only \$8,000 under no-tying.

\(^{43}\) A survey of these theories can be found in Patrick Rey and Jean Tirole, “A Primer on Foreclosure,” in *Handbook of Industrial Organization*, vol. 3 (University of Toulouse, July 2003).

We refer to the market in which consumers purchase $A$ as the systems market since consumers are interested in purchasing a system of $A$ and $B1$ or $A$ and $B2$. In the systems market, there is one type of consumer whose maximum willingness to pay (MWP) is $v$ for $A/B1$ (which means the system comprised of $A$ and $B1$) and $v - d$ for $A/B2$. There is also a stand-alone market for $B1$ and $B2$. A consumer in that market is interested in $B1$ or $B2$ but not both and doesn’t care about $A$. Her MWP for $B1$ is $w$ and for $B2$ is $w + e$. Since $d \geq 0$ and $e \geq 0$ then $B2$ is superior to $B1$ as it is more desired by consumers either in isolation or as part of a system.

There are two firms in these markets. Firm 1 produces $A$ and $B1$, while firm 2 produces $B2$. Firm 1 then has monopoly power in the systems market, while the two firms compete in the stand-alone market. Assume there are $m$ consumers in the systems market and $n$ consumers in the stand-alone market. The cost to producing $A$ is $c_A$ and to $B1$ or $B2$ is $c_B$.

First consider an equilibrium when there is no tying which means that firm 1 offers $A$ and $B1$ (as separate products) and firm 2 offers $B2$. A consumer in the systems market buys either $A/B1$ or $A/B2$. Letting $p_i$ denote the price charged for product $i$, we will show that one equilibrium has prices of $p_{B1} = c_B$, $p_{B2} = c_B + e$, and $p_A = v + d - c_B - e$, and all consumers in both markets buying $B2$. Firm 1 charges cost for $B1$, firm 2 charges a premium for $B2$ by an amount equal to (or just a little below) the additional surplus it provides to consumers over $B1$, and firm 1 charges a price for $A$ so that the price of $A/B2$ just equals the MWP of $v + d$. The profit of firm 1 is $m(v + d - c_B - e - c_A)$ and of firm 2 is $e(n + m) - k$, where $k$ is a fixed cost incurred by firm 2.

To show this is an equilibrium, let us first derive conditions whereby consumer behavior is optimal. In the stand-alone market, a consumer can buy $B1$ and receive surplus of $w - c_B$ or buy $B2$ and receive surplus of $w + e - (c_B + e) = w - c_B$. Being indifferent between these two alternatives, a consumer is content to buy $B2$. (As we mentioned above, we can also have firm 2 pricing a little below $c_B + e$, so that consumers strictly prefer to buy $B2$.) In the systems market, consumers are presumed to buy $A/B2$, and this yields surplus of $v + d - (v + d - c_B - e) - (c_B + e) = 0$. They could instead buy $A/B1$, which has surplus of $v - (v + d - c_B - e) - c_B = e - d$; in exchange for forgoing the additional surplus of $d$ from the superior complement to $A$, consumers pay $e$ less. Let us assume $d > e$, so that consumers prefer to purchase $A/B2$. This assumption says that the incremental value of $B2$ (compared to $B1$) is higher in the systems market than in the stand-alone market.

We next need to show that each firm’s price is optimal given the other firm’s price and that both firms earn non-negative profit. Toward this end, first note that consumers in the systems market are buying $A/B2$ and paying a price exactly equal to their MWP. Since a higher price

---

45. A variant of the ensuing argument works even if the products are independent rather than complements.

46. There are other equilibria in which $p_{B1} = c_B$, $p_{B2} = c_B + \lambda d$ and $p_A = v + (1 - \lambda)d - c_B$ where $0 \leq \lambda \leq 1$, so that firm 1 extracts a bigger share of the surplus from system $A/B2$. To simplify the analysis, we have set $\lambda = 1$. 
for either $A$ or $B_2$ would induce them to demand zero, firm 1 does not want to price any higher, as then its demand and profit would be zero. And it does not want to price any lower since it is already selling to everyone in the systems market. Clearly, firm 1’s price for $A$ is optimal.

In considering the optimality of the prices in the stand-alone market, firm 1 does not want to price below its cost, as that would incur losses, and pricing above cost would still result in zero demand. Given that firm 1 prices at cost, a consumer always has the option of buying $B_1$ and receiving surplus of $w - c_B$. Since the surplus from buying from firm 2 is $w + e - p_{B_2}$, a consumer prefers to buy $B_2$ only when

$$w + e - p_{B_2} \geq w - c_B \text{ or } c_B + e \geq p_{B_2}.$$  

Thus, firm 2 does not want to price above $c_B + e$, as then its demand is zero, and there is no reason for it to price any lower, since it is already selling to all consumers in both markets when it prices at $c_B + e$. Firms’ prices in the stand-alone market are then optimal.

The final condition to check is that both firms are earning non-negative profit. For firm 1, this requires $v + d - c_B - e - c_A > 0$ which is true as long as $v$ is sufficiently large. For firm 2, it requires $e(m + n) - k > 0$. Assume both of those conditions hold. This completes the proof that, in the absence of tying, it is an equilibrium for firm 1 to price $A$ at $v + d - c_B - e$ and $B_1$ at $c_B$ and firm 2 to price $B_2$ at $c_B + e$. At those prices, all consumers in the systems market buy $A/B_2$ and all consumers in the stand-alone market buy $B_2$.

Now suppose firm 1 engages in tying by offering a physically bundled product of $A$ and $B_1$, as well as $B_1$ separately. A consumer in the systems market either buys $A/B_1$ or nothing. It should be fairly clear that the equilibrium prices are $p_{B_1} = c_B$, $p_{B_2} = c_B + e$, and $p_{A/B_1} = v$, with all consumers in the systems market buying $A/B_1$ and all consumers in the stand-alone market buying $B_2$. Consumers in the systems market can only buy $A/B_1$, so firm 1 prices it at their MWP of $v$. Prices of $p_{B_1} = c_B$ and $p_{B_2} = c_B + e$ form an equilibrium in the stand-alone market for the same reason as without tying. The resulting profit for firm 1 is $m(v - c_A - c_B)$ and for firm 2 is $ne - k$.

Assume $(m + n)e - k > 0 > ne - k$ or, equivalently, $(m + n)e > k > ne$, so that firm 2 earns positive profit without tying but negative profit with tying. Hence, if firm 1 engages in tying, then firm 2 exits (or, if firm 2 is contemplating entry, it chooses not to enter when it sees that product $B_1$ is tied with $A$). In that case, firm 1 is left as a monopolist in both markets, in which case its profit-maximizing prices are $p_{B_1} = w$ and $p_{A/B_1} = v$, with profit of $m(v - c_A - c_B) + n(w - c_B)$.

47. Given the assumption $d > e$, one can prove that firm 1 does not find it optimal to provide a discount to consumers who buy $A$ and $B_1$.

48. To simplify matters, we will assume that a consumer in the systems market cannot buy $A/B_1$ and $B_2$ and then replace $B_1$ with $B_2$. Doing so incurs additional expenditure but would increase the value of the system from $v$ to $v + d$. Our conclusions are robust to that possibility, but allowing for it would introduce unnecessary complications.
Firm 1 earns higher profit by tying when
\[ m(v - c_A - c_B) + n(w - c_B) > m(v + d - c_B - e - c_A) \text{ or } n(w - c_B) > m(d - e). \] (8.7)

Equation 8.7 holds as long as the value to the stand-alone product, \( w \), is sufficiently high and/or the size of the stand-alone market, \( n \), is sufficiently large relative to the size of the systems market, \( m \). Both conditions make sense, as the payoff to tying by firm 1 is that it becomes a monopolist in the stand-alone market, and the associated profit from doing so depends on how much consumers are willing to pay in that market as well as how big that market is. The cost to firm 1 of tying is that it does not make as much profit in the systems market since, with the elimination of the superior complement \( B_2 \), systems consumers are not willing to pay as much for \( A \). This incremental value of \( B_2 \) to a system is measured by \( d \), but firm 2 is able to extract \( e \) of that, so the per unit net effect on firm 1’s profit is \( d - e \).

In sum, we have derived conditions whereby the monopolist earns higher profit by tying products \( A \) and \( B_1 \). What makes tying work is that it forecloses firm 2 from the systems market; those consumers no longer buy component \( B_2 \) to use with component \( A \). This leads to a reduction in firm 2’s variable profit to a level insufficient to cover its fixed cost. As firm 2 then exits, firm 1 is left with a monopoly in the systems market but also in the stand-alone market. If the latter market is sufficiently big and the profit margin, \( w - c_B \), is sufficiently large, then tying is profitable.

We have shown that a monopolist may engage in tying in order to extend its monopoly, but is it anticompetitive? In addressing that question, note that, in this simple model, all consumers buy, whether or not there is tying.\(^49\) Since all consumers always buy, welfare equals the value of the goods produced less the cost of producing them. It is then lower with tying when:
\[ m(v + d - c_A - c_B) + n(w + e - c_B) - k > m(v - c_A - c_B) + n(w - c_B) \text{ or } md + ne > k. \]

The welfare-reducing effect of tying is that the superior product of firm 2 is not supplied. This reduces total surplus by \( md + ne \), which is the reduction in value from consumers using \( B_2 \) rather than \( B_1 \). Tying does have the beneficial effect of avoiding the fixed cost \( k \) incurred by firm 2. Tying then reduces welfare when \( md + ne > k \).

**Tying That Protects a Firm’s Primary Market**

Essential to the above anticompetitive rationale for tying is the existence of a stand-alone market for the tied good. As we will now show, a stand-alone market is not necessary for a monopolist to find tying profitable when it is threatened with entry in its primary market as

49. In a more general model, tying would influence how many consumers buy by affecting the price of a system and the stand-alone product.
well. Tying can occur not to extend monopoly power but rather that to protect existing market power.\textsuperscript{50}

If there is no stand-alone market for \(B1\) and \(B2\) in the preceding model \((n = 0)\), an equilibrium is of the form: \(p_{B1} = c_B\), \(p_{B2} = c_B + \lambda d\), and \(p_A = v + (1 - \lambda)d - c_B\), and all consumers buy \(A/B2\).\textsuperscript{51} System \(A/B2\) is worth \(v + d\) to consumers and they pay a price equal to that value, with \(v + (1 - \lambda)d - c_B\) of the revenue going to firm 1 and \(c_B + \lambda d\) going to firm 2. To reduce the amount of notation, assume \(m = 1\). The resulting profit is then \(v + (1 - \lambda)d - c_B - c_A\) for firm 1 and \(\lambda d\) for firm 2.

Recall that without the stand-alone market, tying is unprofitable. If firm 1 ties, then firm 2 is entirely foreclosed. Firm 1 sells \(A/B1\) at a price of \(v\) and its profit is \(v - c_B - c_A\) which is lower than \(v + (1 - \lambda)d - c_B - c_A\). Tying is unprofitable because firm 1 no longer shares in the additional surplus created by firm 2’s superior complement to firm 1’s product \(A\).

Allowing for multiple periods will be important for the ensuing argument, so let us assume there are two periods (with no discounting). For when there is no tying, the total profit to firm 2 from entry is then \(2\lambda d - k\), which is assumed to be positive. Thus, in the absence of tying, entry by firm 2 occurs.

Now we come to the major change in the model. Firm 2 can offer not only a version of product \(B\) but also of product \(A\). For clarity, we will now refer to firm 1’s version of \(A\) as \(A1\) and use \(A2\) to denote firm 2’s version. For technological reasons, firm 2 cannot begin supplying \(A2\) until the second period, though it can supply \(B2\) in both periods. Thus, if firm 2 enters, it will provide a superior complement to firm 1’s product in period 1, but it could provide a superior system, by offering the bundle \(A2/B2\), in period 2. Assume consumers attach value \(v + f\) to \(A2/B2\), where \(f \geq d\). (Thus, \(A2\) does not need to be superior to \(A1\).) Finally, the unit cost of producing \(A2\) is also \(c_A\).

Suppose there is no tying, and firm 2 enters. It supplies \(B2\) in period 1 and earns profit of \(\lambda d\) in that period. Come period 2, each firm can supply a system, so we can think about them competing in terms of systems. Firm 1 offers \(A1/B1\), to which consumers attach value \(v\), and firm 2 offers \(A2/B2\), with higher value \(v + f\). The equilibrium then has firm 1 pricing at cost and firm 2 charging a premium equal to the additional surplus that its system provides: \(p_{A1/B1} = c_A + c_B\) and \(p_{A2/B2} = c_A + c_B + f\). Hence, in period 2, firm 1’s profit is zero and firm 2’s profit is \(f\), so its total profit from entry is \(\lambda d + f - k\), which is positive (since it was already assumed that \(2\lambda d - k > 0\) and \(f \geq d \geq \lambda d\)).

Now consider tying in this setting. Tying prevents firm 2 from selling \(B2\) in period 1. Its profit from entry is then \(f - k\), as it dominates the systems market in period 2. Let us then


\textsuperscript{51} The equilibrium analyzed for the preceding model was the case of \(\lambda = 1\).
assume $f - k < 0$, which has the implication that tying deters entry by firm 2, as it cannot make enough money in period 2 to cover its entry cost. With firm 2 absent from the market, firm 1 prices at $v$ in each period so its total profit is $2(v - c_A - c_B)$. Tying is then profitable if

$$2(v - c_A - c_B) > v + (1 - \lambda)d - c_B - c_A \text{ or } v - c_A - c_B > (1 - \lambda)d$$

which holds if $\lambda$ is close enough to one, which means that, in the absence of tying, firm 1 gets a sufficiently small share of the additional surplus from firm 2 supplying $B2$.

As with the first theory of anticompetitive tying, tying is profitable for an incumbent firm because, by foreclosing part of an entrant’s market, it makes entry unprofitable. Here, tying forecloses the market for the entrant’s complement in the first period while, in the previous model, it foreclosed the systems market. A more important distinction is that the objective of tying is not to extend monopoly power into another market but rather to protect the incumbent firm’s monopoly position in market $A$. Finally, note that the change in welfare from tying is $k - f - d$, as the value of the system being supplied is reduced by $d$ in period 1 and $f$ in period 2. If $f + d > k$, then tying deters entry and reduces welfare.

Instrumental in tying as an entry deterrence device is that an entrant has a limited number of periods for which it can earn profit to cover its entry cost. Where such a force may be relevant is in highly innovative markets where a new product may already have a relatively short technological life span as newer products come along to replace it. Tying that reduces the lifetime of an entrant’s product, where the lifetime is already short, may indeed serve to deter entry. In chapter 9, we will see how this tying argument pertains to the Microsoft case.

**Antitrust Law and Policy**

*International Salt Co., Inc., v. United States*, decided in 1947, is a landmark case and has usually been cited in subsequent tying cases. The International Salt Company had a patent over salt-dispensing machines used in food processing. The company required all users of the machines to buy their salt from the company as well. They argued that only their salt was of sufficient quality to function properly in their machines, and the tie-in was necessary to preserve goodwill. The Supreme Court, however, disagreed:

If others cannot produce salt equal to reasonable specifications for machine use, it is one thing; but it is admitted that, at times, at least, competitors do offer such a product. They are, however, shut out of the market by a provision that limits it, not in terms of quality, but in terms of a particular vendor. Rules for use of leased machinery must not be disguised restraints of free competition.

In a 1958 case, *Northern Pacific Railway Company et al. v. United States*, the Court spelled out its rule toward tying as what might be termed a “modified” per se rule. The case

---

In the case of vertical mergers and vertical restraints, a railroad selling land along its right of way on the condition that the buyer ship over the railroad’s line. A difference between this case and *International Salt* is that the market power in the salt dispenser case was due to a patent. Here, the tying product was land, and the railroad was found to have sufficient market power to find the tie-in sale illegal:

Tying arrangements deny competitors free access to the market for the tied product. . . . At the same time buyers are forced to forgo their free choice between competing products. For these reasons “tying agreements fare harshly under the laws forbidding restraints of trade.” They are unreasonable in and of themselves whenever a party has sufficient economic power with respect to the tying product to appreciably restrain free competition in the market for the tied product and a “not insubstantial” amount of interstate commerce is affected.

An interesting and important case decided in 1984 by the Supreme Court involved the tying of anesthesiological services to hospital services.54 The so-called *Hyde* case was a private lawsuit in which Dr. Edwin Hyde, an anesthesiologist, charged the East Jefferson Hospital in the New Orleans area with tying the services of a firm of anesthesiologists, Roux & Associates, to its hospital services. As a result of the tie-in, Dr. Hyde was denied permission to practice at the hospital.

The hospital had signed an exclusive contract with Roux & Associates for all of its anesthesiological services. An important institutional fact was that every patient who had surgery there paid Roux & Associates directly for their services. This is important because anticompetitive tying is supposed to involve the monopolist of the tying product (East Jefferson Hospital) increasing its profits through the tie of a second product (the services provided by Roux) over which it does not initially have a monopoly. As Lynk explains in his analysis of this case: “[anticompetitive tying] is not likely to involve the channeling of that profit into someone else’s pocket.”55 In short, there is a real problem in the application of the standard leverage theory of tying to this case.

The Court unanimously found for the defendant, though there was a difference of opinion among the Justices as to the reason. The majority opinion of five Justices, written by Justice Stevens, held that there was no basis for a finding that the hospital had market power. There was actually very little evidence in the record on this issue. Recall that the modified per se rule for tying requires that there be market power over the tying good. The four other Justices issued a concurring opinion, written by Justice O’Connor, but disagreed on the market power issue. They found that even though there was market power, there “is no sound economic reason for treating surgery and anesthesia as separate services” and therefore “the Hospital can acquire no *additional* market power by selling the two services together.”

The four Justices also indicated their desire to drop the per se rule approach and move to a rule of reason for tying. Given the theoretically diverse welfare outcomes for tying, this change would seem to be a good idea. Alternatively, an approach that was contained in the DOJ’s 1985 guidelines regarding tying would seem sensible. That is, according to those guidelines, it would use a two-step approach and drop any investigation in which the tying product’s market share was less than 30 percent. If the share was over 30 percent, they generally would follow a rule of reason analysis. Because they were controversial, President Clinton’s assistant attorney general for antitrust, Anne Bingaman, dropped the 1985 guidelines regarding tying and other vertical restraints in August 1993.

A 1992 tying case involving Kodak was also a Supreme Court decision, and it left the per se approach intact. The tie-in was a tie of repair services to parts for Kodak photocopiers. The tie-in excluded independent service companies from repairing Kodak photocopiers, and they brought the suit against Kodak. Though the market for photocopiers was agreed by all to be unconcentrated, the majority of Justices held that it was possible for Kodak to have a monopoly in parts. We will examine this controversial decision in detail in chapter 9.

The rule is that tying is illegal when the seller possesses sufficient market power in the market of the tying product and the amount of commerce involved is substantial. Of course, it is possible for the tying to be “reasonable” in the sense that it was in the Jerrold case described earlier.

Manufacturer-Retailer Restraints

While the previously discussed vertical restraints are applicable to many upstream-downstream relationships, here we turn to two practices that are especially if not exclusively relevant to manufacturers (or wholesalers) and retailers.

Resale Price Maintenance

As we have mentioned, RPM can either be a requirement by the supplier that the dealer not sell below a minimum price or a requirement that the dealer not sell above a maximum price. It is possible to describe certain situations where the supplier would prefer the minimum resale price, and others where the supplier would prefer the maximum resale price.

The simplest explanation is perhaps that pertaining to the desire of a supplier to require maximum resale prices. To understand this case the reader should refer to the discussion earlier in this chapter on double marginalization. There we discovered that vertical integration could lead to a lower final price and higher combined profits of successive monopolists. If the supplier and dealer both have market power, it is clear that the ability of the supplier to limit the dealer’s price will increase its own profitability. Incidentally, this will also improve society’s welfare, given that the supplier’s monopoly cannot be eliminated.

The explanation for the opposite type of RPM—setting minimum resale prices—is more complex. After all, it seems counterintuitive to think that a manufacturer might prefer higher distribution costs than competition among its retailers could bring about. However, minimum-price RPM might be wise in cases where the supplier wants to ensure the provision of certain presale informational services that are necessary for the marketing of technically complex products.

Consider a personal computer. Before buying an Apple computer, the consumer would like to learn as much about it as possible. A retail computer store that sells Apples is ideal—the consumer can consult with technically trained salespersons and operate the computer. Then, however, the consumer might decide to purchase the Apple through a mail-order outlet that has lower prices for Apples because the outlet need not provide floor space for demonstrations and technically trained salesmen. In other words, the mail-order outlets are free-riding on the retail computer stores. The concern of the supplier is that the mail-order outlets may make it unprofitable for the retail stores to continue to provide the informational services that are necessary for Apple to compete against IBM and other computer suppliers. This is a case where setting minimum resale prices may be sensible from the point of view of the supplier.

A graphical explanation may be helpful. In figure 8.9 a retailer is shown whose marginal cost \( MC \) is equal to the price charged by the manufacturer (for simplicity, we assume the retailer’s only cost is the manufacturer’s product). Vigorous competition from other retailers brings price down to \( MC \). Assume now that the manufacturer sets a minimum resale price at \( p_1 \), while continuing to sell to the retailer at \( MC \). This might seem irrational because the demand \( D \) would imply that quantity sold would fall to \( q_1 \). Hence the manufacturer would necessarily lose money.

The point, however, is that the retail margin \( p_1 - MC \) enforced by the manufacturer is expected to lead to promotional activities by the retailer, which shifts the demand to \( D' \). The net effect is shown in figure 8.9 as an increase in quantity to \( q' \).

While the particular case shown in figure 8.9 is one possibility, other cases are also plausible, since the demand shift can be of various magnitudes. A detailed analysis of various cases suggests that RPM can be either efficiency increasing or decreasing, depending on the magnitude of the assumed demand shift.

57. The “free-riding” argument was originally made by Lester G. Telser, “Why Should Manufacturers Want Fair Trade?” *Journal of Law and Economics* 3 (October 1960): 86–105. For an analysis of cases that do not seem to fit the free-rider theory of Telser, see Howard P. Marvel and Stephen McCafferty, “Resale Price Maintenance and Quality Certification,” *Rand Journal of Economics* 15 (Autumn 1984): 346–59. In particular, they argue that in many cases dealers do not provide tangible presale services; rather, their idea is that certain high-quality retailers—Macy’s, Neiman Marcus, etc.—“serve as the consumer’s agent in ascertaining the quality or stylishness of commodities.” The retailers who invest in “certifying” the quality of the goods are then subject to free riding by other retailers.

There can be other anticompetitive effects of RPM as well. It is conceivable that either a cartel of dealers or a cartel of suppliers might be fostered through RPM. For example, the dealers might get together and insist that the supplier require minimum resale prices for all dealers. This would be very helpful in making the cartel work. Of course, it would not be in the supplier’s best interest. It would also have to be a product that did not face substantial interbrand competition. For example, if Apple dealers could raise their prices, it might be quite profitable if they did not have to reckon with IBM, Compaq, Toshiba, Dell, and so on.

The point about interbrand competition should be amplified. That is, many cases of minimum RPM have to do with intrabrand competition. If all dealers selling Apples eliminate intrabrand competition, it is not likely to be an effective cartel because consumers can shift to other brands of computers. And reduced intrabrand competition might make Apple a more effective interbrand competitor.
Since 1911, however, minimum RPM has been a per se violation of Section 1 of the Sherman Act, despite the possible procompetitive arguments that we have presented. The key case was *Dr. Miles Medical Co. v. John D. Park & Sons*.\(^{59}\) Dr. Miles, a manufacturer of proprietary medicines, had established a set of minimum resale prices that applied throughout its distribution chain. John D. Park, a drug wholesaler, refused to enter into the restrictive agreements and instead was able to buy Dr. Miles’s products from other wholesalers at discounted prices. Dr. Miles brought a suit against John D. Park for interfering with the contracts between Dr. Miles and the other wholesalers. The Supreme Court held that the contracts were illegal, observing that “agreements or combinations between dealers, having for their sole purpose the destruction of competition and the fixing of prices, are injurious to the public interest and void.”

It should be pointed out that there must be a conspiracy between the manufacturer and dealer to fix the prices; it is not illegal for a manufacturer unilaterally to set resale prices and refuse to deal with retailers who do not comply. Two recent cases\(^{60}\) have established the standards necessary to infer such a conspiracy. The first case, *Monsanto v. Spray-Rite Service*, involved a Monsanto herbicide dealer selling at discount prices. There was evidence that other dealers had complained to Monsanto, and Monsanto subsequently terminated the dealer. The Court said that evidence of complaints was not sufficient unless additional evidence tended to exclude the possibility of independent action by Monsanto.

The second case, a Supreme Court decision in 1988, also supports the view that the conditions for RPM to be found illegal per se are quite restrictive. In *Business Electronics v. Sharp*, two retailers of electronic calculators, Business Electronics and Hartwell, and the manufacturer, Sharp, were involved in a dispute in the Houston area. Hartwell complained to Sharp about Business Electronics’ low prices—below Sharp’s suggested retail prices—and in June 1973 gave Sharp an ultimatum that Hartwell would terminate his dealership unless Sharp ended its relationship with Business Electronics. Sharp then terminated Business Electronics’ dealership in July 1973; Business Electronics then brought suit alleging a conspiracy that was illegal per se. Business Electronics won at the District Court level, but the decision was reversed by the Court of Appeals. The Supreme Court then agreed with the Circuit Court’s decision, which found that

to render illegal per se a vertical agreement between a manufacturer and a dealer to terminate a second dealer, the first dealer must expressly or impliedly agree to set its prices at some level, though not a specific one. The distributor cannot retain complete freedom to set whatever price it chooses.

---

Before leaving minimum RPM, it is worth describing its interesting history. During the Depression, individual states enacted “fair trade” laws that allowed minimum RPM. The reason was to “protect” small, independent retailers from the fierce price competition of the newly emerging chains. However, such laws generally affected goods in interstate commerce and as a result were in conflict with federal antitrust laws. In 1937, Congress passed the Miller-Tydings Act, which exempted state fair trade laws from the reach of the Sherman Act. By 1975 there was a strong lobby to repeal the fair trade laws, and the Consumer Goods Pricing Act was passed, which made the per se illegality of minimum RPM under the Sherman Act once again clear.

There is one other form of RPM that has drawn antitrust scrutiny, and that is when an upstream supplier limits the maximum price that a retailer can charge. In spite of the lack of compelling arguments at the time that maximum RPM could be anticompetitive, the Supreme Court ruled in 1968 in Albrecht v. Herald Co. that it was per se illegal! The Supreme Court changed its view in 1997 with State Oil Company v. Khan in concluding that maximum RPM is neither per se lawful nor unlawful, and that such practices should be evaluated under the rule of reason. It also took the opportunity to reaffirm the per se illegality of minimum RPM.

Though some economists thought that per se legality of maximum RPM was instead appropriate, there is at least one argument by which the welfare effects of maximum RPM are ambiguous. Consider the situation described above, in which retailers free-ride in terms of some nonprice variables that augment demand. The manufacturer would like to encourage more of that activity. One solution to this problem, which we will investigate next, is to give each retailer its own geographic monopoly. In that case, a retailer is assured of reaping the additional sales from its spending on promotion and service. But recall from our discussion of double marginalization that a downstream monopolist tends to charge too high a price from the perspective of the manufacturer. While the manufacturer likes marking up its price, it does not like the retailer doing it as well! Thus the creation of local monopolies to prevent free-riding has spawned double marginalization. A solution to it is maximum RPM. Given there are already territorial restrictions, the imposition of maximum RPM has two counteracting welfare effects: it results in a lower price but also less service (as the retailers now receive a lower price than when there was no maximal RPM). The net effect on consumers is ambiguous, and therein lies a rationale for using the rule of reason.

Territorial Restraints

Territorial restraints or, as the practice is also known, vertical market division, can be closely related to RPM. That is, earlier we discussed a hypothetical case of why Apple Computer might find minimum-price RPM a wise business decision, the reason being that the dealers would then be able to provide presale informational services to consumers without the free-rider problem. Allocating exclusive marketing territories to its dealers can operate in much the same way—each dealer would have some market power in its territory.

An important social benefit of territorial restraints is that distribution costs might be lowered by enabling each dealer to obtain scale economies. That is, by spreading fixed costs over a higher volume of sales, the costs of distribution can be reduced.

The potentially anticompetitive effects of territorial restraints are also similar to RPM—the fostering of collusive behavior among dealers or manufacturers. An interesting case of territorial restraints exists in the soft drink industry. The major soft drink syrup manufacturers—Coke, Pepsi, Dr. Pepper, and so on—allocate exclusive territories to their bottlers. A 1973 study by the FTC said that the reduced intrabrand competition had been costly to consumers because concentration at the syrup level was high. That is, Coke buyers could not benefit from competition among Coke bottlers, but only from competition among Coke, Pepsi, Dr. Pepper, and so on. The FTC was prevented from pursuing the case, however, when Congress passed legislation that specifically exempted the soft-drink industry’s territorial restrictions.

The cases on territorial restraints have led to a rule of reason approach as opposed to the per se illegality of RPM. Many economists believe this asymmetric treatment of economically similar practices to be wrong. Generally, the view of the critics is that both should be rule of reason offenses.

The major case is Continental T.V., Inc., et al., v. GTE Sylvania, Inc., which was decided by the Supreme Court in 1977. GTE Sylvania was a manufacturer of television sets that in 1962 had only 1 or 2 percent of the national market. In 1962, Sylvania began a new marketing plan. It phased out its wholesale distributors and began to sell its television sets directly

65. There is also an argument that reduced competition among sellers of the same brand can result in lessened competition among firms offering different brands; see Patrick Rey and Joseph Stiglitz, “The Role of Exclusive Territories in Producers’ Competition,” RAND Journal of Economics 26 (Autumn 1995): 431–51.
66. For a model allowing manufacturers to choose between minimum RPM and territorial restrictions (as well as some other forms of vertical restraints), see G. Frank Mathewson and Ralph A. Winter, “An Economic Theory of Vertical Restraints,” Rand Journal of Economics 15 (Spring 1984): 27–38. This theory provides a formal argument as to why these two practices should be treated comparably in terms of antitrust practice.
to a smaller and more select group of franchised retailers. The objective was to decrease the number of competing Sylvania retailers in the hope of attracting a smaller group of more aggressive and competent retailers who could increase Sylvania’s market share. To accomplish this end, Sylvania limited the number of franchises granted for any given region and required each dealer to sell only from the location at which the dealer was franchised. Interestingly, Sylvania retained the discretion to increase the number of retailers in a region depending upon the success of the retailers in developing the market.

In 1965, a Sylvania dealer in San Francisco, Continental T.V., wanted to open a store in Sacramento but was prohibited from doing so by Sylvania; Sylvania was doing exceptionally well in Sacramento and did not believe another dealer would be beneficial. As a result, Continental filed suit against Sylvania under the Sherman Act, Section 1.

The Supreme Court decided in favor of Sylvania and in so doing made it clear that the case should be decided on a rule of reason basis:

Vertical restrictions promote interbrand competition by allowing the manufacturer to achieve certain efficiencies in the distribution of his products. . . . For example, new manufacturers and manufacturers entering new markets can use the restrictions in order to induce competent and aggressive retailers to make the kind of investment of capital and labor that is often required in the distribution of products unknown to the consumer. Established manufacturers can use them to induce retailers to engage in promotional activities or to provide service and repair facilities necessary to the efficient marketing of their products. . . . Certainly there has been no showing . . . that vertical restrictions have or are likely to have a “pernicious effect on competition” or that they “lack . . . any redeeming virtue.”

Summary

This chapter has examined the benefits and costs of vertical integration and vertical restraints toward identifying when such activities are anticompetitive and should be prohibited. Vertical integration can increase the profit of the merging parties and also welfare more generally due to enhanced productive efficiency and reduced transaction costs by replacing market transactions with internal ones. Vertical integration can also make firms and consumers both better off by reducing the double markup that occurs when an input supplier with market power charges a price above cost and then the purchaser of the input marks it up again when it sets the final product price to exceed cost. Offsetting these benefits are the potentially anticompetitive effects of vertical integration. When both upstream and downstream markets are oligopolistic, a vertical merger can raise the cost faced by the remaining unintegrated downstream firms. This can make the merger profitable and result in consumers being worse off due to a higher final product price.

As is appropriate, the rule of reason must be used to trade off these benefits and costs in deciding whether a vertical merger should be allowed. The last fifty years have seen a tremendous change in antitrust policy as it has evolved to discouraging many vertical mergers in
the 1960s to where few were challenged by the authorities in the 1980s. Currently, the sentiment is that vertical integration is not a serious antitrust concern except where in one or both markets there is (or will be) a high level of market power.

We then turned to contractual arrangements, which can have many of the same effects as vertical integration. Exclusive dealing can foreclose a market to competitors by limiting a firm’s demand to one supplier, while tying achieves foreclosure by a firm using its market power in one market to require consumers to buy another product from it. Recent theories have characterized circumstances under which these vertical restraints are profitable and welfare-reducing. In spite of the anticompetitive effects of exclusive dealing and tying not being universal, antitrust policy has treated them harshly. Other vertical restraints examined included resale price maintenance and territorial restraints, which are of particular relevance to relations between manufacturers and retailers.

Questions and Problems

1. Consider a business firm that has an organization that you understand well. Try to explain why the firm buys certain inputs and makes the remainder (in the context of minimizing transaction costs).

2. Assume that firm $M$ (the manufacturer) sells an input (a lawn mower) to firm $R$ (the retailer). Now $R$ sells the lawn mower to the public, incurring a constant cost of $5 per lawn mower for its services. Fixed-proportions production holds for $R$. Let $X$ therefore represent the number of lawn mowers. If both $M$ and $R$ are monopolists, and $P_L$ is the lawn mower price charged to the public with the demand $X = 100 - P_L$, answer the following questions.
   a. Find the derived demand for lawn mowers facing $M$. Hint: Find the marginal revenue equal marginal cost condition for $R$, where $R$’s marginal cost is the sum of the $5 services cost and the price $P_s$ that it must pay $M$ per lawn mower. Solving for $X$ gives the derived demand.
   b. If $M$’s total cost function is $10 + 5X + X^2$, find the equilibrium prices and quantity, $P_L$, $P_s$, and $X$, and the profits of the two firms.
   c. Assume now that $M$ and $R$ form a single vertically integrated firm, $M$-$R$. Find the equilibrium values of $P_L$ and $X$ and the profit of $M$-$R$.
   d. Compare the unintegrated case in part b with the integrated case in part c. Is it true that both the firms and the public would prefer the case in part c? Explain.

3. Assume the same facts as in problem 2, except that monopolist $R$ is now replaced with competitive industry $R$.
   a. Find the derived demand for lawn mowers facing the manufacturing monopolist $M$. Hint: Make use of the fact that perfect competition equilibrium is defined by demand equals supply, where supply is simply a horizontal line, $P_L = 5 + P_s$. Solving for $X$ gives the derived demand.
   b. Find the equilibrium prices and quantity, $P_L$, $P_s$, and $X$, and the profit of $M$. 
c. Assume that $M$ vertically integrates forward into the competitive industry $R$, thereby extending its monopoly to cover both manufacturing and retailing. Find the equilibrium values of $P_L$ and $X$ and the profit of the combined firm $M-R$.

d. Compare the unintegrated case in part b with the integrated case in part c. Is it profitable to monopolize forward? What is the intuitive explanation for your result?

4. Assume a situation where a monopolist of input $M$ sells to a competitive industry $Z$ and the competitive industry $Z$ has a production function characterized by variable proportions. A second competitive industry sells its output $L$ to the competitive industry $Z$, and $Z$ combines $M$ and $L$ according to the production function $Z = L^{0.5}M^{0.5}$.

The price of $L$ and its marginal cost are both $\$1$. The demand for the product of industry $Z$ is $Z = 20 - P_Z$. It can be shown that the monopolist will charge $\$26.90$ for $M$ to maximize its profit, given that its marginal cost of $M$ is $\$1$. (This can be found by first obtaining the derived demand facing the monopolist using the price equal marginal cost condition in industry $Z$, and also using the condition for least cost production by industry $Z$.)

The competitive industry $Z$ will have a constant marginal cost of $\$10.37$ and sell 9.63 units at a price of $\$10.37$.

a. Calculate the competitive industry $Z$'s actual combination of $L$ and $M$ that it will use to produce the 9.63 units. Find the true economic cost to society of these inputs (not $Z$'s actual payments to its suppliers, in that its payment to the monopolist includes a monopoly margin). Hint: The optimal input mix can be found by the simultaneous solution of two equations: the equality of the marginal product per dollar of the inputs and the production function equated to 9.63 units.

b. Assume that the monopolist decides to vertically integrate forward into the competitive industry $Z$, thereby extending its monopoly to cover industry $Z$. What will be the least-cost combination of $L$ and $M$ and its true economic cost in producing the 9.63 units? Hint: The vertically integrated firm will "charge" itself the marginal cost for $M$ in determining its input mix.

c. What is the cost saving that the vertically integrated monopolist will obtain if it produces 9.63 units? That is, what is the saving compared to the cost found in part a?

d. What makes this vertical integration profitable? Is it in society's interest if the monopolist holds its output fixed at 9.63 units after vertical integration?

e. In fact, after the vertical monopolization of $Z$, the firm $M-Z$ would have a constant marginal cost of $\$2$. Given this fact, what is the profit maximizing price $P_z$ and output $Z$? Draw a figure to illustrate the overall social benefits and costs of this vertical integration.

5. Assume that the maximum values to theater owners for movies A and B are as follows. The Fox Theater values A at $\$100$ and B at $\$70$. The York Theater values A at $\$60$ and B at $\$50$. Is block booking more profitable than charging a single price for each movie? Explain.

6. For the model of tying, assume there are instead two firms that offer product $B_2$ at cost $c_{B_2}$. Derive the conditions for tying to be profitable.

7. Consider a problem faced by Kamera Company. They have developed a patented new camera that can be produced at a constant unit cost of $\$1$. The film $F$ is available competitively at a zero price. Consumers derive utility only from the combined services of the camera and film—which can be measured by packs of film consumed per period. Assuming two consumers with inverse demands: $p_1 = 8 - 4f/3$ and $p_2 = 12 - 3f/2$. 

Vertical Mergers and Vertical Restraints

a. Consumers will purchase only one camera at most; hence, consumer surplus can be viewed as measuring what they would be willing to pay for a camera. If Kamera must charge both customers the same price for a camera, what is the price it will charge, and what is its profit?

b. Assume now that Kamera decides to tie film to its camera. It requires customers to purchase film from Kamera if they wish to buy a camera. Kamera simply buys film on the market at the zero price and resells it. What are the prices of camera and film that Kamera will charge, and what is its profit? Is tying profitable?

c. Compare total economic surplus in case a with that in case b.

d. If customer 2 has a different demand curve, say, \( p_2 = 10 - 5f_2/4 \), it reverses the result in part c. What is the intuitive reason for this reversal?
Monopolization and Price Discrimination

A major policy concern in the United States has long been the so-called dominant firm. Although few true monopolies exist in real-world markets, there are many industries inhabited by a single, dominant firm. Examples include Intel, Eastman Kodak, Boeing, Xerox, Campbell Soup, and Gillette. In years past, many such firms have been involved in monopolization antitrust cases—for instance, Standard Oil, United States Steel, Alcoa, IBM, and Microsoft. At issue is whether they pursued anticompetitive practices in either sustaining or extending their market power.

On the issue of monopolization, recall the wording of Section 2 of the Sherman Act:

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, shall be deemed guilty of a felony.

It is important to note that the law forbids the act of monopolizing, and not monopoly itself. Judge Irving Kaufman explained the reason for this distinction in a 1979 case in which Eastman Kodak was charged with monopolization: “[O]ne must comprehend the fundamental tension—one might almost say the paradox—that is near the heart of Section 2.” Judge Kaufman then quoted a famous passage from an earlier decision involving Alcoa: “The successful competitor, having been urged to compete, must not be turned upon when he wins.” The challenge is to distinguish dominant firm situations built up and maintained by superior efficiency from those achieved and maintained by, for example, predatory tactics.

After reviewing some general issues related to proving a claim of monopolization, an overview is provided on case law, with detailed discussions of the Alcoa and United Shoe Machinery cases. We then move on to some specific forms of monopolization. Predatory pricing occurs when a dominant firms prices so as to induce exit or deter entry. After developing the theory of predatory pricing, antitrust policy is reviewed with particular attention to the new standards provided by the Brooke decision. A second monopolization practice is refusal to deal. As its name suggests, refusal to deal is when a supplier denies supply of its product to another firm. An example is a photocopy manufacturer not selling parts to an independent service provider, thereby preventing the latter from servicing those photocopiers the manufacturer has sold. An in-depth analysis of the recent Kodak case is provided with attention to its implications for antitrust policy in a wide class of markets. Finally, we conclude our discussion of monopolization with the well-known Microsoft case. A final section of the chapter discusses price discrimination and the Robinson-Patman Act.

2. United States v. Aluminum Co. of America, 148 F. 2d 416 (2d Cir. 1945).
Establishing Monopolization Claims

As we will see in discussing the important monopolization cases, bases for dominant firm status vary considerably. The earlier cases, for example, were concerned with monopolies achieved through mergers. Standard Oil was alleged to have become dominant by both mergers and predatory pricing tactics. Alcoa began its domination of aluminum through patents and economies of scale. More recently, Microsoft achieved its dominance through network externalities that arise when the value of a good to a consumer is increasing in the number of consumers who use that good. Just as scale economies allow a larger firm to produce the same product at lower cost, network externalities allow a firm to produce a better product at the same cost.

Monopolization cases are generally evaluated under the rule of reason. Recall from chapter 5 that there are two parts to establishing guilt under the rule of reason: inherent effect and intent. The currently accepted Supreme Court statement on monopolization was given in the 1966 *Grinnell* case:

The offense of monopoly under Section 2 of the Sherman Act has two elements: (1) the possession of monopoly power in the relevant market, and (2) the willful acquisition or maintenance of that power as distinguished from growth or development as a consequence of a superior product, business acumen, or historic accident.

In this section, various methods for determining whether a firm is dominant are described. Also provided are some preliminary remarks on assessing whether a dominant firm’s behavior is intended to monopolize, although how this is actually done varies considerably from case to case.

Measuring Monopoly Power

The usual graphical depiction of a monopoly is shown in figure 9.1. The monopolist is the sole supplier of the output \( Q \), and therefore chooses the profit-maximizing output \( Q^* \) (where marginal revenue \( MR \) equals marginal cost \( MC \)). Economic profits are shown by the shaded area. This type of analysis contains two strong implicit assumptions: (1) the product is homogeneous, and therefore the market is a “given,” and (2) entry is ignored (or blockaded). Hence, this “monopolist” is easily identified.

A simple definition of monopoly power is the ability to set price above marginal cost. Hence, if we divide the price–marginal cost difference by price, we have one well-known

index of the amount of monopoly power. It is known as the Lerner index. This index, \( L \), is defined as

\[
L = \frac{P - MC}{P} = \frac{1}{\eta}
\]  

(9.1)

where \( P \) is price and \( MC \) is marginal cost. It is important to note that all values are measured at the firm’s profit-maximizing quantity.

Under the assumption of profit maximization, the Lerner index equals the reciprocal of the absolute value of the elasticity of demand, which is denoted \( \eta \). For example, if the firm’s price is double the marginal cost and \( L = 0.5 \), we can infer that the elasticity of demand is 2. Furthermore, it follows that very large elasticities imply very little monopoly power. Recall that competitive firms face infinitely large elasticities and therefore have Lerner indexes of zero.


5. This result was developed in chapter 5, equation 5.9, for the case of \( n \) firms. By setting \( n = 1 \), the monopoly result here is obtained.
The short-run monopoly power measured by \( L \) should not be relevant in antitrust unless it is “large” and is expected to persist for a reasonably long period of time.\(^6\) Hence, barriers, or obstacles, to entry should exist for there to be a serious antitrust concern.

An alternative index of monopoly power is the Bain index, \( B \).\(^7\) The index is essentially a profit index and is defined as

\[
B = R - C - D - iV
\]  
(9.2)

where \( R \) is total annual sales revenue, \( C \) is currently incurred costs for materials, wages, and salaries, \( D \) is depreciation of capital investment, \( i \) is the interest rate for capital funds, and \( V \) is owners’ investment. Hence the Bain index measures economic profits, in that it subtracts all costs from revenues, including the opportunity cost of the owners’ investment, \( iV \). As Bain observed, “Although excess profits are not a sure indication of monopoly, they are, if persistent, a probable indication.”\(^8\)

Estimating the Bain index using accounting data has many well-known difficulties.\(^9\) Accountants do not capitalize certain investments (advertising and research and development) that should be capitalized for economic analysis. Assets that have been sold may also include in their accounting value the present value of monopoly profits, and this makes it impossible to detect economic profits. Excess profits might be attributable to rapidly growing demand or superior efficiency, and not monopoly power. Many other problems could be cited.

One useful lesson from this short review of the economic definition of monopoly power is that it is not an “either-or” concept. It is a matter of degree. As Richard Schmalensee put it in a law review article,

Both short-run and long-run monopoly power are logically continuous variables, in the sense that they can take on a whole range of values. The questions about monopoly power that usually interest economists involve its sources and importance, rather than its existence. Courts, on the other hand, often seem to treat the existence and importance of monopoly power as though they were equivalent.\(^10\)

---

\(^6\) It is useful to observe that \( L \) might be large but might involve very little economic activity. A better measure of monopoly power for purposes of deciding whether the government should bring antitrust charges is probably the deadweight loss that it causes. It can be shown, for example, that for a monopolist with constant marginal cost and linear demand, the deadweight loss, \( DW = L \) (Monopolist’s revenue)/2. Hence it is clear that the deadweight loss would be small if the revenue involved is small, regardless of \( L \). See Richard Schmalensee, “Another Look at Market Power,” *Harvard Law Review* 95 (June 1982): 1789–816.


\(^8\) Bain, “Profit Rate,” p. 274.


In the real world, we do not see monopolies (as in figure 9.1) with homogeneous products and with blockaded entry. Rather, we see firms with large shares of “markets” that contain various products that are imperfect substitutes. In fact, disputes over where to draw the boundaries in defining the “markets” occupies much time and effort in monopolization cases.

In chapter 7 the definition of the relevant market in merger cases was examined in depth. The concept explained there was that a market should include all firms and products that a hypothetical cartel would need to control in order to raise the existing price in a permanent way. Of course, the same general principle is applicable in monopoly situations. A good example is a famous market definition problem involving flexible wrapping materials, to which we now turn.

In the 1956 Cellophane case, the issue turned on whether the relevant market was cellophane only, or whether it was much broader (including other types of flexible wrapping materials: wax paper, greaseproof paper, glassine, foil, and others). If cellophane alone were the market, duPont would have had a 75 percent market share, and the Court would have found monopoly power. Fortunately for duPont, the Court held that the relevant market was the broader flexible wrapping materials market (of which duPont’s share was only 18 percent).

An important point can be illustrated by an economic error made by the Court in the Cellophane case. A high degree of substitution by consumers between two products must exist at competitive prices for the two products to be placed in the same market. Cellophane was indeed a close substitute for other wrapping materials at the going price for cellophane. However, Cellophane’s price contained a monopolistic margin over its marginal cost. A rational monopolist would, in fact, raise price until its product became a substitute for alternatives. Hence, substitutes in consumption should be evaluated at prices that are reasonably close to marginal costs.

Summing up, one should begin with the product at issue—cellophane—and ask what other products, if any, duPont would need to control in order to charge a price, say, 5 percent, above its marginal cost. The evidence that we have cited is that no other products were necessary. Hence the market should be defined as cellophane only.

In addition to the demand-side substitution emphasized in the previous case, supply-side substitution is equally important in defining the relative market. Three types of substitution are possible: (1) competitors currently producing the product may have the ability to increase output from existing facilities; (2) producers of products not considered substitutes in consumption may be able to easily convert to production of relevant products (example: consumers cannot substitute between residential buildings and commercial buildings, but firms

constructing commercial buildings could easily shift to home building); and (3) entry of new competition.

Franklin Fisher has argued persuasively that the tendency of antitrust cases to be focused on determining the relevant market is often seriously misleading. In his view, “monopoly power is present when a firm is sufficiently insulated from competitive pressures to be able to raise prices... without concern for its competitors’ actions because its rivals cannot offer customers reasonable alternatives.”

His complaint is that the courts often focus on defining the markets—which competitors to include and which to exclude—and then on determining whether the firm under investigation has a high market share. But this transforms a continuous question—how close substitutes are two products—to one of inclusion and exclusion from the analysis. The attention paid to market share as the determinant of monopoly power is due in part to the famous opinion of Judge Hand in the 1945 Alcoa case, which we will discuss later in the chapter.

Assessing Intent to Monopolize

Given the existence of monopoly power, the second part of the rule of reason test is to determine whether the monopoly was acquired and/or maintained by practices that cannot qualify as superior efficiency or historic accident. That is, a monopoly over widgets because of superior efficiency in producing widgets is not in violation of the Sherman Act. A widget producer who used predatory pricing to bankrupt all of its rivals in acquiring the monopoly would be in violation.

Distinguishing predatory pricing from vigorous price competition is generally difficult. For example, IBM developed a computer system in the 1960s that consisted of the central processing unit, input-output devices, memory devices, and so on. A number of companies found it profitable to begin selling memory devices that could be “plugged” into the IBM system. When IBM recognized that it was losing sales to these memory suppliers, it responded by vigorously cutting its own prices. These smaller companies suffered losses, and some were forced out of the market. Was IBM maintaining monopoly power by predatory pricing, or was it simply being forced by competition to lower its prices (to the benefit of its customers)?

Distinguishing the two polar cases—superior efficiency and predatory pricing—in the earlier widget examples was clear. However, what about practices that are “in between”? In the United Shoe Machinery case, Judge Wyzanski found that the use of certain leasing practices satisfied the second part of the test for illegal monopolization. The judge viewed leasing

provisions as constituting an “intermediate case where the cause of an enterprise’s success . . . [was not] the skill with which the business was conducted, but rather some practice which without being predatory, abusive, or coercive was in economic effect exclusionary.” Needless to say, Judge Wyzanski’s opinion has been controversial. Some critics have argued that no such category of intermediate practices even exists. We shall return to these issues in our review of some important monopolization cases.

Development of Antitrust Case Law

One antitrust expert has suggested that there have been three more or less distinct eras of Section 2 interpretation. The first period, 1890–1940, was one in which the courts required, in addition to a large market share, evidence of abusive or predatory acts to show intent. In the second period, 1945–1970, the courts did not require evidence of abusive acts to infer intent; it was sufficient “to achieve a monopoly by maneuvers which, though ‘honestly industrial,’ were not economically inevitable.” The third period, 1970 to the present, appears to be characterized by the willingness of the courts to allow more aggressive practices by dominant firms without inferring intent to monopolize.

This section provides a brief overview of the development of antitrust case law with respect to monopolization practices. It is followed by an analysis of predatory pricing and refusal to deal, including discussions of some of the most significant cases in recent years.

1890–1940: Standard Oil and United States Steel

In 1911 the Supreme Court handed down two significant decisions. In the first, the Standard Oil Company, organized by the Rockefeller brothers, was found guilty of monopolization and dissolved into thirty-three geographically separated companies. Two weeks later, James B. Duke’s Tobacco Trust was also found guilty of monopolization and was divided into sixteen successor companies. Although both monopolies were accused of numerous predatory and abusive tactics, we will focus on Standard Oil because it has become a famous example of predatory pricing.

The Rockefellers built Standard Oil by acquiring more than 120 rivals. They also were accused of engaging in predatory pricing to drive competitors out of business, of buying up pipelines in order to foreclose crude oil supplies to rivals, of securing discriminatory rail

freight rates, and of conducting business espionage. Standard Oil obtained a 90 percent share of the refining and sale of petroleum products in the 1880s and 1890s.

The Supreme Court stated in its opinion that the crime of monopolization required two elements. First, the firm must have acquired a monopoly position. The 90 percent market share met this requirement. Second, there must be evidence of intent to acquire the monopoly position. The Court found that intent could be inferred from the predatory tactics described. An in-depth analysis of predatory pricing is provided later.

Turning to United States Steel, the government charged it with monopolization in 1911.20 The company was formed in 1901 through mergers that gave United States Steel control of over 65 percent of the domestic iron and steel business. In 1907 the chairman of United States Steel, Judge E. H. Gary, began a series of dinner meetings with the heads of rival firms. These so-called Gary Dinners were designed to help stabilize pricing and create goodwill among the industry leaders. Apparently they accomplished this mission, because during the trial no competitors had any harsh words for United States Steel’s conduct. Unlike the predatory and abusive tactics that existed in the Standard Oil and American Tobacco cases, United States Steel was viewed as a “good citizen” by its rivals. One result of its price leadership was a gradual loss of market share that reached 52 percent by 1915. United States Steel seemed to hold a price “umbrella” up for its smaller rivals, allowing these rivals to offer lower prices, thereby increasing their share of the business.

The Supreme Court decided in its 1920 opinion that United States Steel was not guilty. The Court concluded that even if the company did have monopoly power, it had not exercised that power: “the law does not make mere size an offense or the existence of unexerted power an offense.” Hence the law seemed to be clear in this period: dominant firms would violate the Sherman Act’s Section 2 only if they engaged in predatory or aggressive acts toward rivals.

1940–1970: Alcoa and United Shoe Machinery

Twenty-five years after the United States Steel decision, another significant monopolization case was decided.21 The Aluminum Company of America (Alcoa) was the sole American producer of primary aluminum until 1940. In 1945, Circuit Judge Learned Hand set forth the opinion that Alcoa was guilty of illegal monopolization even though it had engaged in none of the aggressive and predatory tactics that characterized earlier convictions. Although the decision was not a Supreme Court ruling, it had the effect of one because the Circuit Court of Appeals was empowered by Congress to act as the court of last resort. A quorum of the Supreme Court was not available because several justices disqualified themselves because of a previous connection with the litigation.

Alcoa was the holder of the patents of Charles Hall, who in 1886 invented a commercially feasible electrolytic process for converting alumina (concentrated aluminum oxide) into aluminum. Thanks to the Hall patent and other patents, Alcoa had protection from competition until 1909. After 1909, Alcoa was protected from foreign competition by high tariffs on imported aluminum. In addition, Alcoa protected its monopoly by buying up many of the low-cost deposits of bauxite (the ore containing aluminum) and cheap electric power sites (aluminum production is electricity-intensive). Many economists also believe that Alcoa made entry less attractive by limit pricing—unlike United States Steel’s strategy, discussed earlier.

Alcoa’s price and output performance in its years as a monopoly were impressive. Aluminum prices generally declined over this period and output grew rapidly. One Alcoa official explained that their profits “were consistently held down in order to broaden markets. Aluminum has no markets that have not been wrested from some other metal or material.”

The aluminum industry can be divided into the four vertical stages, as shown in figure 9.2. Each stage involves a distinct technology and can be located in different regions. For example, bauxite is mined largely in the Caribbean area, but bauxite is processed into alumina near the Gulf Coast ports. Large-scale economies exist in the alumina plants (stage 2 in figure 9.2)—in fact, until 1938 there was only one alumina plant in the United States. Because the reduction of alumina into aluminum ingots requires large amounts of electricity, these plants are

located near cheap hydroelectric power sites (in the Northwest and in the Tennessee Valley). The first three stages in figure 9.4 constitute the production of primary aluminum.

The fourth stage, fabrication, is technically similar to fabrication of other metals. Hence there are independent fabricators who buy aluminum ingot from primary producers and compete with the fabricated output of the primary producers. Again, until the 1940s Alcoa was the only primary producer in the United States.

It is important to understand the vertical structure of the industry in order to evaluate the market definition selected by Judge Hand. Also, one further technical fact is needed. As a durable good, scrap aluminum can be reprocessed and used by fabricators as a substitute for primary aluminum. This so-called secondary aluminum output was approximately one-quarter of the primary aluminum output in the 1940s.

Judge Hand considered three market share definitions for Alcoa:

1. \[
\frac{\text{Alcoa's sales}}{\text{Total primary + Secondary + Imports}} = 33 \text{ percent}
\]

2. \[
\frac{\text{Alcoa's sales + Internal use}}{\text{Total primary + Secondary + Imports}} = 64 \text{ percent}
\]

3. \[
\frac{\text{Alcoa's sales + Internal use}}{\text{Total primary + Imports}} = 90 \text{ percent}
\]

In the first definition, Alcoa’s consumption of its own primary aluminum production for fabrication purposes is excluded. By adding it back, the second definition is attained, leading to an increase in market share from 33 percent to 64 percent. The third definition yields a share of 90 percent by excluding secondary aluminum from the denominator.

Judge Hand stated that 90 percent “is enough to constitute a monopoly; it is doubtful whether sixty or sixty-four percent would be enough; and certainly thirty-three percent is not.” He argued that 90 percent is the correct share for the following reasons. First, “All ingot—with trifling exceptions—is used to fabricate intermediate, or end, products; and therefore all intermediate, or end, products which Alcoa fabricates and sells, pro tanto reduce the demand for ingot itself.” Hence, “Internal use” should appear in the numerator. Second, Alcoa in the past had control of the primary aluminum that reappears as secondary aluminum in the present. Hence, Alcoa “always knew that the future supply of ingot would be made up in part of what it produced at the time, and if it was as far-sighted as it proclaims itself, that consideration must have had its share in determining how much to produce.” For this reason, Judge Hand excluded secondary from the denominator and concluded Alcoa’s share was 90 percent.

However, according to Judge Hand, “it does not follow that because Alcoa had such a monopoly, that it had ‘monopolized’ the ingot market: it may not have achieved monopoly; monopoly may have been thrust upon it.” And, “a single producer may be the survivor out
of a group of active competitors, merely by virtue of his superior skill, foresight and industry.” Judge Hand ruled out these possibilities in the Alcoa case. As he put it,

The only question is whether it falls within the exception established in favor of those who do not seek, but cannot avoid, the control of a market. It seems to us that that question scarcely survives its statement. It was not inevitable that it should always anticipate increases in the demand for ingot and be prepared to supply them. Nothing compelled it to keep doubling and redoubling its capacity before others entered the field.

The Alcoa decision was then a major change in the legal definition of monopolization. Predatory and aggressive acts were no longer necessary. Simply building capacity ahead of demand could be sufficient to indicate intent to monopolize by a dominant firm.23 The remedy in the Alcoa case was to create two new competitors (Reynolds Metals and Kaiser Aluminum) by the sale of government-owned plants. The plants were built at government expense during World War II for military purposes. Divestiture of Alcoa was not a feasible alternative in any case since Alcoa had only two alumina plants, and one was almost obsolete.

The 1953 United Shoe Machinery case24 was important in providing another example of a business practice that could indicate illegal monopolization by a dominant firm. The leasing practices of United Shoe were found to be exclusionary, and therefore evidence of illegal monopolization.

United Shoe supplied between 75 and 85 percent of the shoe machinery in the United States, and there was no question of its dominance. United would not sell its machines to shoe manufacturers, but leased them for ten-year terms. There was a requirement that lessees had to use United machines if work was available. Also, the practice of United was to provide free repair services, and the Court viewed this as restricting entry, since rivals of United would have to offer repair services as well. That is, independent repair firms would not exist, given United’s zero charges. Having to offer repair services in addition to shoe machinery would raise a capital requirement barrier.

The remedy in the United Shoe case was to purge the leases of their restrictive practices. Though divestiture into three separate manufacturing plants was proposed by the government, the Court held this to be infeasible, because all of United’s manufacturing was conducted in a single plant in Massachusetts.

1970 to Present: Kodak, IBM, Microsoft, and Others

The last several decades witnessed a number of “big cases” brought by antitrust agencies. IBM, Xerox, AT&T, Microsoft, Intel, Kodak, and three breakfast cereal manufacturers were charged with monopolization. There were also a number of private suits that have received

23. There is a discussion in chapter 5 on strategic investment in capacity so as to deter entry.
much attention, including a number of cases by smaller computer firms against IBM, and Berkey Photo against Eastman Kodak.

The Berkey–Kodak case involved a photofinisher (Berkey) charging monopolization against Eastman Kodak—with its 60 to 90 percent shares of most segments of the photography industry. Berkey claimed that when in 1972 Kodak introduced its 110 Pocket Instamatic photographic system, which required a new Kodacolor II film, rival film and photofinishing suppliers were foreclosed from that market. According to Berkey, Kodak should have predisclosed its innovation to rivals so that they could compete immediately on introduction. Judge Kaufman concluded that Kodak “did not have a duty to predisclose information about the 110 system.” He went on to explain his reasoning:

It is the possibility of success in the marketplace, attributable to superior performance, that provides the incentives on which the proper functioning of our competitive economy rests. If a firm that has engaged in the risks and expenses of research and development were required in all circumstances to share with its rivals the benefits of those endeavors, this incentive would very likely be vitiated.

Some of the other “big cases” mentioned at the beginning of this section were terminated by consent decrees (Xerox, AT&T, and Intel) and by a decision by an FTC administrative law judge (breakfast cereals case). A consent decree is a negotiated settlement between the two parties subject to court approval. In the Xerox settlement in 1975 it was agreed that Xerox would license patents, supply “know-how” to competitors, sell as well as lease copy machines, and alter its pricing policies. In the AT&T settlement in 1982, AT&T divested its telephone-operating companies. Among other things, this separated the regulated telephone utilities from their main unregulated equipment supplier, Western Electric.

The “shared monopoly” theory presented by the FTC against Kellogg, General Mills, and General Foods was a novel approach to a highly concentrated oligopoly. The three firms had collectively 81 percent of the ready-to-eat cereals market, with Kellogg, the largest of the three, holding 45 percent. The FTC charged that the companies had engaged “in certain interdependent acts and practices in order to achieve a highly concentrated, noncompetitive market structure and shared monopoly power.”

An example of one of these acts was termed “brand proliferation.” By introducing some 150 brands between 1950 and 1970, the companies were alleged to have left “no room” for new entrants. The economic theory follows closely the Bain-Sylos limit-pricing model. The key idea is that brands are located in a “product characteristics space,” and each competes only with brands located nearby exhibiting similar product characteristics. Fixed launching costs make it necessary to achieve a relatively large sales volume to be economically viable.

26. This important divestiture is described in detail in chapter 15 on telecommunications.
So, if the “room” between brands is small (because of brand proliferation), new entrants will be deterred because they cannot expect to attain large enough sales.28

The FTC judge, however, saw brand proliferation as “nothing more than the introduction of new brands, which is a legitimate means of competition. . . . There is no evidence of a conspiracy or intent to deter entry by means of new product introductions.” After considering other acts, and rejecting them as being illegal, the judge dismissed the complaint. In 1982 the FTC commissioners decided to let the dismissal stand rather than appeal the decision to a higher court.

The Department of Justice (DOJ) filed its case against IBM in 1969, following a lengthy internal investigation of the leading computer firm. The case, dismissed in 1982, involved huge resource costs—over $200 million in legal costs, 950 witnesses, 726 trial days, 17,000 exhibits, and 104,400 trial transcript pages.29 The government argued that IBM had about 70 percent of the market for medium-size, business-oriented, “general purpose,” electronic, digital computer systems. Not unexpectedly, IBM argued that the market was much broader and should include companies selling “parts of systems” and computer systems for scientific and military purposes. IBM’s share under the broader definition was less than 50 percent.

IBM’s dominant position in the computer industry at the time is usually attributed to its strong position before computers existed in the punched-card office equipment business. This gave IBM strong ties to the potential computer users. IBM delivered its first computer, the IBM 650, in 1954. Thus, while not being the technical leader in computers, IBM supplied both hardware and software that performed well.

According to the government, IBM engaged in numerous practices that enabled it to maintain its monopoly power. These practices were argued to be neither “inevitable nor honestly industrial.” They were leasing, bundling, differentiation of software, fighting machines, tying of products, manipulation of purchase-to-lease price ratios, and education allowances.

In the last decade, there have been major cases against Microsoft and yet another one against Kodak. Though the Microsoft case received more media attention, the decision against Kodak is likely to be more important. Both cases will be discussed in detail later.

**Predatory Pricing**

In 1994, Frontier Airlines entered the Billings–Denver route with a fare of around $100, about half of the fare offered by the incumbent firm, United Airlines. United responded with a comparable fare. After about a year, Frontier withdrew from the route, in response to which

28. This type of entry barrier is discussed in more detail in chapter 6.
United’s fare rose precipitously. (To see the path of price, go to figure 17.10.) Frontier Airlines then logged a complaint with the Department of Justice (DOJ) that United had acted in an anticompetitive manner by engaging in predatory pricing.  

What exactly is anticompetitive about what just happened? Might one say that this is just “normal” competition at work? To get a better handle on these questions, let us examine matters a bit more closely. Start with a monopolist for which the firm (and market) demand curve is $D(p)$ (see figure 9.3(a)). Initially, it is pricing at the profit-maximizing level of $p^m$, where its marginal revenue curve, denoted $MR$, is equated with its marginal cost, denoted $c_I$. Now suppose entry occurs by firm $E$, which has marginal cost $c_E$. The entrant prices at $\hat{p}_E$, so that the incumbent firm’s demand is now $D(p_i; p_E = \hat{p}_E)$, which is less than its demand before entry. The competitor takes part of the incumbent firm’s demand and, in fact, takes more when it prices lower (this is why the incumbent firm’s demand is shown to depend on the entrant’s price.) Given that demand curve, the new marginal revenue curve is $MR'$, for which $\hat{p}_i$ is the profit-maximizing price. An analogous situation prevails for the entrant and is depicted in figure 9.3(b). Note that $\hat{p}_E$ is optimal for the entrant given it faces demand curve $D_E(p_E; p_i = \hat{p}_i)$. 

Instructor Evaluation Copy
Not for Classroom Use

307 Monopolization and Price Discrimination

$D_E(p_E; p_I = \hat{p}_I)$. This is then an equilibrium in that each firm’s price maximizes its profit given its rival’s price. As we would expect, entry causes prices to fall—from $p^m$ to $\hat{p}_I$—as a part of normal profit maximization and competition. The entrant earns positive profit, measured by the shaded rectangle.

That describes the type of competition we would like to see. Now consider the incumbent firm pursuing a more aggressive pricing policy with the intent of reducing the entrant’s profit. This situation is depicted in figure 9.4. The incumbent firm now prices at $\tilde{p}_I$, which is below $\hat{p}_I$. Relative to when its competitor’s price is $\hat{p}_I$, the entrant’s demand curve is shifted in to $D_E(p_E; p_I = \tilde{p}_I)$ and the price at which it maximizes profit, $\bar{p}_E$, is lower than $\hat{p}_E$. It now incurs losses as price is less than average cost. Further note that the incumbent firm’s profit-maximizing price, denoted $\tilde{p}_I$ in figure 9.4(a), exceeds $\tilde{p}_I$, which is the price that maximizes its current profit. This scenario may describe an episode of predation. The incumbent firm prices below that which maximizes current profit. The potential benefit is in terms of higher future profit by inducing exit by the new rival or deterring future entry in other markets. 31

In sum, when a firm enters a market, it does so at a price below the preexisting price so as to lure business away from existing suppliers. In response, we expect incumbent firms to

31. Also see the discussion in chapter 17 on predatory pricing in the airline industry.
also lower their prices. This is defined to be *predatory pricing* only when the intent is exclusionary; that is, it is profitable only because it will result in a higher future price in this or some other markets by virtue of inducing exit or deterring entry in those other markets.

Predatory pricing has also been defined as a price designed to induce an equal or more efficient firm to exit. Unfortunately, this definition is flawed. As we will see, a price can be exclusionary even when it does not induce exit. By pricing aggressively and making entry sufficiently unprofitable ex post, a firm may be able to discourage future entry. (The reason that a firm may not exit but regret having entered is that there are some sunk, unrecoverable costs associated with entry.) Furthermore, entry by even a less efficient firm can be welfare-improving. Though it is producing at higher cost, this may be compensated by the lower price it induces through intensified competition. So it is not clear that we would want to allow strategic behavior that drives out a mildly less efficient competitor.

The relevance of predatory pricing does not rest on whether an incumbent firm is able to impose losses on a rival. The real issues are instead the following. First, when is it that an incumbent firm’s current price can influence entry and exit? Here it is important to recognize that entry and exit are based on *future* anticipated profit and not directly on what a firm is currently earning. Even if a firm is incurring losses, it can borrow against future earnings if the capital market thinks it will have future earnings. Thus, the issue comes down to the beliefs of firms and the capital market as to the future profit stream. When can an incumbent firm’s pricing behavior substantively influence those beliefs? Second, presuming that predatory pricing can “work,” when is it optimal for an incumbent firm to use it? When does predation generate a higher profit stream than accommodating rivals or acquiring them?

At the heart of the issue of the optimality of predatory pricing is a simple intertemporal trade-off. A necessary condition for predation is that a firm is pricing differently from that which maximizes current profit in order to reduce future competition and thereby yield higher future profit. This trade-off is depicted in Figure 9.5. In response to entry, an incumbent firm can accommodate by pricing in a manner to maximize current profit and accepting the existence of a competitor. Alternatively, it can price aggressively—which reduces profits in the near term—and be rewarded with higher future profits upon exit. Predation is preferred over accommodation when the value of the near-term forgone profit is smaller than the value of the long-term profit gain.

Associated with any predatory pricing scheme is a demonstrative market and a recoupment market. The *demonstrative market* is the one in which aggressive pricing takes place. That is where an incumbent firm incurs a cost through forgone profit. The return to that strategy is reaped in the *recoupment market*. As depicted in figure 9.5, the demonstrative market is the incumbent’s firm market in the periods prior to exit and the recoupment market is that

---

32. This is shown more fully in chapter 16.
same market after exit. Or one could think of this in a multimarket context. A retail chain can price predatorily in Cleveland—the demonstrative market—with the anticipation of such behavior deterring entry in Cincinnati, Detroit, and Indianapolis—the recoupment markets.

Before moving on, the reader should be warned that predatory pricing is the “black hole” of antitrust. As with black holes, one cannot directly observe predatory pricing in that what defines it is the intent to exclude. The intent of a price cut is not observed; at best it is inferred. And, as with black holes, the evidence of their existence is elusive and contentious. Some economists believe predatory pricing is exceedingly rare if not nonexistent. Others believe there are well-documented episodes. The debate over the importance of predatory pricing has raged for more than a century and does not appear to be abating.

**Theories of Predatory Pricing**

The theory of predatory pricing has been slow in developing. Though predatory pricing has been an antitrust issue since the turn of the twentieth century, it came under careful

---

theoretical scrutiny only in the late 1950s with a reexamination of the *Standard Oil* case.\(^{34}\) That study came to the conclusion that Standard Oil had not priced predatorily after all. The reason was that such a strategy would not have been profitable. This famous study led to rising skepticism as to the relevance of predatory pricing as a phenomenon. But in the last two decades, game theory has been used to show that predatory pricing can be quite rational. These theories are reviewed here after we discuss the Standard Oil study that launched the theoretical examination of predatory pricing. The section concludes with a discussion of various pricing theories that produce paths that look predatory but are actually not. In considering any candidate predatory price path, it is important to consider these alternative efficiency rationales for low pricing.

**Standard Oil and the Claimed Irrationality of Predatory Pricing**

In reexamining the trial record of the 1911 Standard Oil case, John McGee came to the conclusion that predatory pricing had not occurred.\(^{35}\) In substantiating this claim, he provided a number of challenges for which any theory of predation must overcome.

In the Standard Oil case and in most litigated cases of predatory pricing, the predator firm has much greater sales than the prey firm. If firms are pricing below cost, the cost of predation is then actually more severe for the firm inflicting it. If so, shouldn’t the prey outlast the predator in this price war? But then how is predation effective? One common response is that the predatory has “deep pockets” to finance these losses. But Professor McGee’s reply is that if the prey is more efficient, the capital markets can provide the necessary financing, so it can withstand the price war. A theory of predation would have to explain why that does not occur.

Let us even suppose the predator succeeds in driving out its competitor, after which it raises price to reap the rewards. Why wouldn’t another firm enter in response to that high price? Isn’t then success at best brief? Finally, even if such a pricing strategy did succeed in causing exit and something prevented other firms from subsequently entering, predation is a very costly way in which to enhance one’s market power. Wouldn’t it be more profitable for the predator to buy out the prey? Indeed, this would seem to be a Pareto improvement, as both firms avoid losses from doing so.

While the last criticism can be dismissed these days by the retort that the antitrust authorities would not allow a dominant firm to acquire or merge with a competitor, the other criticisms of the predatory story are not so easily dismissed. This will take some care, and to that task we turn next.

---

35. Ibid.
Overview of the Modern Theory of Predatory Pricing

The McGee analysis erected a formidable challenge to those who believe predatory pricing is a cogent theory. They had to show that predatory pricing could work and that it was the most profitable response by an incumbent firm to entry. Few challengers initially came forward, but, beginning in the 1980s, with the aid of game theory, the sword was eventually pulled from the stone. We now know that predatory pricing is a logically valid means toward maintaining dominance in an industry. To what extent it actually occurs is much more of an open question.

There are at least three classes of predatory pricing theories. The theory of financial market predation explains why imperfect capital markets may not provide adequate financing for a more efficient firm to survive predation. The argument is as follows. Suppose the capital market is unable to perfectly monitor a firm’s management. There is then an inference problem when losses are observed. It may be due to factors outside of the management’s control, such as a weak economy or predation, but may also be due to incompetent or negligent management (for example, one that makes bad project choices or doesn’t work hard). If the capital markets automatically financed any losses, this would provide weak incentives to management to do its job well. So the capital market does not do that, but this then provides the opportunity for a predatory pricing strategy to bankrupt its prey. By pricing aggressively and imposing losses, it makes investors inclined to limit financing since it does not know to what extent the losses are due to the competitor’s prices or to mismanagement. This argument is a direct response to the claim that a more efficient firm will have the necessary financial resources to withstand a predatory pricing. It needn’t.

The second class of theories shows how an incumbent firm can develop a useful reputation by aggressively responding to entry. This can enhance the incumbent firm’s profitability by inducing exit or deterring entry in other markets. The basic mechanism is one of signaling. A relevant trait of the market, such as the incumbent firm’s cost, is known to the incumbent firm but not to competitors. The incumbent firm’s price can then influence the profitability of entry and exit because it contains information about this unknown trait. An example of this theory will be examined in detail below.

These reputational models have also responded to the criticism that acquisition of a competitor is more profitable than predation. Theoretical analysis has shown that aggressive behavior can influence the takeover price so that predatory pricing can be used in conjunction with acquisition. Aggressive pricing can signal that the incumbent firm’s cost


is low, and therefore a competitor’s profit in the event that there is no acquisition would be low, which encourages the firm to sell out at a lower price. An examination of American Tobacco found that the price at which it acquired competitors over 1891 to 1906 was indeed significantly reduced because of its reputation for aggressive pricing.\(^{38}\)

The third class of theory shows how predation can reduce the information garnered by a new firm and this can deter expansion; it is referred to as *test market predation*. For example, in 1971 Proctor & Gamble was considering selling its Folger brand of coffee in the eastern part of the United States. As was standard practice for P&G, it introduced the coffee in some test markets in order to gauge demand and, in particular, how it fared in comparison with Maxwell House, the leading brand, sold by General Foods. For this to be an informative experiment, P&G would need to observe demand at the prices that were expected to prevail in the long run. General Foods muddied up the experiment by engaging in an intense price war where price was set below average variable cost. Though P&G learned about demand at very low prices, it did not learn about demand at the relevant prices. By limiting the amount of information that a test market experiment reveals, predatory pricing can make full-scale entry more risky and thus potentially deter it. In this case, P&G chose to postpone its entry plans. It is worth noting that the Federal Trade Commission (FTC) ultimately decided that this was not predatory pricing.

Both the reputation and test market predation theories show that the cost of successful predation need not be high, contrary to the claim of McGee. In those analyses, the objective is not to bankrupt a new firm but rather to signal that future profit is low, which can serve to induce exit or deter entry in other markets of the incumbent firm.

**A Reputational Theory of Predation**

One mechanism by which predatory pricing works is by establishing a reputation for aggressive pricing. Suppose, for example, that a competitor is uncertain of a dominant firm’s cost. It knows, however, that firms with lower cost tend to price lower. Hence, if the dominant firm prices low in response to entry, this may suggest that its cost is low, and thus one can anticipate similarly aggressive behavior in the future. Inducing such beliefs can then prove profitable to the incumbent firm by inducing exit—as a new firm becomes convinced that the incumbent firm will persist with low pricing because it has low cost—or deterring future entry. Of course, its competitors are no fools and will realize the incumbent firm will be trying to induce false beliefs. Regardless, we will show that the incumbent firm can strategically deter entry: that is, by pricing below that which maximizes its current profit, the incumbent firm deters entry that would have occurred had it priced to maximize current profit.

---

Consider an incumbent firm that is currently a monopolist in two identical markets: A and B. An entrant is considering entering these two markets (or, alternatively, there are two entrants and each is considering entry into one of them). Entry requires incurring a fixed (and sunk) cost of $k$. To simplify matters, assume that, in response to entry, a firm can either set a price of $L$ (low) or $M$ (moderate). The profit to an entrant is denoted $\pi_E(p', p'')$ when the incumbent firm prices at $p'$ and the entrant prices at $p''$. For example, an entrant earns $\pi_E(M, L)$ if it prices at $L$ and the incumbent firm prices at $M$.

Assume the entrant’s optimal pricing strategy is to match the price of the incumbent firm:

$$\pi_E(M, M) > \pi_E(M, L) \text{ and } \pi_E(L, L) > \pi_E(L, M).$$

In addition, assume that entry is profitable when the incumbent firm does not set a low price:

$$\pi_E(M, M) > k > \pi_E(L, L).$$

(9.3)

The incumbent firm can be of three possible types. It can have low cost (denoted $l$), moderate cost (denoted $m$), or high cost (denoted $h$). Though the incumbent firm knows its cost, the entrant does not know. When it is low cost, a price of $L$ maximizes the profit in that market, while, if it is moderate or high cost, it is optimal for it to price at $M$. That is, these are the short-run profit-maximizing prices. Let $\pi_i(p', p'')$ denote the incumbent firm’s profit when its cost is $c$, the incumbent firm prices at $p'$, and the entrant prices at $p''$. We have then assumed that

39. In the parlance of game theory, price $L$ strictly dominates $M$ when the incumbent firm is low cost and price $M$ strictly dominates $L$ when the incumbent firm is moderate or high cost.

Finally, let $\hat{\pi}_l$, $\hat{\pi}_m$, and $\hat{\pi}_h$ be monopoly profit for the incumbent when it is low cost, moderate cost, and high cost, respectively.

If there was complete information, so the potential entrant knew the incumbent firm’s cost, entry would occur if the incumbent firm was moderate or high cost but would not occur if it was low cost. We know that pricing low is best for the incumbent firm when it has low cost and, given that the incumbent firm sets a low price, the profit-maximizing response of the new firm is to also price low. This yields profit of $\pi_E(L, L)$ to the new firm which, by equation 9.3, is insufficient to cover its cost of entry. Entry (in either market) is unprofitable. By analogous reasoning, one can show that entry is profitable when the incumbent firm is moderate or high cost, as in both cases it prices at $M$. It follows that, under complete

39. In the parlance of game theory, price $L$ strictly dominates $M$ when the incumbent firm is low cost and price $M$ strictly dominates $L$ when the incumbent firm is moderate or high cost.
information, the incumbent firm earns, in each market, $\hat{\pi}'$ when it is low cost and $\pi_l^c(M,M)$ and $\pi_l^h(M,M)$ when it is moderate cost and high cost, respectively.

Let us return to when there is incomplete information. Suppose that, if entry occurs, it takes place in market $A$ and then, in response to what happens in market $A$, there is possible entry in market $B$. Consider the following strategy for the incumbent firm in response to entry in market $A$:

If the incumbent firm has low or moderate cost, it prices at $L$ in market $A$.
If it is high cost, it prices at $M$ in market $A$. \hfill (9.4)

We will later show that this pricing strategy is optimal for the incumbent firm. In the meantime, we want to derive optimal entry decisions given the incumbent firm uses this strategy.

Suppose all prospective firms know this is the pricing strategy deployed (so that, if there is successful predation, it is not due to entrants being misinformed). Then entry into market $A$ is optimal if

$$(\rho_l + \rho_m)\pi_E(L,L) + (1 - \rho_l - \rho_m)\pi_E(M,M) > k \hfill (9.5)$$

where $\rho_l$ is the probability the entrant assigns to the incumbent firm being low cost and $\rho_m$ is the probability that the incumbent firm is moderate cost.\(^{40}\) For example, if the three cost types are equally likely then $\rho_l = 1/3$ and $\rho_m = 1/3$. Given (9.3), (9.5) holds when the probability that the incumbent firm is high cost is sufficiently high.

Given that market $A$ has been entered, what happens in market $B$ in the event of entry is straightforward. Since there are no future considerations for the incumbent firm, it will choose price to maximize its profit from market $B$, which means pricing low when it is low cost and moderately otherwise.

Let us next consider the decision to enter market $B$, depending on how the incumbent firm priced in response to entry in market $A$. If the incumbent firm prices at $M$, then the potential entrant should infer that its cost is high, since, according to the strategy in (9.4), only a high-cost firm prices at $M$. Entry into market $B$ will then occur since $\pi_E(M,M) - k > 0$. If the incumbent firm instead prices at $L$, then the potential entrant cannot infer its cost, because both a low-cost and a moderate-cost incumbent firm prices aggressively in market $A$. In that situation, let $\beta$ denote the probability that the entrant assigns to the incumbent firm having low cost given it believes the incumbent firm is either low cost or moderate cost.\(^{41}\) Assume that

\(^{40}\) 1 $- \rho_l - \rho_m$ is the probability that it is high cost, since probabilities must sum to one.

\(^{41}\) If the reader is familiar with Bayes’ rule, then $\beta = \rho_l/(\rho_l + \rho_m)$. For example, if all three cost types are initially equally likely, $\rho_l = 1/3$ and $\rho_m = 1/3$, then $\beta = 1/2$. 
so that, in expectation, entry is unprofitable if the potential entrant believes the incumbent firm has low or moderate cost.\textsuperscript{42} Thus, if the incumbent firm prices at $L$ in market $A$, entry into market $B$ will be deterred.

Let us assume (9.5) is true—so that entry occurs in market $A$—and consider the optimal pricing response of the incumbent firm in market $A$. In other words, we want to show that the strategy in (9.4) is optimal. It is clear that pricing at $L$ is optimal when the incumbent firm is low cost. Doing so maximizes profit in market $A$ and, in addition, deters entry into market $B$ (as we argued above). A low price is optimal when it is moderate cost only when

\[
\beta \pi_E(L, L) + (1 - \beta) \pi_E(M, M) < k
\]

(9.6)

Compared to pricing at $M$, a price of $L$ yields lower profit in market $A$ of $\pi_I^m(L, L)$, compared to $\pi_I^m(M, M)$, but generates higher profit in market $B$ of $\hat{\pi}^m$, compared to $\pi_I^m(M, M)$, by deterring entry. Assume equation 9.7 holds. Finally, pricing at $M$ is optimal for a high-cost incumbent firm (as prescribed in (9.4)) when

\[
2\pi_I^h(M, M) > \pi_I^h(L, L) + \hat{\pi}^h.
\]

(9.8)

This is true when a high-cost firm finds it too costly to price low in response to entry. It prefers to price moderately and allow entry to occur into market $B$. Assume (9.8) holds. Both (9.7) and (9.8) can hold when the difference between high cost and moderate cost is sufficiently large that it is too costly for a high cost firm to price aggressively but not too costly for a firm with moderate cost.

Let us compare what happens under incomplete information with what happens under complete information (and thus there is no opportunity for predatory pricing). When the incumbent firm has low or high cost, the outcome is the same under both information scenarios. With incomplete information, a low-cost incumbent firm prices aggressively in response to entry in the first market, and this serves to deter any further entry. This is the same as under complete information; entry ought not to occur when the incumbent firm is highly efficient. With incomplete information, a high-cost incumbent firm prices moderately in response to entry, which induces yet more entry. This is the same as under complete information, and indeed, entry ought to occur when the incumbent firm is highly inefficient. In sum, there is no anticompetitive behavior when the incumbent firm has either low or high cost.

Anticompetitive behavior in the form of predatory pricing does arise, however, when the incumbent firm has moderate cost. Under incomplete information, the incumbent firm prices below that which maximizes the profit it earns in market $A$. This aggressive response to entry

\textsuperscript{42} Note that both equations 9.5 and 9.6 can hold simultaneously if $\rho_l + \rho_m$ is sufficiently small and $\rho_l$ is sufficiently large relative to $\rho_m$. 

is rewarded with the deterrence of entry into its other market. Thus, compared to when there is complete information, price is lower in market $A$—which is good—but entry into market $B$ does not occur—which can be bad.$^{43}$ The demonstration market is then market $A$ and recoupment occurs in market $B$.

In sum, by pricing aggressively in response to entry, an incumbent firm with moderate cost is able to signal that it is not high cost. By leaving sufficient uncertainty about whether it is low or moderate cost, it is able to deter any further entry. Predatory pricing can then be anticompetitive and profitable for an incumbent firm.

**Efficiency Rationales**

If an incumbent firm prices below that which maximizes short-run profit, it does not necessarily mean that its intent is to induce exit or deter entry. There are other ways in which a low price in the short-run can benefit a firm in the long-run. In evaluating the predatory intent of a price path, it is important to consider legitimate business justifications for such pricing.

When a firm gives out free samples, it is clearly pricing below that which maximizes short-run profit; a zero price is certainly below marginal cost! But rarely is such an activity anticompetitive. *Promotional pricing* is just a way in which to expand future demand by informing consumers of the value of one’s product.

In some products, such as aircraft manufacturing, it is well documented that the marginal cost of production declines with experience. This is attributed to *learning-by-doing*, whereby through the actual production of a product, the firm learns about more effective practices that serve to lower cost. In competing with other firms, it may then behoove a firm to produce at a high rate and price low, even below current marginal cost, because it’ll gain experience and benefit by virtue of lower future marginal cost.

Network externalities are a third rationale for a low price in the short-run. *Network externalities* are present when the value of a good to a consumer depends on how many other consumers use that good. For example, the value of a word processing package like Microsoft Word depends on how many people use Word. The more people who use it, the more people with whom one can exchange documents without compatibility problems. Other common examples are communication networks, like telephone and e-mail. The more people connected to the network, the more valuable it is to everyone.

Now consider a market with network externalities in which firms are competing. An example is the video market in the 1980s with the competing formats of VHS and Beta. The more people who have VHS video cassette recorders, the more movies will be available in the VHS format and thus the more valuable is VHS over Beta. It can then be optimal for a firm to price low in selling VHS machines in order to expand its customer base. It will be

---

43. Whether entry deterrence reduces welfare depends on whether entry would have been welfare-enhancing and that depends on whether the rise in consumer surplus from a lower price compensates for the additional entry cost.
rewarded in the future with higher demand (and the ability to charge higher prices), because those additional sales make purchasing a VHS machine more valuable. Like learning by doing, network externalities can lead to low short-run prices as firms are engaging in a long-run competitive battle.

**Antitrust Policy**

The legal definition of predatory pricing is necessarily of great importance. In effect, it will be the rule by which firms must play the “competitive game.” For example, suppose that price cutting is deemed predatory when a price is set below average total cost, as a number of courts have ruled in the past. A dominant firm, faced with entry by an aggressive new rival, must then be careful not to price below its average total cost in responding to that new rival. The result may be that healthy price competition is stifled. Of course, errors can be made in the other direction as well. If the test for predatory pricing is made too permissive, monopoly power might be fostered through predatory pricing. It is then useful to bear in mind that two types of errors are possible in searching for the optimal definition of predatory pricing.

Toward developing a legal definition of predatory pricing, there have been two key events in recent decades. In 1975, two Harvard Law School professors, Phillip Areeda and Donald Turner, published an article in the *Harvard Law Review* in which they proposed a per se definition of predatory pricing. A particularly strong economic argument favoring the Areeda-Turner rule was given by Circuit Judge Stephen Breyer (now on the Supreme Court) in a 1983 decision, *Barry Wright v. ITT Grinnell*. The second event occurred in 1993 with *Brooke Group v. Brown and Williamson Tobacco*. The Supreme Court’s decision laid out a two-tier test for predatory pricing that has proven difficult for plaintiffs to pass.

**The Areeda-Turner Rule and Other Single-Parameter Rules**

In considering the Areeda-Turner rule, it is helpful to refer to the typical short-run cost curves shown in Figure 9.6. Its creators argued that any price below marginal cost ($MC$) will cause the monopolist to lose money on some units of output which is consistent with the predatory pricing strategy. Also, pricing below short-run marginal cost is well known to be economically inefficient (from a short-run perspective). For these reasons, they would classify such prices as predatory and therefore illegal. However, for quantities to the right of $Q^*$, average cost ($AC$) is less than marginal cost. Because prices above average cost (but below marginal cost) would not exclude equally efficient rivals, they would permit such prices (even though they would be economically inefficient).


Figure 9.6
Region Showing Predatory Prices under ATC Rule That Are Not Predatory under Areeda-Turner Rule

Chapter 9

318
Given that estimating marginal cost is difficult, Areeda and Turner propose to substitute average variable cost (AVC) for marginal cost. Their conclusion is as follows:

1. A price at or above reasonably anticipated average variable cost should be conclusively presumed lawful.
2. A price below reasonably anticipated average variable cost should be conclusively presumed unlawful.

The Areeda-Turner rule stimulated a large number of criticisms and alternative proposals. A number of economists independently proposed some version of what we term the average total cost (ATC) rule. It was thought that the Areeda-Turner rule would allow too many predators to escape prosecution. It was claimed that “a monopolist with abundant financial reserves could, under the [Areeda-Turner] rule, drive less financially secure but equally efficient rivals from the market without fear of prosecution merely by pricing below ATC and above AVC.”

The shaded region in figure 9.6 illustrates how the Areeda-Turner rule differs from the ATC rule.

An alternative to cost-based rules was put forth by Oliver Williamson. The idea is that since entry shifts in a firm’s demand, a legitimate competitive response should be to reduce quantity. Thus, an incumbent firm increasing quantity in response to entry is consistent with predation. The Output Restriction Rule states that in the period after entry occurs, the incumbent firm cannot increase quantity above the pre-entry level. One suggestion for the “period after entry” was twelve to eighteen months. Other rules were proposed, though all had serious weaknesses. As described next, predation is far too complex for a simple rule to work.

The Brooke Case and the Two-Tier Rule

The first attempt at a more nuanced approach was proposed by Paul Joskow and Alvin Klevorick. They suggested using a two-stage approach to identifying predatory pricing. The first stage would require an examination of the market structure to determine if the structure is likely to permit predation to be successful. For example, if entry barriers are low, the finding would be that predation is not likely to be a viable strategy, and the case would not be pursued. The second stage would use the type of cost-based or pricing behavior tests described above.

A more sophisticated view of predation also began to develop with the Supreme Court’s decision in *Matsushita v. Zenith* in 1986. The case began in 1970 with charges made by

American television set manufacturers that seven large Japanese firms were conspiring to drive them into bankruptcy. By setting monopoly prices in Japan, the seven firms were argued to use those profits to subsidize “below-cost” U.S. prices. Ultimately, the Japanese were supposed to set monopoly prices in the United States too.

A careful economic analysis showed, however, that the purported strategy is, under reasonable assumptions, unprofitable for the Japanese firms.\(^{49}\) Even with an eventual monopoly that lasts forever, the Japanese firms would fail to break even which raises serious doubts about the credibility of a claim of predation. In coming to that conclusion, the analysis assumed a ten-year predation period with the price of television sets being 62 percent of the price that would have occurred if there had not been predation. With an assumed demand elasticity of about 1.2, the postpredation price ranged from 119 to 138 percent of the “but for predation” price. Finally, a 12.2 percent opportunity cost of capital was assumed, equal to the average return on assets in Japan at that time. Under these assumptions, a lifetime of monopoly profits could not compensate for ten years of predatory profits.

The Supreme Court observed that “it is not enough simply to achieve monopoly power, as monopoly pricing may breed quick entry by new competitors eager to share in the excess profits. The success of any predatory scheme depends on maintaining monopoly power for long enough both to recoup the predator’s losses and to harvest some additional gain.” Then, the Court noted that “if predatory pricing conspiracies are generally unlikely to occur, they are especially so where, as here, the prospects of attaining monopoly power seem slight.”

This provides the backdrop to a landmark Supreme Court decision in 1993. In *Brooke Group v. Brown and Williamson Tobacco*, the Court established a new standard for judging predatory pricing.\(^{50}\) Liggett (owned by Brooke) had introduced a generic, low-price cigarette in a simple black-and-white package. Brown and Williamson, a competitor to Liggett, responded with a similarly packaged generic cigarette and proceeded to continually undercut Liggett’s price to the point that price was below average variable cost. Price remained below cost for eighteen months, after which Liggett raised its price. In evaluating this case, the Supreme Court found that the market structure was such that Brown and Williamson could not recoup its predatory losses. As Brown and Williamson had only a 12 percent share of the cigarette market, the Court argued that recoupment would require coordinated pricing with other companies which was perceived as unlikely.

The *Brooke* decision laid out a clear two-tier policy for judging predatory pricing. First, price must be set below an appropriate measure of cost. Second, there must be a “dangerous probability” (under the Sherman Act) or “reasonable possibility” (under the Clayton Act) of

---


subsequent recoupment of lost profits. Necessary conditions for recoupment are market concentration, entry barriers, and capacity to supply the demand created by the prey’s exit.

The requirement that predatory pricing is shown to be profitable—a firm can recoup the cost of aggressive pricing—is a clear distinction from the standards applied to other forms of exclusionary behavior where one must only show it harms competition. In the ensuing six years since this decision, plaintiffs had not won a single case in federal court and, more striking, all but three of thirty-nine reported decisions were dismissed or failed to survive summary judgment.51 (Summary judgment is granted when, even if all disputed facts are in favor of one party, that party cannot, as a matter of law, win the case.) It appears this new standard has set a high hurdle for establishing predatory pricing.

Recent Developments

Part of the failure of plaintiffs in predatory pricing cases has been attributed to continued skepticism among judges about the plausibility of predatory pricing and the “judicial neglect of modern strategic theories of predatory pricing.”52

Exemplifying this point is the American Airlines case in 2001.53 American Airlines had responded aggressively to entry by low-cost carriers on routes out of its Dallas–Fort Worth (DFW) hub. After initially just matching the lower fares of these entrants on a limited basis, it expanded the availability of these lower fares and also capacity by increasing the number of flights and using bigger planes. It even entered the DFW–Long Beach route, which it had previously abandoned but now was served by one of the low-cost carriers. Entering a market now that there are more competitors is certainly suspect!

The cost of this strategy was estimated to be around $41 million, which one could not reasonably argue could be recouped on the affected markets. The DOJ then used a “reputation for predation” argument to contend that American Airlines could reap the benefits by deterring entry and expansion in other route markets. The District Court was unconvinced.

However, other recent cases suggest that the courts are beginning to recognize these strategic theories. Reputation as a method of recoupment was recognized in 1996 by the Third Circuit in Advo, Inc. v. Philadelphia Newspaper, Inc. Also, the recently proposed Department of Transportation guidelines for evaluating the presence of predatory pricing considered reputation for aggressive pricing as a mechanism by which to recoup profits from a predatory price war.

51. Brodley et al., “Predatory Pricing.”
52. Ibid, p. 2259.
Recognizing that no simple rule is going to work for a complicated activity like predatory pricing, a recent proposal is that the establishment of predatory pricing would require proof of (1) a facilitating market structure; (2) a scheme of predation and supporting evidence; (3) probable recoupment; (4) price below cost; and (5) absence of a business justification or efficiencies defense.54

**Refusal to Deal and the Essential Facilities Doctrine**

Antitrust laws are constructed on the premise that competition among firms is to be promoted. A tension with this premise emerges with a monopolization practice known as refusal to deal. The courts have identified circumstances under which a firm with market power has a duty to deal with a competitor, whether it is to supply a product—either an input or a complement to the rivals’ product—or provide valuable information. When a firm refuses to deal and that refusal is deemed to have anticompetitive intent, the firm can be found in violation of Section 2 of the Sherman Act.

Refusal to deal can arise in various circumstances.55 Two firms can provide products in the same market and one firm refuses to engage in a joint venture with the other. Exemplifying this case and representing a landmark decision by the Supreme Court in 1985 is *Aspen Skiing Company v. Aspen Highlands Skiing*.56 Aspen Skiing Company owned three of the four major mountain ski facilities in its geographic market, with Highlands owning the fourth. The two companies had sold a pass that allowed a skier to ski on any of the four mountains, with revenue being allocated on a pro rata basis. After several years, Aspen Skiing demanded a higher share of revenue. After Highlands refused to concede it, Aspen Skiing discontinued the joint venture. Soon thereafter, Highland’s share of skiers fell from 15 to 11 percent and it sought, without success, to reestablish its arrangement with Aspen Skiing. The Supreme Court found that there were no valid business reasons for Aspen Skiing’s refusal to return to the joint venture and that its refusal was anticompetitive in intent. However, in a recent decision,57 the Court stated that *Aspen Skiing* is “at or near the boundary of Section 2 liability.”

A second instance of refusal to deal is when a dominant firm refuses to deal with any customer or firm who is supplied by a competitor. This is, of course, exclusive dealing. As it is covered in chapter 8 as part of vertical restraints, we will refer the reader to the discussion there, with special emphasis on the *Standard Fashion* case.

54. Ibid, p. 2264.
A third type is when a firm has a monopoly over an input and also competes in the final product market. Refusal to deal arises when the input monopolist denies the input to competitors in the downstream market. The economics of this situation are explored in the discussion of vertical integration in chapter 8. There are many refusal-to-deal cases of this sort, including the landmark 1992 case involving Kodak, which we explore later.

**Essential Facilities Doctrine**

From a legal perspective, an important type of refusal to deal is when it involves an *essential facility*. The essential facilities doctrine originated in 1912 with the *Terminal Railroad* decision by the Supreme Court. A company controlled a key bridge at the St. Louis crossing of the Mississippi River. It was acquired by some railroads that then controlled all railway bridges and switching yards into and out of St. Louis. They then proceeded to deny access to competing railroads, which effectively shut them out from offering rail services to and through St. Louis. The court concluded that this was an attempt to monopolize and mandated that rival railroads be given access to the bridge, terminals, and approaches.

The conditions to apply the essential facilities doctrine were most clearly laid out in the more recent case between AT&T and MCI. MCI was competing with AT&T in the long-distance telephone market and the essential facility was AT&T’s local telephone network. For customers to use a non-AT&T long-distance supplier, they had first to connect to the local network, so as to access MCI’s network, and then connect with the local network of the person they were calling. The Seventh Circuit Court of Appeals found in favor of MCI and mandated access to the monopolist’s local network by competing long-distance providers.

In its decision, the Court stated that to establish antitrust liability under the essential facilities doctrine, one must prove: “(1) control of the essential facility by a monopolist; (2) a competitor’s inability practically or reasonably to duplicate the essential facility; (3) the denial of the use of the facility to a competitor; and (4) the feasibility of providing the facility to competitors.” Of course, the doctrine requires monopoly control of the facility and its denial to rivals. The other two conditions warrant comment. If it is reasonable for a competitor to create its own facility, such as its own bridge in the case of *Terminal Railroad*, then there is no monopolization. There must then be some entry barriers or it must be excessively costly to build. The fourth condition recognizes that there may be legitimate business rationales for refusing service. These four conditions are considered relatively stringent in that liability is rarely found under this doctrine.

60. *MCI Communications Co. v. AT&T*, 708 F.2d 1081 (7th Cir. 1982).
61. Ibid.
The essential facilities doctrine tends to arise in markets involving networks and natural monopolies. Whether it is a railroad network or a telecommunications network or an electric power network or yet some other type, provision of a service may require the cooperation of a competitor that owns a portion of the network. The real issue is whether that portion is a bottleneck; there is no reasonable way around it. A bottleneck is more likely when it is a natural monopoly; that is, the least-cost way in which to supply that part of the network is to have a single firm.62 In the case of long-distance telephone service in the 1980s, there was no feasible way around the local telephone network in that it would not be profitable for a long-distance supplier to build its own local network.

Another example of a network arose in Otter Tail Power Co. v. United States.63 In that case, municipalities sought to establish their own municipal power distribution systems, but to do so required purchasing electric power. Otter Tail Power, the local regulated utility currently distributing power, chose to withhold access to its power transmission lines. Those transmission lines were the essential facility, as they were the only means to deliver power to the municipalities. Otter Tail Power was found in violation of Section 2 of the Sherman Act because it refused to transmit competitive electricity over its lines and refused to sell wholesale power to the municipal systems.

**Intellectual Property Rights**

The basis for intellectual property rights such as patents and copyrights is to reward innovators with a temporary monopoly. This does not mean, however, that a patent holder can do as it wishes. With a legal monopoly does not come the right to further monopolize.

That antitrust law places limits on the use of intellectual property rights is well established. The Supreme Court had an opportunity to reaffirm this position in a recent case involving Kodak that we will look at in greater detail shortly. Kodak refused to sell its patented replacement parts for its photocopiers to independent service operators who were competing with Kodak in serving and repairing Kodak copiers. In supporting the Ninth Circuit’s overruling of a lower court’s summary judgment in favor of Kodak, the Supreme Court stated that.

> [t]he Court has held many times that power gained through some natural or legal advantage such as a patent, copyright or business acumen can give rise to [antitrust] liability if “a seller exploits his dominant position in one market to expand his empire into the next.” 64

---

62. For more on natural monopoly, the reader is referred to chapter 11.
Intellectual property rights were at the center of a highly publicized case against Intel in the late 1990s. By a 3-1 vote, the FTC chose to pursue an antitrust action alleging that Intel had engaged in monopolization practices in violation of Section 5 of the FTC Act. Intel was the dominant producer of the microprocessors used in personal computers and workstations. Intergraph made computer workstations and, as a customer of Intel, received advance information about the next generation of microprocessors from Intel. This business relationship began to sour when Intergraph sued Intel for patent infringement. In a retaliatory act, Intel withheld the type of proprietary information that was essential to Intergraph. Intel engaged in similar practices against Compaq and Digital, two other computer manufacturers.

On the eve of the trial in March 1999, the FTC settled with Intel. However, Intergraph continued with its litigation. Reaffirming the limits of the essential facility doctrine and refusal to deal as a monopolization practice, the court of appeals for the Federal Circuit ruled in Intel’s favor on the grounds that Intel was not a competitor to Intergraph in the market in which Intel was alleged to have monopoly power, and thus there could be no anticompetitive intent:

Intergraph argues that the essential facility theory provides it with the entitlement, in view of its dependence on Intel microprocessors, to Intel’s technical assistance and other special customer benefits, because Intergraph needs those benefits in order to compete in its workstation market. However, precedent is quite clear that the essential facility theory does not depart from the need for a competitive relationship in order to incur Sherman Act liability and remedy. . . . [T]here must be a market in which plaintiff and defendant compete, such that a monopolist extends its monopoly to the downstream market by refusing access to the facility it controls. Absent such a relevant market and competitive relationship, the essential facility theory does not support a Sherman Act violation.

Intergraph also phrases Intel’s action in withholding access to its proprietary information, pre-release chip samples, and technical services as a “refusal to deal,” and thus illegal whether or not the criteria are met of an “essential facility.” However, it is well established that “[i]n the absence of any purpose to create or maintain a monopoly, the [Sherman] act does not restrict the long recognized right of a trader or manufacturer engaged in an entirely private business, freely to exercise his own independent discretion as to parties with whom he will deal.” *United States v. Colgate & Co.*, 250 U.S. 300, 307 (1919).65

Though this case fell outside the realm of the Sherman Act, the Court did note that mandated access to intellectual property may be imposed where the defendant has demonstrated anticompetitive intent in refusing to license access to it. Establishing how intellectual property rights and antitrust law mesh is an important issue for determining the returns to innovation and thereby the incentives to invest in research and development.

Kodak and Monopoly Power in Aftermarkets

After taking ten years to resolve *Eastman Kodak Co. v. Image Technical Services, Inc.*, the ultimate resolution is one of the most important antitrust decisions in recent years. Though there is considerable controversy regarding the appropriateness of the decision, there is general agreement that it has the potential to spawn many antitrust cases and to influence firm behavior in a wide array of markets.

The case concerned the aftermarket, in the form of service and repair, for micrographic equipment and high-volume copiers sold by Kodak. Originally, the aftermarket was served by both Kodak and independent service organizations (ISOs). To operate in the aftermarket, ISOs needed Kodak’s patented replacement parts, which Kodak chose to sell to them. Soon after losing a service contract in a vicious price war with Image Technical Services, Kodak changed its policy regarding the sale of parts. Purchasers of parts now had to provide proof of ownership of the equipment, which meant that ISOs could no longer buy them. Seventeen of these ISOs brought suit on the grounds that Kodak’s refusal to deal was intended to leverage its monopoly over parts to the service market. After a series of judicial events, which are summarized in table 9.1, the courts ultimately found Kodak guilty of monopolization practices.

Part of the importance of the Kodak decision is that it is relevant to any aftermarket. For the purposes of this decision, a market is an aftermarket if it has three elements: (1) a consumer is purchasing a system comprised of multiple components; (2) the components are purchased at different points in time; and (3) to some degree a consumer is locked in to buying a company’s system after having bought some of the components. In the case of Kodak, a “copier system” is comprised of the copier and service because if it breaks down and is not repaired then the consumer receives no value. The second criterion is satisfied because the copier is typically purchased before service is purchased. The essence of lock-in is that some of the cost incurred in association with the equipment cannot be recovered if the customer purchases another company’s equipment. For example, if employees are trained on a particular company’s equipment, then there are additional training costs associated with purchasing new equipment. Or if the resale market is imperfect, then much of the original expenditure on equipment may not be recoverable. Lock-in is also described as the presence of switch-

68. Shapiro, “Aftermarkets.”
69. Although service comes after the purchase of the equipment, equipment and service could, in principle, be purchased simultaneously if the company offers, at the time of purchase, a completely specified service contract such as an unconditional warranty.
Monopolization and Price Discrimination

327

Table 9.1
Timeline for Eastman Kodak Co. v. Image Technical Services, Inc.

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1987</td>
<td>Seventeen independent service organizations (ISOs), including Image Technical Services, sue Eastman Kodak in district court for using its monopoly over parts to monopolize the service market.</td>
</tr>
<tr>
<td>1988</td>
<td>The district court of the Northern District of California grants summary judgment in favor of Kodak.</td>
</tr>
<tr>
<td>1990</td>
<td>After the plaintiffs appeal the district court’s decision, the 9th Circuit Court of Appeals overturns the summary judgment and remands the case for trial.</td>
</tr>
<tr>
<td>1992</td>
<td>After Kodak appeals the circuit court’s decision, the Supreme Court agrees with the circuit court.</td>
</tr>
<tr>
<td>1995</td>
<td>Jury trial begins in the district court of the Northern District of California.</td>
</tr>
<tr>
<td>1996</td>
<td>Kodak is found guilty of monopolization. Damages of $24 million are assessed, which are then trebled to $72 million. The Court issues a ten-year injunction requiring Kodak to sell parts to ISOs at nondiscriminatory prices.</td>
</tr>
<tr>
<td>1997</td>
<td>After Kodak appeals the District Court’s decision, the 9th Circuit Court of Appeals upholds liability but requires a new trial for calculation of damages.</td>
</tr>
</tbody>
</table>

ing costs, which are the costs associated with a customer’s changing the product being used. Lock-in gives a company an advantage over competitors in that a customer won’t leave unless the aftermarket price premium is sufficiently high so as to exceed switching costs.

Aftermarkets include more than markets for service and repair of equipment. By this criteria, computer software is an aftermarket for hardware. The two form a system, are typically bought at different points in time, and there is lock-in, as, for example, previously purchased software may only work with that hardware. Another example is computer operating systems (primary good) and applications (aftermarket good). The Kodak decision is then relevant to many markets.

There are several antitrust issues at work here. If Kodak has market power in the equipment market, to what extent can it leverage it to the market for service? In this context, refusal to deal is then just like tying in that Kodak is, in essence, requiring customers to buy both its equipment and its service. As the economics of tying is covered extensively in chapter 8, we will not review it here. Though the plaintiffs originally alleged that Kodak had tied service with its monopoly over parts, they dropped that claim at the end of the District Court trial.

The antitrust issue that draws our attention is whether a firm that has minimal market power in the primary market (for example, equipment) can have extensive market power in the aftermarket (for example, service). To analyze this question, consider the following simple scenario. Suppose there is one consumer type (or firms can price discriminate and are competing for each individual customer) who attaches value \( V(s) \) to the primary good when she anticipates purchasing \( s \) units in the aftermarket. For example, \( V(s) \) is the value of a copier when

a customer anticipates buying \( s \) units of service. A consumer decides from which firm to buy the equipment and then how much service to purchase in the aftermarket. Due to lock-in, a consumer buys both goods from the same firm.

There are two firms, firm 1 and firm 2, and for firm \( i \), \( p_i \) denotes the price of its equipment and \( r_i \) the per unit price of service. Firms have identical primary and aftermarket goods with per unit cost of \( c \) for the primary good and \( f \) for the aftermarket good. Since firms’ products are identical, they do not have much market power in the equipment market.

Given a consumer has purchased a primary good, let us derive the demand for the aftermarket good. A consumer wants to choose a level of service, \( s \), to maximize her net surplus, \( V(s) - rs \). The value to purchasing \( s \) units, \( V(s) \), and its cost, \( rs \), are depicted in figure 9.7(a). Maximizing net surplus then means maximizing the vertical distance between these two curves. The optimal value is shown as \( s^*(r) \). Note that it occurs where the slope of \( V(s) \) (that is, the marginal value of another unit of service) is equated to the slope of \( rs \) (that is, the price of service). If they are not equated (so that, for example, the marginal value of service exceeds its price), then a consumer can increase net surplus by purchasing more, as it adds to surplus by an amount equal to the marginal value of surplus less price.

\( s^*(r) \) is the demand for service as it states the amount of service demanded as a function of its price. As \( r \) rises, the curve \( rs \) shifts up, and the point at which the slope of \( V(s) \) is equated to the slope of \( rs \) falls. This is shown in figure 9.7(b), where the price of service rises from \( r' \) to \( r'' \) and the amount of service demanded declines from \( S^*(r') \) to \( S^*(r'') \).

Initially, let us assume that firms compete by simultaneously setting prices for equipment and service. These prices are contractually guaranteed. The consumer will purchase one unit of equipment (and some level of service) from the firm offering the highest net surplus. Suppose firm 1 expects firm 2 to offer equipment and service prices that give the consumer net surplus of \( W_2 \). Firm 1 will then want to choose \( p_1 \) and \( r_1 \) so as to maximize profit subject to getting the customer’s business (assuming that it yields at least zero profit). We can then represent firm 1’s problem as:

Choose \( p_1 \) and \( r_1 \) to maximize \( p_1 + r_1 s^*(r_1) - cs^*(r_1) - f \) subject to \( V(s^*(r_1)) - r_1 s^*(r_1) - p_1 \geq W_2 \).

\( p_1 + r_1 s^*(r_1) - cs^*(r_1) - f \) is firm 1’s profit from selling the primary good and \( s^*(r_1) \) aftermarket units. As \( V(s^*(r_1)) - r_1 s^*(r_1) - p_1 \) is the net surplus to the consumer buying from firm 1, the constraint requires it to be at least as great as the net surplus from firm 2. Otherwise, the consumer will not buy from firm 1.\(^{71}\)

---

71. If the net surpluses are equal, then the consumer is indifferent. To ensure the sale, a firm would then want to provide a little more net surplus than its rival. Making this change would only marginally change the results and would not affect any of our conclusions.
Figure 9.7
(a) Net Surplus to a Consumer from Equipment and Service. (b) Change in Demand for Service in Response to a Change in the Price of Service
In describing the solution to firm 1’s problem, the first step is to show that, at the profit-maximizing prices, the constraint is binding; that is, the net surplus provided by firm 1 equals $W_2$. Suppose it were not true, so that firm 1 offered net surplus in excess of $W_2$. By raising the price of the primary good, $p_1$, just a little bit, net surplus still exceeds $W_2$, so that the consumer still buys from firm 1. But now firm 1’s profit is higher by virtue of selling the primary good at a higher price. We conclude that, at a profit-maximizing solution, the net surplus from buying from firm 1 equals the net surplus offered by its rival:

$$V(s^*(r_1)) - r_1 s^*(r_1) - p_1 = W_2.$$  

Solving this expression for $p_1$, it follows that the price of the primary good must be set so that:

$$p_1 = V(s^*(r_1)) - r_1 s^*(r_1) - W_2.$$  \hspace{1cm} (9.10)

Substituting the right-hand side of equation 9.10 for $p_1$ in (9.9), the firm’s problem is now:

Choose $r_1$ to maximize $V(s^*(r_1)) - r_1 s^*(r_1) - W_2 + r_1 s^*(r_1) - cs^*(r_1) - f$,

which simplifies to

Choose $r_1$ to maximize $V(s^*(r_1)) - cs^*(r_1) - W_2 - f$.

As part of a profit-maximizing solution, firm 1 then chooses a service price to maximize the value of the good, $V(s^*(r_1))$, less the cost of producing it, $cs^*(r_1)$. But that just means maximizing total surplus, and we know the solution to that problem is to price at marginal cost. Hence, the profit-maximizing service price is $c$. Using equation 9.10, the profit-maximizing prices are then:

$$r_1 = c, \ p_1 = V(s^*(c)) - r_1 s^*(c) - W_2.$$  \hspace{1cm} (9.11)

Intuitively, a firm wants to set the service price so as to maximize the total surplus between the customer and the firm. Though it results in no profit from selling service, the firm can extract profit through the equipment price.

The final step is to show that competition results in the social optimum of price equaling marginal cost in both the primary market and aftermarket. To prove that marginal cost pricing is an equilibrium, suppose firm 2 prices at cost: $r_2 = c$ and $p_2 = f$. It follows that the net surplus from buying from firm 2 is $W_2 = V(s^*(c)) - cs^*(c) - f$. Substituting this into (9.11), the optimal prices for firm 1 are $r_1 = c$ and $p_1 = f$. In other words, if firm 2 prices at cost, then firm 1’s optimal response is to do so as well. By the symmetry of the model, it follows that firm 2 optimally prices at cost given firm 1 does. Competition drives firms to offer the highest surplus they can while covering their costs. We conclude that if firms offer identical products and they can simultaneously commit to both equipment and service prices, equilibrium has marginal cost pricing.
In the above setting, firms did not have market power in the primary market because they had identical products. In this situation, the lack of market power resulted in no market power in the aftermarket and thus no basis for an antitrust case. In fact, this is the argument originally made by Kodak’s attorneys:

Very early in the fact-finding process, Kodak moved for “summary judgment.” Kodak argued that there was no allegation that it had market power in the equipment market. Kodak claimed that equipment customers had many alternatives available to them and made purchase decisions based on the total cost of ownership, and thus any attempt by Kodak to extract higher profits from maintenance customers would result in equipment customers' taking their business elsewhere. Thus, because it could not have service market power, Kodak argued that as a matter of law it could not be found guilty of tying or monopolizing service markets.72

Though the district court judge did grant summary judgment, the Circuit Court reversed the decision.

A key assumption in the above analysis is that a firm commits to its service price at the time the customer is purchasing the equipment. However, it is often difficult to write a complete service contract. When there is no such commitment, which we will now assume, firms will not set the aftermarket price at marginal cost. There may then be a role for the antitrust authorities.

To enrich the setting a bit, suppose firms face a series of customers; one each period for simplicity. In each period, a firm sets its primary good price for the new customer and its service price for old customers who are now in the aftermarket. Furthermore, we will assume some lock-in. Let us show that pricing at marginal cost is not an equilibrium by showing that a firm’s optimal response to its rival pricing at cost is to set its service price above cost. Thus, firms pricing at cost is not an equilibrium.

If a firm prices both equipment and service at cost, then a firm earns zero profit in each period. Now suppose it instead prices service above cost. As long as service is not so much above cost that an equipment owner would choose to buy new equipment (in other words, there is some lock-in), the firm can earn positive profit from selling service. Because its future profit stream cannot be negative (for the firm can assure itself of non-negative profit by pricing at or above cost in the future), its total profit is now positive. Thus, it cannot be optimal for it to price at cost.73

The firm is able to generate profit through what is called installed base opportunism. There are those consumers who have purchased in the past—the installed base—and, due to lock-in, the firm has some market power over them. This allows it to price service above cost.

the earlier scenario, the service price was set before lock-in, at which point the firm had no market power.

The general point is that unless customers contract the aftermarket good’s price at the time of purchase of the primary good, and assuming there is lock-in, a firm will have market power in the aftermarket even when it has no market power in the primary good market. Of course, setting a high service price can develop a reputation for “ripping consumers off with service.” This reputation is costly in that the price that consumers are willing to pay for equipment is reduced by the prospect of a higher service price. While this reputational effect constrains how high a firm will set its service price, it will still set it above marginal cost.

Though firms do have market power in the aftermarket, it is unclear what the role of antitrust law should be here. Simply requiring Kodak to supply ISOs is insufficient, as Kodak can perform a “price squeeze” by pricing its parts so high that ISOs are unable to profitably compete with Kodak. To avoid that outcome requires the judiciary to ensure reasonable pricing of parts, but then the judiciary would be performing a regulatory role, a task for which it is ill-suited.

Microsoft Case

What Standard Oil was to the start of the twentieth century, Microsoft is to its end. As we all know, Microsoft supplies OSs for the PC—initially MS-DOS and then Windows—and many software applications for that OS, such as Word, Excel, and Internet Explorer. The Microsoft case is full of subtle antitrust issues and clever economic analysis. Besides involving one of the most important companies and one of the most important industries in the United States, at issue is the suitability of the Sherman Act for dealing with anticompetitive practices in the information economy.

Network Externalities

In any monopolization case, there must be a firm that is dominant. The source of Microsoft’s dominance is network externalities. Before examining the antitrust case, let us first review some of the central economic forces at work in the computer software industry.

A product has network externalities when its value to a consumer is increasing in how many other consumers use it. Communication networks, such as telephone systems and e-mail, have network externalities. The more people who are connected to a telephone network, the more valuable is it to have a telephone, because there are more people with whom to talk. Software applications, such as word processing packages, have network externalities because people like to exchange files. And, most relevant to the case at hand, computer OSs have network externalities, though a bit indirectly. Developers of software applications for OSs are more inclined to write programs when there are more users who might buy it. Hence, the
more people that use an OS, the more software is written for it, and, therefore the more consumers are willing to pay for that OS.

Products with network externalities have a unique set of features. To appreciate this point, let us derive a demand curve for such a product. To simplify matters, suppose there are two types of consumers, denoted $L$ and $H$, and presume that type $H$ consumers value the good more. When a total of $Q$ consumers use the product, $V_H(Q)$ will denote the valuation that a type $H$ consumer attaches to the product and $V_L(Q)$ the valuation of a type $L$ consumer. As the product is presumed to have network externalities, $V_H(Q)$ and $V_L(Q)$ are both increasing in $Q$, as depicted in figure 9.8. For a population of 1,000 consumers, assume there are 250 type $H$ consumers and 750 type $L$ consumers.

In deriving the demand curve, first note that it can take three possible values—0, 250, and 1,000—corresponding to no one buying, only type $H$ consumers buying, and everyone buying (note that if a type $L$ buys, then a type $H$ buys, too). As usual with demand curves, demand is zero if price is sufficiently high. If price exceeds $V_L(1,000)$ then no type $L$ consumer will buy, in that their value is below the price even in the best-case scenario of maximal network externalities. Furthermore, if only type $H$ consumers buy, then a type $H$ consumer is only willing to pay $V_H(250)$ which, as can be seen in figure 9.8, is less than $V_L(1,000)$. Thus, if price exceeds $\bar{p} (= V_L(1,000))$, then the only value for demand is zero.
What about if price is less than $\bar{p}$? Demand can still be zero, but it can also be positive; demand is not uniquely defined! If consumers expect no one to buy then this belief will be self-fulfilling in that no one will buy. That is, if each consumer expects $Q = 0$, then, since $V_H(0) = 0$ and $V_L(0) = 0$, they are not willing to buy at any positive price. Thus, there is a vertical segment at $Q = 0$ in figure 9.9 that depicts the relationship between demand and price. Another possibility, however, is that demand is 1,000. If all consumers expect $Q = 1,000$, then type $L$ (and therefore type $H$) consumers will buy as long as $p \leq \bar{p}$. The remaining possibility is for demand to be 250. For that to occur, price must satisfy $V_H(250) \geq p \geq V_L(250)$, so, when $Q = 250$ is expected, type $H$ consumers want to buy but type $L$ consumers do not. As shown in figure 9.8, this holds when $\bar{p} \geq p \geq p$.

Summing up, we have the odd-looking demand curve in figure 9.9. When price exceeds $\bar{p}$, demand is zero. When price is between $\bar{p}$ and $\hat{p}$, demand is 0 or 1,000. When price is between $\hat{p}$ and $\bar{p}$, demand is 0, 250, or 1,000. And when price is less than $\hat{p}$, demand is once again 0 or 1,000.

There are several key points to make. First, consumers’ expectations are crucial in determining demand for products with network externalities. For example, when price lies between $p$ and $\hat{p}$, demand is zero if each consumer expects no one else to buy, 250 if each consumer expects only those with high valuation to buy, and 1,000 if each consumer expects everyone

---

74. This is not exactly true, since if a consumer buys, then $Q = 1$. The analysis will be simpler if we gloss over this point.
to buy. This contrasts with more standard products in which each consumer’s purchasing decision is independent of what other consumers do, and thus a consumer’s expectations over other consumers is irrelevant. As an example, IBM introduced OS/2, which was generally considered to be a superior to Microsoft’s OS. However, IBM failed to convince people that it would succeed. Consumers didn’t buy it because they didn’t expect others to do so, and if not enough people bought it then there wouldn’t be much software written for it. Any behavior that influences consumer expectations can have a long-run impact on the demand for a product with network externalities.

Second, a critical mass of consumer support can be instrumental in the success of a product with network externalities. In the context of figure 9.9, suppose a firm must get all 1,000 consumers to buy in order to achieve profitability. Further suppose that consumers base their decision on the size of the installed base, that is, the number of consumers who previously purchased and thereby are using the product. If the firm is able to induce the type $H$ consumers to buy, so that initially $Q = 250$, it can induce the type $L$ consumers to buy by setting price below $V_L(250)$. However, if it cannot get that initial mass of consumers to buy, there is no price that will induce any consumer to buy. At work is a mechanism known as positive feedback: the more consumers that buy, the easier it is to induce additional consumers to buy. The trick is getting the critical mass to jumpstart the positive feedback process.

Third, markets with network externalities naturally lead to market dominance. Due to positive feedback, there can be tipping, whereby once a firm gets enough of a lead in its installed base it goes on to dominate. The firm with a higher installed base will offer a more attractive product and thus tend to add new consumers at a higher rate. This then mean that an initial advantage in installed base steadily grows: the better keep getting better. Though market dominance is to be expected for products with network externalities, it is worth noting that competition for monopoly—that is, firms competing to get an early installed base advantage—can, as a result, be intense.

Fourth, such markets are also prone to sustained dominance. The initial firm in an industry can persist in being dominant, even when superior technologies come along, because it has a large installed base. An entrant with a better technology faces the disadvantage of not having an installed base. A superior technology can then fail when consumers do not coordinate in making a move to adopting it and, as we will elaborate shortly, when a dominant firm engages in certain anticompetitive practices.

A number of antitrust issues emerge in the presence of network externalities. Since market dominance is to be expected, the first condition for monopolization is typically satisfied: the presence of monopoly power. In chapter 8, we argued that exclusionary contracts, such as exclusive dealing and tying, can deter more efficient entry by preventing an entrant from gaining the necessary critical mass to achieve profitability when there are scale economies. For example, if there is a large fixed entry cost then it must earn variable profit on enough consumers so as to cover that fixed cost. Though it is generally unprofitable for a dominant
firm to offer terms sufficient to induce all consumers to commit to buying exclusively from it, entry can be deterred by having enough (though not all) consumers agree to buy only from the dominant firm. A similar logic applies to markets with network externalities. Exclusionary contracts can keep a new firm’s installed base from reaching the critical mass required for the product to be attractive to consumers. This can drive out a more efficient firm or deter one from entering.\textsuperscript{75}

The effectiveness of exclusionary contracts is accentuated by the critical role of consumer expectations. It need not take much for a dominant firm to cause a new firm to fail. If it can convince consumers that the new technology will fail, then it will indeed fail. A reputation that a dominant firm will aggressively respond to entry can make such aggression unnecessary if it induces consumers not to buy out of fear of being stuck with an unpopular product. There are potentially many instruments available for a dominant firm to adversely influence consumer expectations over a new technology.

Predatory pricing is also potentially relevant in that aggressive pricing can keep an entrant from acquiring critical mass. But complicating matters is that penetration pricing is a legitimate business strategy in markets with network externalities. Penetration pricing refers to initially pricing low, perhaps even below marginal cost, so as to build an installed base. One should then expect below-cost pricing even when predation is not the intent. An antitrust challenge is to distinguish penetration pricing from predatory pricing.

In concluding, the video game market exemplifies many of the issues raised. Network externalities are present because the more consumers who own a particular video game platform, the more games will be written for it, since potential demand for those games is higher. The industry was originally dominated by pioneer Atari, but then in 1985 Nintendo launched its 8-bit system, the Nintendo Entertainment System (NES), which became the dominant platform. The later entry of Sega’s 8-bit system failed because NES had more games—an example of the power of network externalities at work. However, Nintendo also engaged in exclusionary contracts. A software developer who wrote a game for NES was prohibited from making that game available to other systems for a period of two years after the game’s release.

**Antitrust Case**

On May 30, 1990, the Federal Trade Commission opened an antitrust investigation of Microsoft.\textsuperscript{76} Though the FTC eventually dropped its case three years later, any thoughts that

\textsuperscript{75} Indeed, the effectiveness of exclusionary contracts in deterring entry may generally be greater with network externalities than with scale economies; see Carl Shapiro, “Exclusivity in Network Industries,” *George Mason Law Review* 8 (Fall 1999): 1–11.

the authorities were through with Microsoft were quickly dispelled as the DOJ immediately took over the case. It was to be the beginning of many years of litigious battles.

The DOJ claimed Microsoft used anticompetitive terms in its licensing and software agreements in order to perpetuate its monopoly in the OS market. As an example, Microsoft would require a computer manufacturer (also known as an original equipment manufacturer, or OEM), such as Compaq, IBM, or Dell, to pay Microsoft a fee for all computers that it sold, regardless of whether the manufacturer installed Windows or a rival’s system. Thus, an OEM would, in effect, be paying for two OSs for every computer in which it elected to install a non-Microsoft OS.

In July 1994, the DOJ and Microsoft resolved matters with a consent decree\(^\text{77}\) that put restrictions on the types of contracts Microsoft could use; for example, the practice just mentioned was outlawed. It also prohibited Microsoft from tying the sale of products to its OS. This decision became known as Microsoft I.\(^\text{78}\)

Three years later the DOJ filed another suit claiming that Microsoft had violated the provision of the settlement that prohibited tying. The claim was that Microsoft required OEMs to install its Internet Explorer (IE) browser along with Windows 95. In what became known as the Microsoft II decision, the Circuit Court concluded that technologically bundling did not violate the consent decree.\(^\text{79}\)

Just prior to the Microsoft II decision being delivered, the DOJ and numerous states filed suit that Microsoft had violated the Sherman Act with its recent behavior in the browser market. As later summarized by the Circuit Court:

Relying almost exclusively on Microsoft’s varied efforts to unseat Netscape Navigator as the preeminent internet browser, plaintiffs charged four distinct violations of the Sherman Act: (1) unlawful exclusive dealing arrangements in violation of Section 1; (2) unlawful tying of IE to Windows 95 and 98 in violation of Section 1; (3) unlawful maintenance of a monopoly in the PC operating systems market in violation of Section 2; and (4) unlawful attempted monopolization of the internet browser market in violation of Section 2.\(^\text{80}\)

The trial began in district court in October 1998 and a verdict was delivered in April 2000. Judge Thomas Penfield Jackson found Microsoft not guilty on the charge of exclusive dealing but guilty on the other three charges. He ordered that Microsoft be broken up into two companies, one with the OS and the other with applications (such as Word). Microsoft appealed the decision and, in June 2001, the Circuit Court upheld the third charge regarding

“maintenance of monopoly” but reversed the decision on monopolization and remanded the 
decision on tying back to the district court. The proposed remedy was also remanded.

Due to improper behavior, the Circuit Court removed Judge Jackson and he was replaced 
with District Judge Colleen Kollar-Kotelly. Now under a Republican presidential adminis-
tration, the DOJ chose not to pursue a breakup of Microsoft, after which the two sides reached 
a settlement agreement in November 2001 that restricted Microsoft’s future behavior but left 
the firm intact.

Tying and Monopolization of the Browser Market

Most of our attention will be focused on the monopolization charge upheld by the Circuit 
Court. However, let us briefly deal with the other two guilty charges found by the District 
Court.

As regards the tying charge, the plaintiffs claimed that Microsoft’s technological integra-
tion of Windows and IE, along with certain features of contractual arrangements between 
Microsoft and intermediate suppliers of browsers such as OEMs, were anticompetitive and 
a per se violation of Section 1.81 Microsoft disputed this claim and contended that the inte-
gration was done to enhance quality by making Windows a better applications platform. In 
light of the novel role of technology associated with this tying arrangement, the Circuit Court 
remanded the guilty charge on the grounds that a rule of reason  should be applied. It noted 
that “Applying per se analysis to such an amalgamation creates undue risks of error and 
of deterring welfare-enhancing innovation.”82 Because many of the facts pertaining to this 
charge are similar to those for the maintenance of monopoly charge, we will not discuss the 
tyling claim any further.

With the fourth charge, regarding monopolization, the government claimed and the 
District Court concluded that Microsoft had tried to leverage its monopoly in the OS market 
so as to monopolize the browser market. Microsoft engaged in a number of contractual 
arrangements with OEMs and Internet service providers (ISPs) to promote IE at the 
expense of Netscape Navigator. These arrangements will be reviewed shortly (though in 
the context of maintaining monopoly power in the OS market). The government also accused 
Microsoft of pricing predatorily by distributing IE free (even at a negative price in 
some instances). The Circuit Court reversed this decision because, in order to establish 
such a monopolization claim, the plaintiffs must show that the browser market could be 
monopolized. However, the Court found that the plaintiffs had neither defined the relevant 
market nor argued that there are the necessary entry barriers to protect a monopoly once 
gotten.

81. The reader is encouraged to go to the anticompetitive theory of tying discussed in chapter 8. It is quite relevant 
to the Microsoft case.
Maintenance of Monopoly in the Operating Systems Market

Arguably the most serious charge, and the one that the Circuit Court upheld, was that Microsoft engaged in anticompetitive practices to maintain its near-monopoly with Windows in the OS market.

Before diving into the economic arguments, some knowledge of computer technology will prove useful. As we all know, an OS runs software applications. To do so it uses application programming interfaces (APIs), which allow the application to interact with the OS. A platform is a collection of such APIs. In writing software, a developer needs to write it so that it works with an OS’s APIs. If the software is also to work on a different OS, then the code must be rewritten for its APIs. Such a process is known as “porting” and can be costly. Given the large number of Windows users, software developers generally write their programs for the Windows platform. Such is the advantage of having the largest installed base, as we know from our analysis of network externalities. Given fewer Mac users, not all of this software written for Windows’ APIs gets ported to the Mac OS.

A threat to Windows’ dominant position would be a technology that allows the same programs to run on all OSs. In that case, if a superior OS came along, it would not be a disadvantage to Microsoft because the existing software applications would run on it. This was the threat that Netscape Navigator and Java posed. Written by Sun Microsystems, Java was a cross-platform language that allowed a program to run on many OSs. Navigator relied on Java and could run software applications independently of Windows. For example, it ran on seventeen different OSs, including Windows, Mac, and various versions of the UNIX operating system. Referred to as middleware, Navigator and Java was a potential challenge to Windows’ position as the dominant platform.

To establish a monopolization claim, the plaintiffs must argue that (1) the accused firm has monopoly power in a relevant market, and (2) it has sought to maintain that monopoly through anticompetitive behavior (for example, harming a rival’s product as opposed to making one’s product better). In addressing the first part, the District Court concluded, and the Circuit Court affirmed, that the relevant market is the global Intel-compatible PC OS market. In that market, Windows’ market share exceeded 95 percent. Microsoft argued for a more encompassing notion of a market. For example, it wanted to include Mac OS as a competitor (in which case Microsoft’s market share would be around 80 percent). However, the District Court concluded that even if the price of Windows was raised substantially, few Windows users would switch to the Mac OS because of the cost of new hardware and fewer software applications. Microsoft also sought to include hand-held devices and middleware, a point we will return to shortly.

Of course, high market share is necessary but not sufficient for monopoly power. An airline may be the only carrier to fly from Des Moines to Omaha, but it knows it does not need to raise price by very much to induce another airline to enter and compete. But Windows did have a barrier to entry in the form of network externalities and a large installed base. A new OS would lack such an installed base, which meant that not much software might be written for it, and if there is not much software, consumers will not be inclined to buy it, even if the OS is vastly superior.

Having established that Microsoft did have monopoly power, the next issue was anticompetitive conduct. Here, the government rattled off a long litany of practices designed to prevent rival browsers from becoming a viable alternative platform for software development. One cannot do justice to the many offenses put forward by the government, but we can give a flavor of their arguments.

The two primary avenues for distributing browsers are OEMs (such as Dell), which install programs on the computers they sell, and ISPs (such as America Online), which offer browsers when someone signs up for Internet access. In both cases, Microsoft was found to have used contracts that severely disadvantaged competing browsers. OEMs were prohibited from altering the Windows desktop and the initial boot sequence. So, for example, an OEM could not take IE off the desktop and replace it with Navigator, even if the customer so desired it. As a case in point, in 1996 Compaq wanted to load the more popular Netscape browser on its machines and to remove the icon for IE from Windows 95. Microsoft informed Compaq that if it removed IE, Compaq would lose its license for Windows. Compaq complied with Microsoft’s wishes.

With regard to ISPs, Microsoft agreed to provide easy access to an ISP’s services on the Windows desktop in exchange for them exclusively promoting IE and keeping shipments of Navigator under 25 percent. Figure 9.10 gives some indication of the impact of these exclusionary agreements. AOL and CompuServe had contracts with Microsoft that restricted them in their ability to promote non-IE browsers. The market share of IE at those two ISPs rose from 20 percent to 87 percent in less than two years. By comparison, “IE Parity” ISPs refer to those ISPs in the top eighty whose browser choice was not known to be restricted and that had 10,000 or more subscribers. IE’s market share among them rose only to 30 percent, while the market share among all ISPs rose to 49 percent. In other words, those ISPs that had a choice chose IE much less often than those who were restricted. The contractual restrictions appear to have made a difference.

The District Court concluded that these restrictions did not improve the performance of IE but rather just lowered the usage of rival browsers. Microsoft argued that it was “exercising its right as the holder of valid copyrights” in imposing these restrictions on those to whom

84. There were also anticompetitive contracts with other companies such as Apple Computer.
Figure 9.10
Microsoft’s Share of the Browser Market (Three-Month Moving Average of Usage by ISP Category)

it licensed Windows. As reviewed earlier in the section Refusal to Deal, the courts have long noted that intellectual property rights do not provide an exemption to antitrust laws.

The government also argued that the integration of Windows and IE was anticompetitive in that it involved excluding IE from the “Add/Remove Programs” utility, commingling code so that deletion of IE would harm the OS and so that Windows could override the user’s default browser when it wasn’t IE. Here matters are not as clear as Microsoft argues that there are technical reasons for these features. The Circuit Court did not find them liable because the plaintiffs failed to argue that any anticompetitive effects outweighed the benefits identified by Microsoft.

Quite distinct from the previous practices but arguably the most serious in terms of its potential impact on the incentives to innovate is what Microsoft did to Java. We have already explained that Java was a potentially serious threat to the Windows platform. Though Microsoft announced its intent to promote Java, it actually crippled it as a cross-platform threat by designing a version of Java that ran on Windows but was incompatible with the one written by Sun, and then using its Windows monopoly to coerce parties (including Intel) to adopt its version.

Before moving on to the issue of remedy, there is an argument made by Microsoft that is worth mentioning. Microsoft attorneys contended that the District Court was being contradictory in excluding middleware such as Java and Navigator in defining the market for Windows but accusing Microsoft as engaging in monopolization practices against middleware in order to maintain a Windows monopoly. The Circuit Court responded that because the threat was “nascent,” that there was no contradiction. They then went on to say,

\[ \ldots \text{the question in this case is not whether Java or Navigator would actually have developed into viable platform substitutes, but (1) whether as a general matter the exclusion of nascent threats is the type of conduct that is reasonably capable of contributing significantly to a defendant’s continued monopoly power and (2) whether Java and Navigator reasonably constituted nascent threats at the time Microsoft engaged in the anticompetitive conduct at issue. As to the first, suffice it to say that it would be inimical to the purpose of the Sherman Act to allow monopolists free reign to squash nascent, albeit unproven, competitors at will—particularly in industries marked by rapid technological change and frequent paradigm shifts. As to the second, the District Court made ample findings that both Navigator and Java showed potential as middleware platform threat.}^{85} \]

**Remedies and Harm**

Three remedies were being widely discussed. One was to restrict conduct, as done in the Microsoft I decision. This would entail prohibiting some or all of the anticompetitive practices, such as limiting what Microsoft could put in a contract with an OEM. The other two

remedies were structural and much more draconian. In the line of Standard Oil and AT&T, the strategy was to reduce monopoly power by breaking up the company. One plan was to create three identical Microsoft companies, each with Windows and applications.\textsuperscript{86} Dubbed “Baby Bills,” as a play on the Baby Bells that rose from the ashes of the AT&T breakup, the objective was to inject competition into the OS market. The second structural remedy was to create two companies, one with Windows and one with applications (such as Word and Excel). This is what the government proposed and the District Court mandated. However, the Circuit Court remanded the remedy because the District Court had not adequately considered the facts or justified its decision.

At that point, the government stopped pursuing a structural remedy. A settlement restricting the conduct of Microsoft was reached in November 2001. For example, the remedy prohibits Microsoft from retaliating against an OEM for using or promoting software that competes with Microsoft products and requires uniform licensing agreements to OEMs (so as to make it more difficult for Microsoft to hide rewards and punishments in discriminatory agreements). It also seeks to make it more difficult for Microsoft to hamper the development of middleware and more generally to promote interoperability by, for example, mandating that it make available the APIs and related documentation that are used by middleware to interact with a Windows OS product. Whether this conduct remedy will prove any more successful than the first one remains to be seen.

### Price Discrimination and the Robinson-Patman Act

The Clayton Act provision that price discrimination is illegal where it substantially lessens competition or tends to create a monopoly was amended in 1936 by the Robinson-Patman Act. The resulting, quite lengthy Section 2 of the Clayton Act is reproduced in the appendix of chapter 3.

According to Judge Richard Posner, “The Robinson-Patman Act is almost uniformly condemned by professional and academic opinion, legal and economic.”\textsuperscript{87} It was passed during the Great Depression largely to protect small, independent retailers from the newly emerging chains. For example, it outlaws brokerage fees unless an independent broker is involved. A&P was found to be in violation for buying directly from suppliers and performing its own wholesaling function. The “discrimination” was seen as A&P paying lower prices for supplies than independent stores pay that buy through brokers or wholesalers. The fact

\textsuperscript{86} For a description of this structural remedy, see Robert J. Levinson, R. Craig Romaine, and Steven C. Salop, “The Flawed Fragmentation Critique of Structural Remedies in the Microsoft Case,” Antitrust Bulletin (Spring 2001): 135–62.

that A&P was lowering its distribution costs—an economic benefit for society—by performing its own wholesaling function was irrelevant.

The economic definition of price discrimination—charging different customers prices that are not in proportion to marginal costs—is almost completely turned on its head in the Robinson-Patman Act and its enforcement. For example, in cases where some injury to competition has been found, the act has been interpreted as holding simple price differences to be illegal, regardless of cost differences. (Strictly, cost differences are a possible defense in the language of the law, but in practice it is virtually impossible to employ this defense. A second possible defense is that the discrimination was required to “meet competition.”)

We begin by reviewing the economics of price discrimination. The traditional analysis considers a monopolist and examines equilibrium prices, profitability, and efficiency. This form is referred to here as systematic (or persistent) discrimination, to distinguish it from unsystematic (or temporary) discrimination. Unsystematic discrimination means situations of disequilibrium in which shifts of demand or cost lead to price changes that are not uniform to all buyers. In other words, normal competitive processes of adjustment to new equilibria can involve differences in prices that might be viewed (incorrectly) as price discrimination by antitrust authorities.

**Systematic Discrimination**

Price discrimination was discussed in chapter 8, where tying and block booking were examined. There are three types of price discrimination: first-degree or perfect discrimination (in which the monopolist obtains all of the consumer surplus), second-degree (such as tying), and third-degree (in which demanders are partitioned into groups and each group is charged a different price, such as discounts for children at the movie theater).

The distinction between second- and third-degree discrimination is that in second-degree discrimination all demanders confront the same price schedule, but they pay different average prices, depending on their preferences;\(^{88}\) in third-degree discrimination the seller separates demanders into different groups based on some external characteristic (such as age) and confronts the groups with different prices. Clearly, in third-degree situations the seller must be able to keep resales from occurring. If children’s tickets could be used by adults, no adults would pay the higher price of an adult ticket. It is also necessary for the seller to possess some degree of market power; otherwise there is no way prices can differ (inasmuch as price equals marginal cost in competitive markets).

---

88. For example, in the tying illustration discussed in chapter 8, the two customers faced the same price schedule (a copying machine price of $2,812.50 and a price per package of copying paper of $25); however, one customer paid an average price per “unit of copying services” of $62.50, compared with only $43.75 for the other customer. The amounts differed because the two customers’ demands for copying services differed.
The economics of third-degree discrimination can be explained by example. Consider duPont’s patented superstrength synthetic fiber Kevlar. To simplify, assume that there are only two uses: in undersea cables and in tires. Because tire companies have the option of using low-cost substitutes in tires—steel and fiberglass—the demand for Kevlar by tire companies is more elastic (at a given price) than that by cable companies. That is, because of its technical characteristics, Kevlar is far superior to possible substitutes in undersea cables.

Assume hypothetically that the demand curves for Kevlar are

\[ q_c = 100 - p_c \quad \text{for use in undersea cables} \]
\[ q_t = 60 - p_t \quad \text{for use in tires}. \]

For simplicity, let the marginal cost (MC) be constant at $20. This situation is shown in figure 9.11, where cable demand is shown on the right and tire demand on the left. That is, we have “flipped” tire demand to the left of the origin, and so we measure its output as increasing left to right. This procedure is followed to make the diagram less cluttered.

Figure 9.11
Price Discrimination That Decreases Total Surplus

First, consider the profit-maximizing solution assuming that duPont can charge different prices to tire and cable companies if it is profitable to do so. The logic is simple. Set the marginal revenue \( (MR_c) \) from cable companies equal to the marginal revenue \( (MR_t) \) from tire companies, and set both equal to \( MC \). Clearly, if the marginal revenues differed, duPont would find it profitable to shift a unit from the lower marginal revenue market to the higher. The solution \( (MR_c = MC \text{ at point } N \text{ and } MR_t = MC \text{ at point } G) \) is to sell 40 units to the cable market at a price of $60 and 20 units to the tire market at a price of $40. The profits in the two markets are found by computing revenues less costs, or $1,600 from cable users and $400 from tire users, for a total of $2,000.\(^90\)

It is instructive to note that the higher price is charged in the market in which the elasticity of demand is lower. That is, at the equilibrium, the elasticity of demand is 1.5 in the cable market and 2 in the tire market.\(^91\) If the elasticities were not different, the prices would be the same and discrimination would not be profitable.\(^92\)

Next, consider the discrimination-disallowed, or single-price, equilibrium. Here it is necessary to aggregate the two demands by adding them horizontally to get the total demand curve. This is a kinked curve where the kink is at point \( A \). Above point \( A \), the total demand curve corresponds to cable demand only, because the tire users will pay no more than $60. The marginal revenue curve associated with the total demand curve, \( SMR \),\(^93\) intersects \( MC \) at a total output of 60 (at point \( T \)). Hence the single price is $50, and at this price cable users buy 50 units and tire users buy 10 units. Finally, the profit is again found by computing revenues minus costs, or $1,800. As anticipated, profit is lower if discrimination is not permitted.

Summarizing,

90. Profit from cable users = \( pq_c - (MC)q_c \)
   \[ = (60)(40) - (20)(40) = 1,600 \]

91. The elasticity of demand is \( -(dq/dp)(p/q) \), so substituting \( p_c = 60, q_c = 40 \), and \( dq/dp = 1 \), we get 1.5. Similarly, in the tire market the elasticity is 2.

92. This point can be understood by recalling the standard formula \( MR = p(1 - 1/\eta) \), where \( \eta \) is the absolute value of the elasticity. Hence, if \( MR_c = MR_t \), then \( p_c(1 - 1/\eta_c) = p_t(1 - 1/\eta_t) \), and if \( \eta_c = \eta_t \), then \( p_c = p_t \).

93. The SMR is actually \( MNPTR \), not just \( PTR \). There is a discontinuous jump upward at point \( N \) where the kink in total demand occurs. We denote this SMR for “simple marginal revenue,” following the exposition in Joan Robinson, The Economics of Imperfect Competition (London: Macmillan, 1933), chap. 15. The SMR should not be confused with the aggregate \( MR \) schedule, which is the horizontal sum of \( MR_c \) and \( MR_t \). We do not need that schedule here because of the assumption that \( MC \) is constant. The aggregate \( MR \) schedule is used in the discrimination case for determining total output when \( MC \) is not constant.
347  Monopolization and Price Discrimination

<table>
<thead>
<tr>
<th>Discrimination</th>
<th>Single Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_c = 60$</td>
<td>$p_c = p_t = 50$</td>
</tr>
<tr>
<td>$q_c = 40$</td>
<td>$q_c = 50$</td>
</tr>
<tr>
<td>$p_t = 40$</td>
<td>$q_t = 10$</td>
</tr>
<tr>
<td>$q_t = 20$</td>
<td>$q_c + q_t = 60$</td>
</tr>
<tr>
<td>$q_c + q_t = 60$</td>
<td>Total profit = $2,000$</td>
</tr>
<tr>
<td>Total profit = $2,000$</td>
<td>Total profit = $1,800$</td>
</tr>
</tbody>
</table>

The Kevlar example makes it clear that third-degree discrimination, the subject of the Robinson-Patman Act, is profitable. The next question is to examine the efficiency of price discrimination. For our Kevlar example, we can compare total economic surplus under the two scenarios, discrimination and single price. It turns out that in situations where there are linear demand curves, and where both demand groups buy positive amounts in the single-price case, total surplus always falls when discrimination is allowed. This is the case in our example.

In figure 9.11 it is easy to see this result graphically. In the cable market, the area of the shaded trapezoid $ANBC$ gives the loss in total surplus in moving from the single price of $50 to the discriminatory price of $60. Total value of Kevlar to cable users is reduced by the area under the demand curve and between outputs of 40 and 50. Subtracting the cost saving from reducing output (the area under $MC$ and between these two outputs) yields the trapezoid $ANBC$. This area equals $350$.

Similarly, the area of shaded trapezoid $HGFE$ in the tire market gives the gain in total surplus in moving from the single price of $50 to the lower discriminatory price of $40. This area equals $250$. Hence the net change in total surplus is a gain in the tire market of $250 minus a loss in the cable market of $350, or a net loss of $100$. For cases like this example (with linear demands), it is always true that the output changes in the two markets are exactly equal and opposite in direction—that is, total output is unchanged. Hence the two trapezoids are equal in width, but the loss trapezoid is taller.

It is important to provide an intuitive explanation for the net loss result in this example. Total output is unchanged, and all that happens is a reallocation of output from high-value users (at the margin) to low-value users. In the single-price case, all users end up with the same marginal valuation. This “gap” is the inefficiency that produces the net loss. Trading between cable users and tire users, which is not allowed by duPont, would make both groups better off. (It is instructive to note that duPont’s enforcement of different prices was attacked by a customer as an antitrust violation.)

94. Because the demanders are firms rather than individuals, the demand curves are marginal revenue product schedules, and the area under these curves represents total revenue to the buying firms.
As noted, we have been examining a rather special case. To show a different welfare result, we need change our example only slightly. Assume now that the demand for Kevlar by cable companies is unchanged, as is the marginal cost. However, now assume that the demand by tire users is smaller than before. In particular, let

\[ q_t = 40 - p_t \] for use in tires

Carrying through the same analysis as before, we find that in figure 9.9 the discrimination solution is where \( MR_c = MC \) and \( MR_t = MC \), so \( p_c = \$60 \) and \( p_t = \$30 \). Now, however, the single price equilibrium where \( SMR = MC \) yields a price of \( \$60 \). Given that the tire users will pay at most \( \$40 \), the tire users will obviously buy zero units of Kevlar. They will simply use fiberglass and steel in their tires. This situation is therefore different from the case in figure 9.11, where the tire users did buy positive amounts of Kevlar when a single price was used.

Summarizing,

<table>
<thead>
<tr>
<th>Discrimination</th>
<th>Single Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_c = $60 )</td>
<td>( p_c = p_t = $60 )</td>
</tr>
<tr>
<td>( q_c = 40 )</td>
<td>( q_c = 40 )</td>
</tr>
<tr>
<td>( p_t = $30 )</td>
<td>( q_t = 0 )</td>
</tr>
<tr>
<td>( q_c = 10 )</td>
<td>( q_c + q_t = 40 )</td>
</tr>
<tr>
<td>( q_c + q_t = 50 )</td>
<td>Total profit = $1,700</td>
</tr>
<tr>
<td></td>
<td>Total profit = $1,600</td>
</tr>
</tbody>
</table>

Discrimination, as before, yields higher profits (\$1,700 versus \$1,600). Now, however, total output increases under discrimination. This increase is, of course, simply the output purchased by tire users who were not in the market when the single price was used. There has been no change whatsoever in the cable market, so the only change is that discrimination has permitted purchases of Kevlar by tire users. The interesting result is that welfare has been improved by discrimination. Not only is total surplus higher, by the shaded area in figure 9.12, but this is an increase in welfare under the Pareto criterion as well. Notice that no one is harmed—the cable market is unchanged—and both duPont and tire users gain. The duPont company gains by a profit increase of \$100 (the square area of the shaded region), and the tire users gain by \$50 (the triangular portion of the shaded area).

Finally, we should note that in cases of nonlinear demand curves the welfare result can be either positive or negative. It is very difficult to provide any general conditions in this

95. Similar to the point made in note 53, the SMR coincides with \( MR \), up to output 60; it then jumps vertically from a negative value to point \( P \) and includes segment \( PR \).
situation. It is true, however, that total output must increase if discrimination is to improve welfare.\textsuperscript{96}

**Unsystematic Discrimination**

Unsystematic discrimination refers to various situations in which prices differ because of temporary changes taking place in the market. Movements toward equilibria are always taking place in competitive markets as shifts in demand or costs occur. That is, the long-run competitive equilibrium described in microeconomics texts is best viewed as a target that is constantly shifting, and competitive markets are always adjusting toward that equilibrium.

For example, prices in one geographic region may rise temporarily because of a demand increase until new supplies can be made available (either through capacity additions or transport from distant regions). These temporary “discriminatory” situations clearly are necessary for markets to function efficiently and should not be forbidden by law.

A different situation that can be considered under this category is the cheating of a cartel member. A cartel member may decide that secretly shading the cartel price to a few large

customers will be profitable. An across-the-board cut to all customers would be too dangerous because fellow cartel members would probably discover the cheating and match the cut. Nevertheless, such cheating has a tendency to spread and has often led to the deterioration of cartel price structures—a socially beneficial result. A good example is probably the nineteenth-century railroad cartels, which were known for frequent breakdowns of their agreed-on freight rates.

In summary, there are cases of unsystematic, or temporary, discrimination that are clearly socially beneficial. The difficulty is to distinguish in practice between this type of discrimination and systematic discrimination practiced by an entrenched monopolist that may be harmful. Hence, laws against price discrimination are difficult to write and enforce if they are to promote competition.

In recent years the FTC has brought few complaints under the Robinson-Patman Act, although there are still actions brought by private parties. Because of this trend, we will discuss only two of the best-known cases here.

Cases

Two main categories of Robinson-Patman cases are those involving primary-line discrimination and those involving secondary-line discrimination. Primary-line discrimination refers to situations in which the seller practicing discrimination injures its own rivals. Predatory pricing is an extreme example, since it requires prices set below costs. Less severe discrimination that harms one’s rivals without being predatory also qualifies as primary-line discrimination.

Secondary-line discrimination occurs when injury to competition takes place in the buyers’ market. The idea is that buyers who get preferentially low prices will have an advantage over their rivals.

We consider two famous cases, one in each category. The primary-line case is a private case known as Utah Pie. A small frozen dessert pie manufacturer in Salt Lake City, Utah Pie Company, brought the suit against three large rivals: Continental Baking, Carnation, and Pet Milk. The three large rivals had manufacturing facilities in California, but not in Utah. Hence, when Utah Pie opened its frozen pie plant in Utah, it obtained a significant cost advantage over its larger rivals. Utah Pie had market-share percentages of 66.5 in 1958, 34.3 in 1959, 45.5 in 1960, and 45.3 in 1961. Also, the market was expanding rapidly over this period, and Utah Pie’s actual sales steadily rose as well.

The Supreme Court noted in its opinion that Continental, for example, set a price in Salt Lake City that was “less than its direct cost plus an allocation for overhead.” Also, the rivals tended to charge less in Utah than they did in other locations. According to the Court,

At times Utah Pie was a leader in moving the general level of prices down, and at other times each of the respondents also bore responsibility for the downward pressure on the price structure. We believe that the Act reaches price discrimination that erodes competition as much as it does price discrimination that is intended to have immediate destructive impact.

Hence, Utah Pie won its case, even though the decision has been regarded by most scholars as a mistake. Justice Stewart, in a dissenting opinion, put the argument clearly when he said that the market should be viewed as more competitive in 1961 than it was in 1958, not less competitive, as the Court held. One reason is that the dominant firm, Utah Pie, had a 66.5 percent share of the market in 1958 but only 45.3 percent in 1961. According to Justice Stewart, “the Court has fallen into the error of reading the Robinson-Patman Act as protecting competitors, instead of competition.”

The 1948 *Morton Salt* case is the landmark case involving secondary-line discrimination.98 Morton Salt sold its Blue Label salt to wholesalers and to chain stores according to the following table of discounts:

<table>
<thead>
<tr>
<th>Price/Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less-than-carload purchases            $1.60</td>
</tr>
<tr>
<td>Carload purchases                      1.50</td>
</tr>
<tr>
<td>5,000-case purchases in any consecutive 12 months 1.40</td>
</tr>
<tr>
<td>50,000-case purchases in any consecutive 12 months 1.35</td>
</tr>
</tbody>
</table>

Only five customers ever bought sufficient quantities of salt to obtain the $1.35 per case price. These were large chain stores, and the Court was concerned that small independent food stores could not compete with the chains because they had to pay, say, $1.60 per case. As the Court put it, “Congress was especially concerned with protecting small businesses which were unable to buy in quantities, such as the merchants here who purchased in less-than-carload lots.”

Morton Salt tried to defend itself by claiming that the discounts were available to all customers. However, the Court rejected this claim, saying that “theoretically, these discounts are available to all, but functionally they are not.”

In conclusion, although economic analysis reveals that there are cases in which prohibiting price discrimination is socially beneficial, there are many cases in which it should not be prohibited. The Robinson-Patman Act appears to be a poor instrument for distinguishing the cases, and reforming the act may be too difficult. This conclusion is similar to the criticism of predatory pricing law in that the attempt to prohibit certain harmful practices may be, on balance, more harmful than doing nothing, because of the chilling effect on socially desirable price competition that such laws may cause.

Former judge Robert Bork has put the point colorfully: “One often hears of the baseball player who, although a weak hitter, was also a poor fielder. Robinson-Patman is a little like that. Although it does not prevent much price discrimination, at least it has stifled a great deal of competition.”

Summary

The law toward monopolization is one of the most difficult antitrust laws to apply. In part this is due to the number of ways (some good, some bad) of winning and maintaining market dominance—superior product, better management, scale economics, predatory tactics, exclusionary contracts, and on and on.

In this chapter we have reviewed the evolution of antitrust case law regarding monopolization. Initially, the courts required, in addition to a large market share, evidence of abusive or predatory acts to show intent. Standard Oil is a case in point. Then, roughly from 1945 to 1970, the courts did not require evidence of abusive acts to infer intent, with United Shoe Machinery being a representative case. Since that time, the courts have been willing to allow more aggressive practices by dominant firms without inferring intent to monopolize. The 1993 Brooke decision meant a more stringent criteria for arguing that a firm has engaged in predatory pricing. In the 2001 Microsoft decision, the Circuit Court applied a strict but fair standard in assessing what was and was not monopolization. A counterexample is the 1992 Kodak decision by the Supreme Court, which has opened many aftermarkets up to antitrust litigation.

The case law over the last century has resulted in the following principles regarding the establishment of a violation of Section 2 of the Sherman Act. To begin, the firm must have monopoly power, and this requires properly defining the market and arguing that there are entry barriers to allow it to maintain that monopoly. With that prerequisite satisfied, a dominant firm must act to harm the competitive process as opposed to competitors. If the plaintiffs have successfully showed that anticompetitive harm has occurred, the defendants can offer a pro-competitive justification for their actions. If that justification withstands rebuttal by the plaintiffs, then they must argue that the anticompetitive effects outweigh the pro-competitive effects. Only then will a firm be found guilty of monopolization.

Questions and Problems

1. If a large firm is found to possess monopoly power, what else is needed to find the firm guilty of monopolization? Why is possessing monopoly power insufficient for illegality?

2. In chapter 7 the merger guidelines were described. In particular, the relevant market was defined as consisting of all products and firms such that a hypothetical cartel could raise its price by 5 percent and not have to rescind it because of unprofitability. In light of the duPont Cellophane case, how might this rule be modified to avoid the “error” of defining the market to be too broad?

3. Which of the three market definitions considered by Judge Hand in the Alcoa case do you think is most defensible?

4. IBM redesigned its disk drive 2319A to make it more difficult for rivals to interface with IBM computer systems. The Justice Department regarded this as anticompetitive. Comment.

5. A three-firm Cournot industry has a demand curve of \( Q = 20 - P \). Each firm has an annual total cost of \( 4q_i + 12 \). Find the equilibrium price, output of each firm, and profit per firm.
   a. The management of firm 1 is considering a strategy of predatory pricing, since management believes monopoly profits would far exceed current profits. What are the potential monopoly profits in this industry?
   b. Ignoring any antitrust concerns, management wants at least to do an investment analysis of predatory pricing. They decide that to drive the other two firms out of the market, a price of $2 per unit is needed, and that it must be maintained for at least three years. Other assumptions were put forth, but the consensus was the $2 price, three-year assumption. Given that a six-year time horizon and a 14 percent cost of capital are standard assumptions for firm 1 in its investment decisions, is predatory pricing profitable? Present-value formulas show that at 14 percent interest, a stream of $1 receipts for three years has a present value of $2.322 and a stream of $1 receipts from year 4 to year 6 has a present value of $1.567.
   c. Assume that whatever the given numbers show, the management could choose numbers that make the investment profitable. If it did appear profitable, say, net present value is $10, should predation be pursued? What other considerations should the management want further information about?

6. A new chemical has been discovered that can be produced at a constant marginal cost of $10 by its patent holder, Johnson, Inc. Two industries, A and B, find the chemical, Cloreen, to be useful in their production processes. Industry A has a demand for Cloreen of \( qa = 100 - pa \). Industry B’s demand is \( qb = 60 - pb \).
   a. If Johnson can prevent resales between industries A and B, what prices will it charge to A and B? It can be assumed that the patent gives Johnson monopoly power. What quantities will be sold to the two industries, and what will be Johnson’s profit?
   b. Assume now that it is illegal for Johnson to charge different prices to A and B. What price will Johnson now charge, and what will its profit be? What is Johnson’s quantity sold?
   c. Is total economic surplus higher in part a or in part b? What is the difference in total surplus in the two cases?
   d. Assume now that the demand for Cloreen by industry B is less than before, and is \( qb = 40 - pb \). Aside from this change, the facts are as given previously. Answer parts a, b, and c given the changed demand by industry B.

7. Assume that you are to decide whether a third-degree discrimination situation is in society’s interest. Assume that under the “no discrimination allowed” case, one group of users (group B) of a new product finds its price too high to buy any of the product. But under discrimination they
would buy a positive amount at the lower price offered them. Further, assume that the price charged to the original group of buyers (group A) remains unchanged after discrimination is permitted.

a. The legal permission to price-discriminate would benefit the monopolist, but the original group of buyers (group A) would be harmed. Do you agree or disagree? Why?

b. The legal permission to discriminate would benefit group B buyers and the monopolist, and group A buyers would be unaffected, so the permission to discriminate is a Pareto superior move compared with the no-discrimination situation. Do you agree or disagree? Why?

c. A Pareto superior move will always pass the “increase in total economic surplus” test, but the reverse is not true. True or false? Why?

8. Justice Antonin Scalia, in his dissent to the 1992 decision *Eastman Kodak Co. v. Image Technical Services, Inc.*, made the observation that if Kodak had required consumers to purchase a lifetime parts and service contract with each machine, the tie-in between service and parts would not have been a violation of Section 2 of the Sherman Act. What is the argument behind this claim?

9. For the example of a product with network externalities used in the chapter, rederive the demand curve when instead $V_H(250) > V_L(1,000)$.

10. Put aside any personal feelings you might have about Bill Gates and Microsoft. What do you think should have been the remedy used? Between the two structural remedies, which would have had more of an effect on market power in the operating systems market? In the applications market?
ECONOMIC REGULATION
10 Introduction to Economic Regulation

What Is Economic Regulation?

The essence of free enterprise is that individual agents are allowed to make their own decisions. As consumers and laborers, each person decides how much to spend, how much to save, and how many hours to work. Firms decide which products to produce, how much to produce of each product, what price to charge, how much to invest, which inputs to use and from which suppliers to buy them. In all modern economies, there is also an entity called government, which decides on such things as the income tax rate, the level of national defense expenditure, and the growth rate of the money supply. Government decisions like these affect both the welfare of agents and how they behave. For example, raising the income tax rate induces some individuals to work fewer hours and some not to work at all. Although an income tax influences how a laborer behaves, the laborer is left to decide how many hours to work. In contrast, in its role as regulator, a government literally restricts the choices of agents. More formally, regulation has been defined as “a state imposed limitation on the discretion that may be exercised by individuals or organizations, which is supported by the threat of sanction.”

As has long been noted, the key resource of government is the power to coerce. Regulation is the use of this power for the purpose of restricting the decisions of economic agents. In contrast to the income tax, which does not restrict the choices of individuals (though it does affect their welfare), the minimum wage is a regulation in that it restricts the wages that firms can pay their laborers. Economic regulation typically refers to government-imposed restrictions on firm decisions over price, quantity, and entry and exit. Economic regulation is to be contrasted with social regulation, which is discussed in part III of this book.

When an industry is regulated, industry performance in terms of allocative and productive efficiency is codetermined by market forces and administrative processes. Even if it so desires, a government cannot regulate every decision, as it is physically impossible for a government to perfectly monitor firms and consumers. As a result, market forces can be expected to play a significant role regardless of the degree of government intervention. For example, under airline regulation, the government controlled price but not the quality of service, and this induced firms to shift competition from the price dimension to the quality dimension. Even in a government-controlled economy like the former Soviet Union, market forces were at work. Although production and price were set by the state, the (effective) market-clearing price was set in the market. If a good was in short supply, people would wait in line for it. The effective price to them was the price paid to the state plus the value of their time spent in line. In equilibrium, people stand in line until the effective price clears the market.

Instruments of Regulation

Although economic regulation can encompass restrictions on a wide array of firm decisions, the three key decision variables controlled by regulation are price, quantity, and the number of firms. Less frequently controlled variables include product quality, advertising, and investment.

Control of Price

Price regulation may specify a particular price that firms must charge, or may instead restrict firms to setting price within some range. If the concern of the government is with a regulated monopolist setting price too high, regulation is apt to specify a maximum price that can be charged. For example, in 1989 the Federal Communications Commission (FCC) instituted price caps to regulate AT&T’s long-distance rates. If the regulated firm has some unregulated competitors, the regulatory agency may also be concerned with the regulated firm engaging in predatory pricing (that is, pricing so as to force its competitors to exit the market). In that situation, regulation is likely to entail a minimum price as well as a maximum price. In some cases, like the control of oil prices in the 1970s, regulation required that a specific price be set.

More often than not, regulation specifies more than a single price. In fact, it can put an entire price structure in place. For example, the regulation of AT&T in the intercity telecommunications market required the FCC to specify long-distance rates for different times of day and for different days of the week. The specification of a price structure as opposed to just a single price greatly increases the complexity of implementing economic regulation and can result in additional welfare losses, as we will observe.

In practice, price regulation may be the means by which a regulatory agency achieves an ultimate objective of limiting industry profit. A regulatory agency often sets price so that the regulated firm earns a normal rate of return. This is standard practice in the regulation of public utilities and has been used in other regulated industries such as the airline industry prior to its deregulation. As firm profit is determined by a variety of factors (with price being just one of them), a regulatory agency may have a difficult time in achieving its goal of a normal rate of return. Regulatory lag in changing price in response to new cost and demand conditions can result in a regulated firm earning either too high or too low a rate of return. During the inflationary period of the 1970s, rising input prices resulted in public utilities often earning a below-normal rate of return because the regulatory agency was slow to adjust price. Alternatively, a regulated firm that experiences an innovation in its production technology will reap above-normal profits until the regulatory agency realizes the cost function has shifted down and responds by lowering price. A detailed discussion of rate-of-return regulation is provided in chapter 12.
Control of Quantity

Restrictions on the quantity of a product or service that is sold may be used either with or without price regulation. From the 1930s up until around 1970, many oil-producing states, among them Texas and Oklahoma, placed maximum production limits on crude oil producers. Although quantity was controlled by the state, price was determined nationally or globally (though obviously these quantity controls influenced the market price). Alternatively, a common form of quantity regulation that is often imposed on a common carrier is that it “meet all demand at the regulated price.” This requirement is used in regulating electric utilities. Finally, regulation may place restrictions on the prices that firms set while leaving their quantity decisions unregulated. Natural gas prices had been regulated, and because these regulated prices were set below their market-clearing levels and firms were not required to meet all demand, the obvious implication was shortages.

Control of Entry and Exit

As we will see in our studies of economic regulation, the two critical variables that regulators have controlled are price and the number of firms, the latter through restrictions on entry and exit. These variables are critical because price and the number of firms are key determinants of both allocative and productive efficiency.

Entry may be regulated at several levels. First, entry by new firms may be controlled, as is typically done in the regulation of public utilities. A key step toward deregulating the intercity telecommunications market was the FCC’s allowing MCI to enter in 1969. MCI was the first entrant in the market since the industry’s regulation at the turn of the twentieth century.

In addition to controlling entry by new firms, a regulatory agency may also control entry by existing regulated firms. These markets may already be served by other regulated firms or may be unregulated markets. As an example of the latter, the FCC placed restrictions on AT&T’s entry into the computer market in the 1980s. The former case is exemplified by airline and trucking regulation. Their respective regulatory agencies made it very difficult for an existing firm to enter a geographic market already served by another regulated firm. As a more recent example of entry restrictions, the Telecommunications Act of 1996 specified that a regional Bell operating company is not permitted to offer long-distance telephone service to its local telephone customers until its local telephone market is deemed sufficiently competitive by the FCC.

A basis for exit regulation is that regulation strives to have services provided to a wider set of consumers than would be true in a free market. Attaining this goal may entail regulated firms serving unprofitable markets, which creates a need for regulations that forbid a regulated firm from abandoning a market without regulatory approval. As we will see, restricting the decision to exit was an important issue in the regulation of the railroad industry.
Our concern is primarily with regulation that significantly limits or outright prohibits entry and that is used in conjunction with price regulation. Though such regulation has been present in some key industries in the United States, most industries are not and have never been subject to such draconian limitations. Still, there are milder regulatory practices that have constrained entry in many industries. For example, state occupational licensing is sufficiently widespread that it was used in over 500 occupations and affected 18 percent of the U.S. workforce in the late 1970s. It is not at all clear whether these licensing requirements produce better service or whether they just serve to protect the income of incumbents by erecting obstacles to entry. But even more ubiquitous is the procedures and fees required of an entrepreneur to start a new business. While the intent of this government intervention is not always clear, its cost can be substantial when it is aggregated over all new businesses.

To gauge both the size and variation of these entry costs, a recent study collected data from 85 countries. Ranked according to the (official) minimum number of days it takes to complete the necessary procedures, the best 12 countries and the worst 12 countries are reported in table 10.1. The table also reports the number of procedures a new business must complete, the fees that must be paid, and the total cost (which is the sum of fees and the value of the entrepreneur’s time). The latter two variables are aggregated over all new businesses and divided by gross domestic product (GDP) in order to get a measure of the fraction of society’s resources that are consumed by this activity. Averaging across all eighty-five countries, it takes a minimum of forty-seven days at a cost of about two-thirds of 1 percent of GDP. But in some countries these costs are quite substantial, as it can take at least several months to complete the process and, in doing so, use up 3 or even 4 percent of GDP. It is also striking how much these costs vary across countries. Whereas it may only take four days and consume about one-sixth of 1 percent of GDP in the United States, it takes 149 days and 1.7 percent of GDP in Mozambique and eighty days and almost 5 percent of GDP in the Dominican Republic. Given the sizable costs involved, it is important to understand the purpose of these regulations and whether they are to improve the well-being of the many or enrich the few.

Control of Other Variables

The essence of economic regulation is the limitation of firm behavior regarding price, quantity, and entry into and exit out of markets. Obviously, firms choose many other decision variables. One of these is the quality of the product or service that they produce. A regulatory agency may specify minimum standards for reliability of a service. If an electric utility has regular blackouts, the regulatory agency is likely to intervene and require an increase in capacity in order to improve service reliability. Although product quality may also be controlled

2. The numbers for the late 1970s are quoted in Morris M. Kleiner and Robert T. Kurlde, “Does Regulation Affect Economic Outcomes? The Case of Dentistry,” *Journal of Law and Economics* 43 (October 2000): 547–82. They find that tougher occupational licensing does not result in fewer cavities but does enhance the income of dentists.
for reasons like product safety, economic regulation does not typically place serious restrictions on it.

One reason for the minimal use of quality regulation is the cost of implementing it. To control any variable, the relevant economic agents have to be able to agree on what the variable is and what restrictions are placed on it. In the case of price and quantity, this is not difficult. The price is the amount paid by the consumer for the good, which is relatively easy to observe. Furthermore, restrictions take the simple form of numbers: a maximum price and a minimum price. Similarly, the measurability of quantity allows a regulatory agency to specify restrictions on it. However, quality is typically neither so well defined nor so easily observed. For example, the quality of airline service encompasses an array of variables, including

Table 10.1
Cost and Time of Government Requirements for Entry: Best Twelve and Worst Twelve Countries

<table>
<thead>
<tr>
<th>Rank</th>
<th>Country</th>
<th>Time</th>
<th>Number of Procedures</th>
<th>Cost</th>
<th>Cost + Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Australia</td>
<td>2</td>
<td>2</td>
<td>0.0225</td>
<td>0.0305</td>
</tr>
<tr>
<td>1</td>
<td>Canada</td>
<td>2</td>
<td>2</td>
<td>0.0145</td>
<td>0.0225</td>
</tr>
<tr>
<td>3</td>
<td>Denmark</td>
<td>3</td>
<td>3</td>
<td>0.1000</td>
<td>0.1120</td>
</tr>
<tr>
<td>3</td>
<td>New Zealand</td>
<td>3</td>
<td>3</td>
<td>0.0053</td>
<td>0.0173</td>
</tr>
<tr>
<td>5</td>
<td>United Kingdom</td>
<td>4</td>
<td>5</td>
<td>0.0143</td>
<td>0.0303</td>
</tr>
<tr>
<td>5</td>
<td>United States</td>
<td>4</td>
<td>4</td>
<td>0.0049</td>
<td>0.0169</td>
</tr>
<tr>
<td>7</td>
<td>Sweden</td>
<td>13</td>
<td>6</td>
<td>0.0256</td>
<td>0.0776</td>
</tr>
<tr>
<td>8</td>
<td>Hong Kong</td>
<td>15</td>
<td>5</td>
<td>0.0333</td>
<td>0.0933</td>
</tr>
<tr>
<td>8</td>
<td>Panama</td>
<td>15</td>
<td>7</td>
<td>0.3074</td>
<td>0.3674</td>
</tr>
<tr>
<td>10</td>
<td>Ireland</td>
<td>16</td>
<td>3</td>
<td>0.1157</td>
<td>0.1797</td>
</tr>
<tr>
<td>10</td>
<td>Switzerland</td>
<td>16</td>
<td>7</td>
<td>0.1724</td>
<td>0.2364</td>
</tr>
<tr>
<td>12</td>
<td>Norway</td>
<td>18</td>
<td>4</td>
<td>0.0472</td>
<td>0.1192</td>
</tr>
<tr>
<td>74</td>
<td>Dominican Republic</td>
<td>80</td>
<td>21</td>
<td>4.6309</td>
<td>4.9509</td>
</tr>
<tr>
<td>75</td>
<td>Spain</td>
<td>82</td>
<td>11</td>
<td>0.1730</td>
<td>0.5010</td>
</tr>
<tr>
<td>76</td>
<td>Peru</td>
<td>83</td>
<td>8</td>
<td>0.1986</td>
<td>0.5306</td>
</tr>
<tr>
<td>77</td>
<td>Bolivia</td>
<td>88</td>
<td>20</td>
<td>2.6558</td>
<td>3.0078</td>
</tr>
<tr>
<td>78</td>
<td>Slovak Republic</td>
<td>89</td>
<td>12</td>
<td>0.1452</td>
<td>0.5012</td>
</tr>
<tr>
<td>79</td>
<td>China</td>
<td>92</td>
<td>12</td>
<td>0.1417</td>
<td>0.5097</td>
</tr>
<tr>
<td>80</td>
<td>Romania</td>
<td>97</td>
<td>16</td>
<td>0.1531</td>
<td>0.5411</td>
</tr>
<tr>
<td>81</td>
<td>Venezuela</td>
<td>104</td>
<td>14</td>
<td>0.1060</td>
<td>0.5520</td>
</tr>
<tr>
<td>82</td>
<td>Vietnam</td>
<td>112</td>
<td>16</td>
<td>1.3377</td>
<td>1.7857</td>
</tr>
<tr>
<td>83</td>
<td>Indonesia</td>
<td>128</td>
<td>11</td>
<td>0.5379</td>
<td>1.0499</td>
</tr>
<tr>
<td>84</td>
<td>Mozambique</td>
<td>149</td>
<td>19</td>
<td>1.1146</td>
<td>1.7106</td>
</tr>
<tr>
<td>85</td>
<td>Madagascar</td>
<td>152</td>
<td>17</td>
<td>0.4263</td>
<td>1.0343</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>47</td>
<td>10.48</td>
<td>0.4708</td>
<td>0.6598</td>
</tr>
</tbody>
</table>

Note: Time is measured in the (official) minimum number of days), Cost is measured by expenses as a fraction of GDP in 1999, and Cost + Time is measured by Time plus the monetized value of the entrepreneur’s time (as a fraction of GDP).

on-time performance, safety, on-board services, seat width, and luggage handling. In principle, a regulatory agency could attempt to control each of these variables and thus control quality, but it would be very costly to do so. In the case of airline regulation, these variables were not controlled except for minimal standards on safety. As a result, airlines competed vigorously in terms of quality. Generally, economic regulation has not placed severe restrictions on the quality of products or services that firms offer, with the notable exception of product safety.

Restrictions on advertising have been quite common for certain products and services. Bans on advertising of “undesirable” goods—cigarettes and alcohol, in particular—are well known. Tobacco companies have been prohibited from advertising on television and radio for decades, and some states do not allow alcohol to be advertised on billboards. This is a form of social regulation in that the intent is to discourage the consumption of goods with externalities or for which consumers are thought to be ill-informed. The Federal Trade Commission also enforces rules on the content of advertising so as to prevent consumer deception. But these various arguments cannot explain why firms are prevented from advertising the prices of alcohol (when alcohol itself can be advertised), prescription drugs, eyeglasses (see chapter 16), optometry services, or legal services. More often than not, the evidence suggests that these prohibitions do not enhance welfare and, not surprisingly, result in higher prices.³

Another variable that is sometimes (though infrequently) regulated is firm investment. In contrast to the other decision variables we have considered, regulation of investment entails government intervention in the production process, that is, a firm’s choice of technology and inputs. A regulatory agency may intervene in the capital decisions of a public utility like an electric utility or a local telephone company. One significant example is state regulation of investment decisions by hospitals. Certificate of Need programs require a hospital to obtain state approval before undertaking certain investment projects. The stated objective is to avoid duplicate facilities.

**Brief History of Economic Regulation**

**Formative Stages**

What is typically meant by economic regulation in the United States began in the 1870s.⁴ Two important events took place around that time. First, a key Supreme Court decision

---

³. For a review of some of the evidence, see Kyle Bagwell, “The Economic Analysis of Advertising” Department of Economics (Columbia University, March 2003, pdf).

provided the basis for the regulation of monopolies. Second, forces were building in the railroad industry that would result in its being the first major industry subject to economic regulation at the federal level.

**Munn v. Illinois (1877)**

In 1877 the landmark case of *Munn v. Illinois* was decided. This case established that the state of Illinois could regulate rates set by grain elevators and warehouses. As stated in the opinion of the majority, the important principle promulgated by this decision was that property does become clothed with public interest when used in a manner to make it of public consequence, and affect the community at large. When, therefore, one devotes his property to a use in which the public has an interest, he, in effect, grants to the public an interest in that use, and must submit to be controlled by the public for the common good.

*Munn v. Illinois* provided the foundation for regulation to be used to prevent monopolistic exploitation of consumers.

**Interstate Commerce Act of 1887**

Around the time of the *Munn v. Illinois* decision, the railroad industry was going through a turbulent period. Throughout the 1870s and 1880s the railroad industry was subject to spurts of aggressive price wars intermixed with periods of relatively stable prices (see figure 5.9 in chapter 5). At the same time, the railroads were practicing price discrimination across different consumers. Those consumers who were charged relatively high prices (because of relatively inelastic demand) were calling for government intervention. At the same time, the railroads were seeking government assistance to stabilize prices (perhaps near the monopoly level). The result of these forces was the *Interstate Commerce Act of 1887*, which created the Interstate Commerce Commission (ICC) for the purpose of regulating rail rates. Although only with later acts of Congress was the ICC given the necessary powers to regulate price, the Interstate Commerce Act represents an important landmark in congressional regulatory legislation.

**Nebbia v. New York (1934)**

A common interpretation of *Munn v. Illinois* was that it was constitutional for government to regulate certain monopolistic industries. A stricter interpretation was that regulation could only be applied to public utilities. However, in its 1934 decision of *Nebbia v. New York*, the Supreme Court outlined a much wider realm for economic regulation. In that case, the state of New York was regulating the retail price of milk. The defense argued that the milk industry was competitive and could not be classified as a public utility, and so there was no basis for state regulation. The majority opinion stated:
So far as the requirement of due process is concerned, and in the absence of other constitutional restriction, a state is free to adopt whatever economic policy may reasonably be deemed to promote public welfare, and to enforce that policy by legislation adapted to its purpose.

The Supreme Court tore down any constitutional barrier to economic regulation as long as, in the state’s judgment, such regulation was in the public interest.

**Trends in Regulation**

Early regulation focused on the railroads and public utilities like electricity, telephone (which encompassed both local telephone and long-distance communications), and city transit. The Massachusetts State Commission began regulating such industries in 1885, but not until the period of 1907–1930 did most state legislatures create public service commissions. In addition to federal regulation of railroads officially dating from 1887, regulation over interstate telephone service came with the Mann-Elkins Act of 1910.

Figure 10.1 depicts the growth of regulatory legislation. Three spurts of legislative activity can be identified. The first two occurred during the periods of 1909–1916 and 1933–1940. During these years, and up through the 1970s, federal regulatory powers were greatly expanded to encompass a large number of vital industries in the United States. The third burst of legislative activity began in the 1970s and entailed the partial or full deregulation of many of the regulated industries. This trend continues to this day.

The economic historian Richard Vietor has put forth the intriguing hypothesis that these regulatory and deregulatory booms are due to a fundamental change in people’s perception of how an economy and its government interact. He attributes the regulatory wave of the 1930s to the downfall of faith in a laissez-faire economy emanating from the Great Depression. The deregulatory period of the 1970s occurred during a period of serious stagflation—high inflation and high unemployment—that, Vietor argues, shook our faith in the ability of the government to provide a constructive influence on the economy. Though this hypothesis is speculative and has not been tested (nor is it clear how one could test it), it is both interesting and plausible.

**1930s: Wave of Regulation**

After the *Nebbia v. New York* decision, and in the midst of the dire economic conditions of the Great Depression, a wave of economic regulation took place over 1933–1940. At the state level, control over the production of crude oil producers was being implemented by oil-

---


producing states. At the federal level, several pieces of major legislation were adopted that greatly expanded the realm of economic regulation.

A list of these legislative acts is provided in table 10.2. With legislative acts in 1935 and 1940, the ICC’s domain expanded from railroads to the entire interstate surface freight transportation industry, which included trucks, water barges, and oil pipelines (the last goes back to 1906). The one key exception was ocean shipping, which was regulated by the Federal Maritime Commission beginning in 1936. Regulation of long-distance passenger transportation was divided between the ICC (railroads and buses) and the newly created Civil Aeronautics Board (airlines).

To deal with the technologically progressive communications market, the Federal Communications Commission was established in 1934 to regulate broadcasting and to take over the duty of regulating the intercity telecommunications market from the ICC. Although electricity and natural gas had long been regulated at the state and local level, federal regulation of interstate commerce with respect to these two energy sources was only established in 1935 (for electricity) and in 1938 (for natural gas). Initially, natural gas regulation only covered its transportation. Regulation of natural gas prices did not take place until the mid-1950s.
The unsatisfactory performance of financial markets in the Great Depression was followed by a wave of federal legislation relating to the banking and securities industries. Among other restrictions, the Banking Acts of 1933 and 1935 created the Federal Deposit Insurance Corporation, forbade commercial banks from paying interest on ordinary checking accounts, and, in what has been referred to as the Glass-Steagall Act, prohibited both commercial banks from participating in investment banking and investment banks from accepting deposits. The Securities Act of 1933 mandated disclosure of information by issuers of securities, and the Securities Exchange Act of 1934 created the Securities and Exchange Commission, the main purpose of which was to monitor the activities of the securities industry.

### 1940s to 1960s: Continued Growth of Regulation

Between the two legislative peaks of the 1930s and 1970s, legislative activity continued on a modest but steady path of expansion of federal regulatory powers. Two sectors, energy and communications, were particularly affected. Although cable television was initially left unregulated at the federal level, it became subject to FCC regulation beginning in 1968. Until 1954 federal regulation of the oil and natural gas industries was only over pipelines and, at the state level, over production of crude oil. Because of a Supreme Court decision in 1954, the Federal Power Commission began controlling the wellhead price of natural gas. Then the price of oil was regulated beginning in 1971. Foreshadowing the deregulation that was to
come, the FCC permitted MCI to enter the intercity telecommunications market in 1969. This action represented a crucial first step in the deregulation of that market.

1970s to 1980s: Wave of Deregulation

The decades of the 1970s and 1980s were characterized by extensive deregulation (see table 10.3). In 1977, fully regulated industries produced 17 percent of the U.S. gross national product. By 1988 this figure had been reduced to 6.6 percent. In the area of transportation, several pieces of legislation over 1978–1982 deregulated airlines (Airline Deregulation Act of 1978), railroads (Staggers Act of 1980), trucking (Motor Carrier Act of 1980), and passenger buses (Bus Regulatory Reform Act of 1982). In communications, entry regulation of the intercity telecommunications market was torn down over the course of several decisions that ranged from the FCC’s Specialized Common Carrier Decision in 1971 to the breakup of AT&T in 1984 as a result of the U.S. Department of Justice’s antitrust case. Also during this period, cable television was deregulated at the federal level. Finally, oil price controls were lifted by President Ronald Reagan in January 1981, partial deregulation of natural gas prices having begun in 1978. Only in 1989 were natural gas price controls removed.

Regulatory Policy in the 1990s

As is evident from table 10.4, the deregulatory wave that began in the 1970s largely continued into the 1990s. In recent years the deregulation of interstate and intrastate trucking was completed. Loosened regulatory controls allowed competition to exert itself in the transmission of natural gas and in the generation and distribution of bulk power. Entry restrictions in banking, which prevented banks from having branches in more than one state and, in some states, prevented a bank from having more than one branch, were largely eliminated by state legislatures. Contrary to this deregulatory trend, cable television rates have oscillated between being regulated and deregulated, and the landmark Telecommunications Act of 1996 is considered to be a mixture of regulation and deregulation. With regard to the future, the primary architect of airline deregulation, Alfred Kahn, sees us at a point of no return:

The evolution of regulatory policy will never come to an end. The path it takes—and we should make every effort to see that it takes—however, is the path not of a full circle or pendulum, which would take us back to where we started, but of a spiral, which has a direction. This is in a sense only an expression of a preference for seeking consistently to move in the direction of first-best functioning of a market economy, rather than the second or third-best world of centralized command and control.8


Table 10.3
Major Economic Deregulatory Initiatives, 1971–1989

<table>
<thead>
<tr>
<th>Year</th>
<th>Initiative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1971</td>
<td>Specialized Common Carrier Decision (FCC)</td>
</tr>
<tr>
<td>1972</td>
<td>Domestic satellite open skies policy (FCC)</td>
</tr>
<tr>
<td>1975</td>
<td>Abolition of fixed brokerage fees (SEC)</td>
</tr>
<tr>
<td>1976</td>
<td>Railroad Revitalization and Reform Act</td>
</tr>
<tr>
<td>1977</td>
<td>Air Cargo Deregulation Act</td>
</tr>
<tr>
<td>1978</td>
<td>Airline Deregulation Act</td>
</tr>
<tr>
<td></td>
<td>Natural Gas Policy Act</td>
</tr>
<tr>
<td>1979</td>
<td>Deregulation of satellite earth stations (FCC)</td>
</tr>
<tr>
<td></td>
<td>Urgent-mail exemption (Postal Service)</td>
</tr>
<tr>
<td>1980</td>
<td>Motor Carrier Reform Act</td>
</tr>
<tr>
<td></td>
<td>Household Goods Transportation Act</td>
</tr>
<tr>
<td></td>
<td>Staggers Rail Act</td>
</tr>
<tr>
<td></td>
<td>Depository Institutions Deregulation and Monetary Control Act</td>
</tr>
<tr>
<td></td>
<td>International Air Transportation Competition Act</td>
</tr>
<tr>
<td></td>
<td>Deregulation of cable television (FCC)</td>
</tr>
<tr>
<td></td>
<td>Deregulation of customer premises equipment and enhanced services (FCC)</td>
</tr>
<tr>
<td>1981</td>
<td>Decontrol of crude oil and refined petroleum products (executive order)</td>
</tr>
<tr>
<td></td>
<td>Deregulation of radio (FCC)</td>
</tr>
<tr>
<td>1982</td>
<td>Bus Regulatory Reform Act</td>
</tr>
<tr>
<td></td>
<td>Garn–St. Germain Depository Institutions Act</td>
</tr>
<tr>
<td></td>
<td>AT&amp;T settlement</td>
</tr>
<tr>
<td>1984</td>
<td>Space commercialization</td>
</tr>
<tr>
<td></td>
<td>Cable Television Deregulation Act</td>
</tr>
<tr>
<td></td>
<td>Shipping Act</td>
</tr>
<tr>
<td>1986</td>
<td>Trading of airport landing rights</td>
</tr>
<tr>
<td>1987</td>
<td>Sale of Conrail</td>
</tr>
<tr>
<td></td>
<td>Elimination of fairness doctrine (FCC)</td>
</tr>
<tr>
<td>1988</td>
<td>Proposed rules on natural gas and electricity (FERC)</td>
</tr>
<tr>
<td></td>
<td>Proposed rules on price caps (FCC)</td>
</tr>
<tr>
<td>1989</td>
<td>Natural Gas Wellhead Decontrol Act of 1989</td>
</tr>
</tbody>
</table>

The Regulatory Process

Overview of the Regulatory Process

Stage 1: Legislation

There are two key stages in the regulation of an industry. The first stage is for the U.S. Congress, a state legislature, or a local government body such as a city council to enact a piece of legislation that establishes regulatory powers over a particular industry. Numerous agents are involved at this stage of the regulatory process. Because regulation restricts firm decisions, it is expected to influence firms’ profits and consumers’ welfare. Hence, one would anticipate that both firms and consumer advocates would lobby the government to try to influence what the piece of legislation looks like, as well as whether or not it passes. Obviously, legislators are key actors during this stage. Depending on their jurisdiction, legislators may represent producers, consumers, or just their electorate at large. Because industry workers are likely to be affected by legislation, one can also expect them to be involved in this process, particularly if workers are organized into a labor union.

Stage 2: Implementation

Once a piece of legislation is passed, the second stage in the regulatory process is the implementation of this legislation. Although the legislature can influence its implementation, the immediate responsibility falls to the regulatory agency. Thus, regulators replace legislators as central actors at the implementation stage, while producers and consumers continue to be relevant. Other important actors may include potential entrants who desire to enter this regulated industry.

Stage 3: Deregulation

There is sometimes a third stage in this process, which is the deregulation of the industry. Although one typically imagines deregulation being achieved via a legislative act, both the regulatory agency and the judiciary have proved to be instrumental forces in deregulating an industry. If the regulatory agency and the judiciary are in favor of deregulation, they may be able to achieve it even if Congress is against it. Long before Congress passed the Airline Deregulation Act, the airline industry was being deregulated by the Civil Aeronautics Board. In this respect, the White House can play a significant role in its choice of regulatory commissioners. Indeed, it was no mistake that President Jimmy Carter appointed the free-market advocate Alfred Kahn as CAB chairman.

Due to its role in deregulating a number of industries, the Circuit Court of Appeals for the District of Columbia has been dubbed the “Supreme Court” for regulations. In addition to the three branches of government, all agents significantly connected with the industry are
Table 10.4

<table>
<thead>
<tr>
<th>Year</th>
<th>Initiative</th>
<th>Provisions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>Federal Deposit Insurance Corporation Improvement Act</td>
<td>Introduced risk-based deposit insurance premia, required early regulatory intervention into failing banks, eased conditions for banking failures by limiting FDIC's ability to reimburse uninsured depositors.</td>
</tr>
<tr>
<td>1992</td>
<td>Energy Policy Act</td>
<td>Opened up wholesale competition by giving FERC the authority to order vertically integrated utilities to act as a common carrier of electrical power.</td>
</tr>
<tr>
<td>1992</td>
<td>FERC Order 636</td>
<td>Required pipelines to unbundle the sale and transportation of natural gas.</td>
</tr>
<tr>
<td>1993</td>
<td>Omnibus Budget Reconciliation Act of 1993</td>
<td>Mandated that the FCC reallocate portions of the electromagnetic spectrum for personal communication and authorized it to use competitive bidding to award licenses. Deregulated cellular telephone rates.</td>
</tr>
<tr>
<td>1993</td>
<td>Negotiated Rates Act</td>
<td>Eliminated regulatory distortions related to trucking rates.</td>
</tr>
<tr>
<td>1994</td>
<td>Riegle-Neal Interstate Banking and Branching Efficiency Act</td>
<td>Codified at the national level the elimination of branching restrictions at the state level.</td>
</tr>
<tr>
<td>1994</td>
<td>Trucking Industry and Regulatory Reform Act</td>
<td>Eliminated remaining interstate and intrastate trucking regulation.</td>
</tr>
<tr>
<td>1995</td>
<td>ICC Termination Act</td>
<td>Abolished the ICC.</td>
</tr>
<tr>
<td>1996</td>
<td>Telecommunications Act</td>
<td>Deregulated cable TV rates, set conditions for local telephone companies to enter long-distance telephone markets, mandated equal access to local telephone systems.</td>
</tr>
<tr>
<td>1996</td>
<td>FERC Order 888</td>
<td>Removed impediments to competition in the wholesale bulk power market.</td>
</tr>
<tr>
<td>1999</td>
<td>Gramm-Leach-Bliley Act</td>
<td>Repealed the prohibition on mixing banking with securities or insurance that had been in place with the Glass-Steagall Act.</td>
</tr>
<tr>
<td>1999</td>
<td>FERC Order 2000</td>
<td>Advocated establishment of independent regional transmission organizations to facilitate competition in wholesale electricity markets.</td>
</tr>
<tr>
<td>2002</td>
<td>Sarbanes-Oxley Act</td>
<td>Instituted new regulations concerning financial practices, corporate governance, and corporate disclosure.</td>
</tr>
</tbody>
</table>

Acknowledgments: The development of this table was aided by suggestions from Timothy Brennan, Robert Crandall, Paul Kleindorfer, Randall Kroszner, Paul MacAvoy, Thomas Gale Moore, and Sam Peltzman. Their assistance is most appreciated. They are not responsible for any errors.
typically involved in the deregulatory process, including producers, consumers, labor, and prospective firms.

**Regulatory Legislation**

**Selection of the Regulatory Agency**

Legislation performs two key tasks in the regulatory process. First, it states which bureaucratic agency has jurisdiction over regulating certain dimensions of an industry. In many cases, such as the Interstate Commerce Act of 1887 and the Federal Communications Act of 1934, legislation actually creates the bureaucratic agency. In other cases, legislation extends the domain of an existing agency, as the Motor Carrier Act of 1935 did in bringing motor carriers within the realm of the ICC.

**Powers of the Regulatory Agency**

The second objective of legislation is to outline the powers of the regulatory agency. The two key powers are control of price and control of entry into and exit from the industry. Although the Interstate Commerce Act of 1887 gave the ICC regulatory jurisdiction over the railroad industry, it took the Hepburn Act of 1906 and the Transportation Act of 1920 for the ICC to have the power to control rail rates. Sometimes it is unclear as to the powers that a piece of legislation gives a regulatory agency. Until a 1954 Supreme Court decision, the Federal Power Commission believed that the Natural Gas Act of 1938 did not give it the power to control the wellhead price of natural gas.

**General Policy Objectives**

Finally, regulatory legislation often specifies some general policy objectives for the regulatory agency to follow. In most cases, legislation instructs the regulatory agency to set “reasonable and just” prices and to see that service is made available to all consumers. Thus, the FCC and the CAB sought to expand long-distance communications and airline service, respectively, to as wide a geographic area as was possible. Another common policy goal is to discourage regulated firms from practicing price discrimination.

**Independent Regulatory Commissions**

An independent regulatory commission at the federal level is typically composed of five or more members. Table 10.5 provides a listing of some major regulatory agencies at the federal level. Federal regulatory commissioners are appointed, though in some states public utility commissioners are elected.\(^9\) The appointment is for a fixed term, and the terms of the com-

---

missioners are staggered. There is an important degree of independence from the executive branch bestowed on regulatory commissioners. A commissioner can be removed for cause, but not at the discretion of the president.

In light of the lack of political accountability of regulatory commissioners, it has been argued that they are set up in the manner of judges. In particular, Section 557 of the 1946 Administrative Procedure Act requires all administrative decisions by a regulatory commission to be substantiated by findings of fact and law.

10. Stone, Regulation.
Members of a Regulatory Agency

Political scientist James Q. Wilson has identified three different kinds of employees of a regulatory agency.11 The careerist is an employee who anticipates a long-term relationship with the regulatory agency and whose major concern is that the regulatory agency continue to exist and grow. Not surprisingly, the careerist frowns on deregulation. The politician envisions eventually leaving the agency for an elective or appointive position; the regulatory agency is then a stepping stone to bigger and better things. Most commissioners are classified as politicians. Finally, the professional is more identified with certain skills than with the regulatory agency and strives to maintain professional esteem to allow career advancement.

The incentives of an employee of a regulatory agency depend very much on the type of employee. Understanding how members of a regulatory agency are motivated is important in explaining the policies that are implemented. For example, consider the implementation of price regulation. The professional may desire to use this opportunity to show technical expertise, and therefore is apt to prefer a highly complex pricing structure. In contrast, the careerist might support a simple pricing structure so as to avoid any major problems that might result in legislative action. Finally, because the politician is concerned with not aggravating interest groups, he would be less inclined to allow price discrimination because it might alienate some consumers. Our ensuing analysis of regulation will not allow for such a rich set of motivations underlying the implementation of regulatory policy, but it is important that we at least recognize their presence.

Regulatory Procedures

Given the general and vague policy objectives provided by legislation, a regulatory agency is often left with considerable discretion as to how it regulates the industry. When a regulatory agency is told to set “reasonable and just” rates, there may be a wide array of rates that one could argue meet these criteria. Alternatively, some legislation is very specific about the duties of a regulatory agency. The Emergency Petroleum Allocation Act (1973–1975) and the Energy Policy and Conservation Act (1975–1981) provided a detailed formula as to the price structure for domestic crude oil. As a result, the Federal Energy Administration had minimal discretion over the regulation of crude-oil prices.

Rulemaking Process

Two basic approaches to rulemaking have been pursued. First, a regulatory agency may act on a case-by-case approach by considering each proposal individually. The most important

proposals concern rate changes and petitions for entry or exit. When the burden of a case-by-case approach becomes too great, a regulatory agency will often turn to substantive rule-making. Hearings are conducted that lead to the formulation of a general rule that is applicable to a wide class of situations. The move to substantive rulemaking from a burdensome case-by-case approach was made by the FPC in regulating natural gas prices (the first case took five years to complete) and by the FCC in deciding on entry into a segment of the intercity telecommunications market (the first case took six years).

If the participants do not agree with the decision of a regulatory decision, they have the right to appeal it in a U.S. court of appeals. This tactic has indeed been used. When the FCC told MCI that it was not allowed to operate in the long-distance telephone service segment of the intercity telecommunications market, MCI went to the U.S. court of appeals, where, in the Execunet I decision (1978), the courts reversed the FCC’s decision. Two years after its 1992 ruling that local telephone companies must allow competitors direct access to the local phone network, the FCC found its decision overturned by a federal appeals court.

**Delay and Strategic Manipulation of Regulatory Proceedings**

An important property of regulatory procedures is that they are biased toward maintaining the status quo. By replacing market forces with administrative processes, regulation imposes due process requirements on any changes. In some sense, producers and consumers have legal rights to the status quo, and it can only be changed through due process. This situation is very much in contrast to the market, where the status quo is regularly overthrown and there is no legal recourse as long as no laws are violated.

Another property of regulation that favors the status quo is the extent of delay in regulatory proceedings. An agent who is interested in maintaining the status quo can pursue tactics such as litigation in order to lengthen the proceedings. Regardless of the reason for delay, its existence is hard to deny. For the CAB and the ICC, licensing proceedings averaged 170 days for the prehearing stage, 190 days for the hearing stage, and 220 days for the agency review stage. The total length of time was in excess of nineteen months. Ratemaking proceedings were even worse, as, on average, ratemaking cases by the CAB, FMC, FPC, and the ICC took over twenty-one months. However, as mentioned earlier, a regulatory agency can reduce delay by replacing a case-by-case approach with substantive rulemaking.

In addition to generating delay in regulatory proceedings, agents can strategically manipulate the regulatory process in other ways. An important avenue for regulated firms is to control the flow of information to the regulators. For example, in considering a ratemaking

---

In case, a regulatory agency usually depends on the regulated firm for estimates of cost and demand conditions. Although outside expert witnesses can be used, their information is simply not as good as that which the firm has at its disposal. Another tactic is for regulated firms to co-opt the experts, for example, by keeping the best law firms on retainer.13

The Theory of Regulation

Why is there regulation? In a free-market economy like that of the United States, why does the government choose to place restrictions on the decisions of agents? One of the objectives of a theory of regulation is to answer this question. Such a theory should make predictions concerning who benefits from regulation, which industries are most likely to be regulated, and what form regulation will take.14 A proper addressing of these issues should allow us better to understand the effects of regulation. For example, if we know that there is a general tendency for price regulation to benefit producers, it is logical to expect price to be set significantly above cost in regulated industries.

In this section, we will outline the evolution of thought that addresses the question, Why is there regulation? There have been three stages in this evolution. The first hypothesis put forth was that regulation occurs in industries plagued with market failures. Originally called the public interest theory, more recently it has been referred to as normative analysis as a positive theory (NPT).15 Largely due to empirical evidence that was inconsistent with NPT, economists and political scientists developed the capture theory (CT). Basically, the CT states that, whether by design or not, the agency that is meant to regulate an industry is “captured” by that industry. The implication is that regulation promotes industry profit rather than social welfare. For reasons described later, NPT and CT are actually not theories but rather hypotheses or statements about empirical regularities. This is to be contrasted to the third stage in this evolution of thought, which is the economic theory of regulation (ET). This is indeed a theory in the proper sense, in that it generates testable hypotheses as logical implications from a set of assumptions. Although ET is an important advancement and explains some of the observed regulatory activity in the United States over the last hundred years, much evidence is still inconsistent with this theory.


Normative Analysis as a Positive Theory

Normative Rationale for Regulation

There is a basis for government intervention in that under certain conditions, unrestrained competition does not work very well. Two common circumstances are that an industry is a natural monopoly or that it is plagued by externalities.

A market is a natural monopoly if, at the socially optimal quantity, industry cost is minimized by having only one firm produce. For the single-product case, if the average cost curve is declining for all quantities, then the cost of producing any industry quantity is minimized by having one firm produce it. In that case, the market is a natural monopoly regardless of market demand. Natural monopolies are likely to exist when there is a large fixed-cost component to cost. For example, most public utilities, such as local distribution of electricity and local telephone, are natural monopolies. In those cases, fixed costs (in particular, the cost of connecting homes and businesses to the distribution system) are large relative to marginal costs. Hence, average cost is declining for a wide range of outputs. For the relevant region of market demand, these markets are natural monopolies.

The problem with a natural monopoly is that there is a fundamental conflict between allocative efficiency and productive efficiency. Productive efficiency requires that only one firm produce, because only then is the value of resources used to supply the market minimized. However, a lone producing firm will be inclined to set price above cost in its objective of maximizing profit. But then allocative efficiency is not achieved. To generate allocative efficiency, we need enough firms that competition drives price down to marginal cost. But then there is productive inefficiency because there are too many firms producing in the market. Thus we have an argument for government intervention when a market is a natural monopoly.

An externality exists when the actions of one agent, say agent A, affects the utility or production function of another agent, say agent B, and agent A does not care how his behavior affects agent B’s welfare. When an externality is present, perfect competition does not result in an optimal allocation of resources. For example, suppose Jared is considering buying an Italian submarine sandwich for lunch. Let us suppose that the restaurant market is competitive, so that the price of the sandwich equals marginal cost. If the input markets are also competitive, then the value of resources used by society in supplying that sandwich equals the price charged for it, which we will denote as $P$. Now suppose the maximum amount that Jared is willing to pay for that sandwich is $V$. If $V > P$, then Jared will buy the sandwich and receive a surplus of $V - P$. If there are no externalities from him consuming the sandwich, then the net welfare gain to society is $V - P$, which is positive. Thus, such a transaction should (and will) take place. Now let us assume that his consumption of the sandwich generates an externality. In particular, suppose the sandwich has onions on it (as any good Italian sub does) and Jared is planning to travel on a crowded subway after eating it. Unfortunately, the indi-
individual who sits next to him on the subway will have to smell his bad breath. Suppose that this individual would be willing to pay up to \( W \) dollars to get Jared to sit elsewhere (however, there are no seats left on the subway). The net welfare effect of him buying and consuming the Italian sub is not \( V - P \) but instead \((V - P) - W\). If \( W > V - P \), then welfare is actually reduced by Jared’s consumption of the sub, even though he is personally better off. Hence, with the existence of an externality, competitive behavior can result in welfare-reducing transactions.

Externalities come in many forms. The example we have just considered is referred to as a negative externality. Other examples of negative externalities are noise and water pollution. In deciding whether to drive to work or take mass transit, the typical automobile driver does not consider the effect of his decision on the quality of the air that everyone must breathe. A common pool problem is a different type of negative externality. It occurs when there are several property owners to a resource: several firms may extract oil from a common reservoir, and several fishermen may fish from the same lake. In their pursuit of utility or profit maximization, these agents do not take into account how their activity reduces the resource and thus raises the cost of production to other agents.

Generally, when there are negative externalities, unregulated competition results in too much of an activity being pursued, whether it is too many Italian subs being consumed or too much oil being pumped out of a reservoir. There are also cases of positive externalities. For example, if I am immunized against a disease, I not only make myself better off but also reduce the spread of the disease, thereby making others better off. Just as there is typically too much activity when there is a negative externality, there is typically too little activity when there is a positive externality.

When a market failure occurs, whether due to natural monopoly, externalities, or some other source, there is a potential rationale for government intervention. In the case of a natural monopoly, price and entry regulation may allow both allocative and productive efficiency. Entry regulation permits only one firm to produce (as required for productive efficiency), whereas price regulation restricts the firm to setting the socially optimal price (as required for allocative efficiency). In the case of externalities, imposition of a tax (subsidy) on an activity that generates a negative (positive) externality can result in a socially preferred allocation. When there is a market failure, in theory, regulation may be able to raise social welfare. Whether it does so in practice is an altogether different issue and will be of central concern to us in the following chapters.

**Description of Theory**

Understanding when regulation *should* occur is normative analysis. This is to be contrasted with a positive theory that explains when regulation *does* occur. Normative analysis as a positive theory (NPT) uses normative analysis to generate a positive theory by saying that regulation is supplied in response to the public’s demand for the correction of a market
failure or for the correction of highly inequitable practices (for example, price discrimination or firms’ receiving windfall profits as a result of some change in industry conditions). According to this theory, if a market is a natural monopoly, then the public will demand the industry be regulated because a first-best solution is not achieved in the absence of regulation. Unfettered competition will result in either too many firms producing and/or price exceeding the socially optimal level. By regulating the industry, net welfare gains result, and it is this potential for welfare gains that generates the public’s demand for regulation. In this way, the public interest theory uses normative analysis (when should regulation occur?) to produce a positive theory (when does regulation occur?).

Critique of Normative Analysis as a Positive Theory

There are at least two major problems with NPT. First, it is at best a very incomplete theory. NPT puts forth the hypothesis that regulation occurs when it should occur because the potential for a net social welfare gain generates public demand for regulation. Lacking in this analysis is a description of the mechanism that allows the public to bring this result about. Regulation occurs through legislative action and the behavior of the regulatory agency. NPT does not address the issue of how the potential for net social welfare gains induces legislators to pass regulatory legislation and regulators to pursue the proper actions. NPT does not generate the testable prediction that regulation occurs to correct a market failure, but rather assumes it.

The second major criticism of NPT, and the key reason why it has lacked supporters for several decades, is the large amount of evidence that refutes it. Many industries have been regulated for which there is no efficiency rationale; some examples include price and entry regulation in the trucking, taxicab, and securities industries. In 1974, Richard Posner concluded, “Some fifteen years of theoretical and empirical research, conducted mainly by economists, have demonstrated that regulation is not positively correlated with the presence of external economies or diseconomies or with monopolistic market structure.”

Further evidence that is difficult to rectify with NPT is that in many cases, firms supported or even lobbied for regulation. This was true with the regulation of the railroads in the late 1880s and of local and long-distance telephone services, where AT&T supported regulation (and thereby eliminated all other competitors from the market). Though firm support is not necessarily inconsistent with NPT, it does not sit comfortably. If a market is a natural monopoly but there are several active firms, competition could be driving price down below average cost so that firms are incurring losses. Regulation would allow at least one of them to earn normal profits. It is unlikely, however, that firms would be in support of regulation if all it could generate would be normal profits. A more plausible explanation is that regulation is

anticipated to provide a stable level of above-normal profits to be earned, and it is for this reason that an industry may be in favor of its regulation.

A third but weaker line of evidence in conflict with NPT is that the regulation of even a natural monopoly does not always really constrain firm pricing behavior. In a well-known study, George Stigler and Claire Friedland examined the effect of regulation on the pricing of electric utilities over the period 1912–1937.\(^{17}\) They found that regulation had an insignificant, though downward, effect on prices. In contrast, NPT would predict that regulation would have a strong downward effect on prices because it forces a monopolist to price at average cost rather than at the profit-maximizing level.

**Reformulation of NPT**

In light of the contradictory evidence, NPT was reformulated. This reformulation says that regulation is originally put in place to correct a market failure but then is mismanaged by the regulatory agency. However, even this reformulated hypothesis is unsatisfactory. First, it is subject to the same criticism of the original formulation in that it merely states a hypothesis rather than generating that hypothesis as a conclusion from a model. To be specific, it does not explain why the regulatory agency is mismanaged. Second, the reformulated hypothesis is still inconsistent with the evidence of industries being regulated that are not subject to significant market failures. The reformulated hypothesis of NPT does not appear to be a substantive improvement on the original hypothesis.

**Capture Theory**

**Genesis of the Capture Theory**

A review of the history of regulation in the United States since the late nineteenth century reveals that regulation is not strongly correlated with the existence of market failures. At least up to the 1960s, one empirical regularity is that regulation is pro-producer in that it tends to raise industry profit. In potentially competitive industries, such as trucking and taxicabs, regulation supported prices above cost and prevented entry from dissipating above-normal profits (also known as rents). In naturally monopolistic industries like electric utilities, there was some evidence that showed that regulation had little effect on price, so that above-normal profit was allowed to be earned. The empirical evidence seemed to support the claim that regulation was inherently pro-producer.\(^{18}\)

These empirical observations resulted in the development of the *capture theory* (CT). In stark contrast to NPT, CT states that either regulation is supplied in response to the indu-


try’s demand for regulation (in other words, legislators are captured by the industry) or the regulatory agency comes to be controlled by the industry over time (in other words, regulators are captured by the industry).19

**Critique of the Capture Theory**

In that it is in greater agreement with regulatory history, CT is more compelling than NPT. Nevertheless, CT is subject to the same two criticisms leveled against NPT. Like NPT, CT has no theoretical underpinnings because it does not explain how regulation comes to be controlled by the industry. In light of there being several interest groups affected by regulation, including consumer and labor groups as well as firms, why should regulation be controlled by the industry rather than by these other interest groups? In its original form, CT does not provide an explanation. Rather, it merely states the hypothesis that regulation is pro-producer.

Although there is much evidence supportive of CT, there are also some empirical regularities that are inconsistent with it. Two common properties of regulation are cross-subsidization and a bias toward small producers. Although we will go into greater detail in later chapters, *cross-subsidization* is when a multiproduct firm prices some goods below average cost and makes up for the losses through revenue collected from the sale of other goods priced above average cost. Such pricing behavior is inconsistent with profit maximization and thus cannot be considered pro-producer. Cross-subsidization has been regularly observed in such regulated industries as railroads, airlines, and intercity telecommunications. It often takes the form of uniform prices being charged to different consumers, even though the marginal cost of supplying these consumers differs greatly. The other property is that regulation is often biased toward small producers. Small producers are allowed to earn greater profits relative to larger firms under regulation than they would have earned in an unregulated market. This was certainly true of small oil refiners under oil price controls.

Perhaps the strongest evidence against CT is the long list of regulations that were not supported by the industry and have resulted in lower profits. The list includes oil and natural gas price regulation and social regulation over the environment, product safety, and worker safety. Finally, CT has a difficult time explaining both why many industries were regulated *and* why they were later deregulated.

**Economic Theory of Regulation**

In summarizing the evidence, one finds that regulation is not strongly associated with the existence of market failure (in conflict with NPT) and is far from being exclusively pro-producer (in conflict with CT). Depending on the regulated industry, the welfare of different

---

interest groups is improved by regulation. One then needs a theory that can explain this phenomenon. In addition, a theory must also explain why we have observed both the regulation and (partial or full) deregulation of such industries as railroads (regulated in 1887, deregulated in 1980), intercity telecommunications (regulated in 1910, partially deregulated starting in 1971), trucking (regulated in 1935, deregulated in 1980), airlines (regulated in 1938, deregulated in 1978), natural gas (price regulated in 1954, deregulated in 1989), and oil (regulated in 1971, deregulated in 1981). It must also tackle the simultaneous decline of economic regulation and rise of social regulation in the latter part of the twentieth century.

The Stiglerian Approach

The major breakthrough in the theory of regulation occurred in a 1971 article by Nobel laureate George Stigler, “The Theory of Economic Regulation.” The value of this contribution was not so much in the predictions that it generated (it basically produced predictions along the lines of CT), but in the way it approached the question, Why is there regulation? In contrast to NPT and CT, Stigler put forth a set of assumptions and generated predictions about which industries would be regulated and what form regulation would take as logical implications of these assumptions.

The initial premise of Stigler’s analysis is that the basic resource of the state is the power to coerce. An interest group that can convince the state to use its power of coercion to that interest group’s benefit can improve its well-being. The next premise is that agents are rational in the sense of choosing actions that are utility maximizing. These two assumptions result in the hypothesis that regulation is supplied in response to the demands of interest groups acting to maximize their income. Regulation is one avenue by which an interest group can increase its income by having the state redistribute wealth from other parts of society to that interest group. As is typically the case, Stigler states it best:

We assume that political systems are rationally devised and rationally employed, which is to say that they are appropriate instruments for the fulfillment of desires of members of the society.

With this fundamental insight, one can construct a theory that will make predictions as to which industries will be regulated and what form regulation will take. The remainder of the section on the economic theory of regulation (ET) describes some of the formal models under this rubric and describes their resulting predictions.

Stigler/Peltzman Model

Stigler’s contribution did not stop with this analysis. He went on to discuss the different factors that determine which interest group(s) will control regulation. A later paper by Sam
Peltzman formalized the analysis of Stigler, and both of these papers have built on the work of Mancur Olson.22

The Stigler/Peltzman formulation has three crucial elements. First, regulatory legislation redistributes wealth. It may do other things, but implicitly, Stigler and Peltzman argue, the primary determinant of the form of regulation is the way in which it transfers wealth among members of society. Second, the behavior of legislators is driven by their desire to remain in office, implying that legislation is designed to maximize political support. Third, interest groups compete by offering political support in exchange for favorable legislation.

The general result that follows is that regulation is likely to be biased toward benefiting interest groups that are better organized (so that they are more effective at delivering political support) and gain more from favorable legislation (so that they are willing to invest more resources in acquiring political support). More specifically, regulation is likely to benefit small interest groups with strongly felt preferences at the cost of large interest groups with weakly felt preferences.

The reasons lie in recognition and implementation. For an interest group to recognize the need for certain legislation, each member must have the potential of gaining a lot from it. Interest group behavior is driven by the desires of its individual members. It is insufficient for some group potentially to realize a large gain from regulation. What is important is that each of its members stand to gain a lot, for only then does each member have the incentive to invest the resources to learn about the issues and about what needs to be done to achieve favorable legislation. This statement argues to the point that interest groups for whom the per capita benefit from regulation is relatively high are more likely to recognize how legislation can be designed to serve their interests. Of course, it is insufficient simply to recognize a desire for a particular piece of legislation. To benefit, that legislation must be implemented. Implementation requires delivering political support—both in terms of votes and money—to legislators who can see that the appropriate bill is written, proposed, and passed. Here, big groups are at a disadvantage because of a free-rider effect. A person who makes a financial donation on behalf of his interest group benefits everyone in the group, though the cost is specific to him. For example, a union worker who contributes dues of $50 incurs the full cost, but all union members share in the increased political power from the additional $50. This tendency to undercontribute is stronger the larger the group, because the marginal impact of one person’s contribution is smaller, though the cost to that person is independent of the group size. Of course, if everyone acts in that manner, contributions will be quite small. The smaller the size of the interest group, the weaker is this free-rider effect, because each member’s contribution has a proportionately bigger impact on the eventual impact of the group. Thus, in terms of both recognition of a need for regulation and the implementation of that regulation,

the advantage rests in small interest groups, for which the per capita benefits from regulation are high.

This argument provides some insight into why much of observed regulation favors producers. Producer groups are typically small in number, with each firm benefiting a large amount from regulation, whereas the primary opposition is consumers, of which there are typically millions, and the harm that regulation creates, while large in the aggregate, is small for each consumer.

**U.S. Peanut Program**

An example of a small group’s benefiting from regulation at the cost of a large group is the peanut quota system. Since 1949 the federal government has run a program that limits the number of farmers who can sell peanuts in the United States. Imports are also severely restricted. On top of these restrictions, price supports are used to guarantee that farmers with peanut quotas can cover their production costs for each year. This system generally results in the minimum selling price being about 50 percent higher than the world price.

For 1982–1987, it was estimated that the average annual consumer-to-producer transfer was $255 million (in 1987 dollars), with an associated deadweight welfare loss of $34 million. In 1982 there were 23,046 peanut farmers so each received, on average, a net transfer of $11,100. In contrast, the cost to the average consumer of this program was only $1.23. Few consumers would be willing to spend their own time and money to dismantle the peanut program when they would gain only $1.23. However, the program is worth $11,100 to the average peanut farmer, and that gain would certainly make it worth one’s while to see that the program continues.

**Predicting the Type of Industry to Be Regulated**

The key assumption of the Stigler/Peltzman model is that the individuals who control regulatory policy (presumably the legislators) choose policy so as to maximize their political support. Although this is not the only assumption one could make, it is certainly a plausible one, inasmuch as legislators desire to be reelected, and this aim is best achieved by maximizing political support. In deciding on government policies (which could include policies other than price and entry regulation), a legislator decides the size of the group to be benefited by regulation and how much wealth is to be transferred to them. For example, a legislator decides on the price structure and, in so doing, decides which consumers are benefited (their price is set below cost), which consumers are hurt (their price is set above cost), and how much firms are benefited (in terms of the level of profits).

Let us address in greater depth the issue of which industries are most likely to be regulated. For this purpose, Peltzman provides a model specifically designed for price and entry regulation. A legislator/regulator chooses price so as to maximize political support. Let the political support function be represented by $M(P, \pi)$ where $P$ is price and $\pi$ is industry profit. $M(P, \pi)$ is assumed to be decreasing in price because consumers increase their political opposition when price is higher, while it is increasing in industry profit because firms respond with greater support. Profit depends on price where $\pi(P)$ denotes the profit function. In particular, $\pi(P)$ is increasing in $P$ for all prices less than $P^m$ (the monopoly price) and is decreasing in $P$ for all prices above $P^m$. The profit function is shown in figure 10.2. For $P < P^m$, note that if a legislator raises price, he raises consumer opposition, since $M(P, \pi)$ is decreasing in $P$, but also raises industry support, since $\pi(P)$ is increasing in $P$ and $M(P, \pi)$ is increasing in $\pi$.

Let us characterize the price that maximizes the political support function $M(P, \pi)$ subject to $\pi = \pi(P)$. To do so, we have put in figure 10.2 indifference curves for a legislator. The curve $M_1$ represents all pairs of price and profit that generate the level $M_1$ of political support. Note that the slope of an indifference curve is positive, reflecting the fact that if price is higher (hence, consumer support is reduced), then profit must be higher (which raises industry support) if the same level of political support is to be achieved. Because $M(P, \pi)$ is decreasing in $P$, note that if a legislator raises price, he raises consumer opposition, since $M(P, \pi)$ is decreasing in $P$, but also raises industry support, since $\pi(P)$ is increasing in $P$ and $M(P, \pi)$ is increasing in $\pi$. 

Figure 10.2
Optimal Regulatory Policy: Peltzman Model
ing in $P$ and increasing in $\pi$, political support is increasing in a northwesterly direction, so that $M_3 > M_2 > M_1$. The optimal price for the legislator, denoted $P^*$, is that which achieves the highest level of political support subject to the constraint that profit equals $\pi(P)$. Note that $P^*$ lies between the competitive price, $P^c$, where profit is zero, and the monopoly price, $P^m$, where industry profit is maximized. Thus, we have formally derived the result that a legislator/regulator will not set a price so as to maximize industry profit. This is a prediction that differs from the capture theory.

The characterization of the optimal regulated price provides important insight into which industries are likely to gain the most from regulation. If the equilibrium price an industry would achieve in the absence of regulation is close to the price that would exist under regulation, $P^*$, then regulation is unlikely. The interest group that would benefit from regulation will not expect to gain a large amount because price would be relatively unaffected. Hence, it would not warrant the investment of resources to get the industry regulated. Because the regulated price lies in between $P^c$ and $P^m$, this argument also suggests that the industries most likely to be regulated are those that are either relatively competitive (so that the unregulated equilibrium price is near $P^c$) or relatively monopolistic (so that the unregulated equilibrium price is near $P^m$). In both cases, some interest group will gain considerably from regulation. Firms will gain in the case of a competitive industry, while consumers will gain in the case of a monopolistic industry.

Casual observation suggests that it is indeed these two extremes that tend to be subject to economic regulation. Monopolistic industries include local and long-distance telephone service, electric and gas utilities, and railroads. Relatively competitive industries include agriculture (regulation takes the form of price supports), trucking, taxicabs, crude oil and natural gas production, and securities.

**Becker Model**

The Stigler/Peltzman modeling of the economic theory of regulation is based on a legislator or regulator choosing regulatory policy so as to maximize political support. In contrast, the formulation of Gary Becker focuses instead on competition between interest groups. He suppresses the role of the legislator/regulator by assuming that “[p]oliticians, political parties, and voters . . . transmit the pressure of active groups.” True to the economic theory of regulation, Becker assumes regulation is used to increase the welfare of more influential interest groups.

For simplicity, suppose there are two interest groups, denoted group 1 and group 2. An interest group can raise its welfare by influencing regulatory policy. The wealth transfer that

25. Ibid., p. 372.
group 1 gets depends on both the pressure it exerts on legislators and regulators (denoted $p_1$) and the pressure exerted by group 2 (denoted $p_2$). The amount of pressure is determined by the number of members in the group and the amount of resources used. Greater pressure by group 1 as well as less pressure by group 2 implies that group 1 has more influence on the political process. Greater influence translates into group 1 receiving a bigger wealth transfer. In particular, if $T$ is group 1’s increase in wealth due to regulation, then $T = I(p_1, p_2)$, where $I(p_1, p_2)$ is called the influence function. It is assumed that $I(p_1, p_2)$ is increasing in the pressure of group 1 and decreasing in the pressure of group 2. In order to transfer wealth of amount $T$ to group 1, it is assumed that group 2’s wealth must be reduced by $(1 + x)T$, where $x \geq 0$. When $x > 0$, more wealth is taken from group 2 than is transferred to group 1. This “disappearing” wealth is measured by $xT$ and is the welfare loss from regulation.

A property of the Becker model is that aggregate influence is fixed. The implication is that what is important for determining the amount of regulatory activity (as measured by the wealth transfer) is the influence of one group relative to the influence of another group. Each group chooses a level of pressure so as to maximize its welfare given the pressure level chosen by the other group. Because greater pressure uses up the group’s resources, each group will not want to apply too much pressure. On the other hand, the less pressure a group applies, the greater the influence of the other group. Hence, by reducing $p_1$, the relative influence of group 1 declines, so that the wealth transfer it gets will be smaller. Taking into account the benefits and costs of pressure, one can derive the optimal value of $p_1$, given any value for $p_2$. This optimal level of pressure for group 1 is denoted $\psi_1(p_2)$ and is plotted in figure 10.3. $\psi_1(p_2)$ is referred to as group 1’s “best response function” because it tells group 1 what level of pressure is best (in terms of its own welfare) in response to group 2’s level of pressure. For example, if group 2 is expected to apply pressure of $\hat{p}_2$ then group 1’s optimal level of pressure is $\psi_1(\hat{p}_2)$, which is denoted $\hat{p}_1$ in figure 10.3. Because the more pressure that group 2 exerts the lower is the influence of group 1, group 1 finds it optimal to apply more pressure to offset the greater pressure of group 2. This response implies that $\psi_1(p_2)$ is increasing in $p_2$, as shown in figure 10.3.

A political equilibrium is defined as a pair of pressure levels such that neither group has an incentive to change its decision. In other words, the pair of pressure levels $(p_1^*, p_2^*)$ is a political equilibrium if, given that group 2 applies pressures $p_2^*$, $p_1^*$, is the pressure that maximizes group 1’s welfare and, given that group 1 applies pressure $p_1^*$, $p_2^*$ is the pressure that maximizes group 2’s welfare. 26 A political equilibrium is then defined by the intersection of the two best response functions $\psi_1(p_2)$ and $\psi_2(p_1)$ as at that intersection both interest groups are simultaneously exerting optimal levels of pressure. The political equilibrium in figure 10.3 is then the pair $(p_1^*, p_2^*)$.

26. For those who read the section on game theory in chapter 5, a political equilibrium is just a Nash equilibrium for a game in which groups simultaneously choose how much pressure to apply.
The political equilibrium has both interest groups investing in pressure so as to influence the political process. The optimal pressure for each group is very much dependent on the level of pressure exerted by the other group, because what determines regulatory policy is relative influence. As a result, the free-riding problem inherent in all groups is not as important as had been previously thought. Because all groups are subject to free-riding, what is important is the relative severity of free-riding. When the free-riding problem is less severe in group 1 than in group 2 (perhaps because group 1 has fewer members), group 1 will have a relative advantage over group 2. This conclusion is true regardless of whether or not group 1 has a severe free-riding problem in some absolute sense.

Another important property to note about the equilibrium is that it is not Pareto efficient. Both groups could invest fewer resources and achieve the same level of relative influence. Because relative influence is all that matters, the political outcome would be the same but at a lower cost for both groups. As an example of this phenomenon, consider the case of competition among cable operators for the cable television franchise in the New York City boroughs of Brooklyn, Queens, Staten Island, and the Bronx:
All the [franchise] applicants have hired influential lawyers and public-relations consultants, a roster of whom reads like a Who’s Who of former city and state officials. . . . [A vice president for one of the applicants] contends that these friends at city hall (who typically command fees of about $5,000 per month) have tended to cancel one another out.27

Competition among groups for influence in the political process uses up economic resources to obtain the wealth transfer, resulting in a Pareto-inefficient outcome. The logic behind this result is exactly the same as that for the Pareto inefficiency of the Cournot outcome in the oligopoly setting (see chapter 5).

Given the theory of a political equilibrium, let us now use it to generate testable hypotheses concerning the properties of regulation. One important result is that if the marginal deadweight loss from regulation, , increases then the amount of regulatory activity decreases (measured by the amount of wealth transfer ). An increase in the marginal deadweight loss means that group 2 incurs a bigger loss for any given transfer received by group 1. This greater potential loss spurs group 2 to apply more pressure for any given anticipated level of pressure by group 1. This effect of a rise in on group 2’s behavior is then represented by a shift in its best response function from to (see figure 10.3). For example, if group 1 is expected to apply pressure , then group 2 now chooses to apply pressure rather than because the welfare loss imposed on group 2 is higher for any given value of (because of a higher value of ). As a result, the new political equilibrium is , which entails more pressure by group 2, as , but also more pressure by group 1, as . Though group 1’s best response function is unchanged, it is reacting aggressively to the more intense pressure applied by group 2. It is generally true, however, that group 1’s rise in pressure is less than that of group 2’s rise so that, on net, there is less pressure for regulation.28

An important implication of the Becker model is that regulatory policies that are welfare-improving are more likely to be implemented than ones that are not. Suppose that industry A is a natural monopoly and industry B is competitive. The deadweight welfare loss from regulating industry B is greater than that for industry A, ceteris paribus, because industry B is already achieving a welfare optimum while industry A is not. The implication of the preceding analysis is that the greater marginal deadweight loss associated with regulation of industry B means that more pressure will be applied for regulation in industry A than in industry B. The Becker model suggests that industries plagued by market failures (so that the marginal deadweight loss from regulation is relatively low or even negative) are more likely to be regulated. The beneficiary groups have greater potential for gains, and so they

28. We are grateful to Sugata Marjit for pointing out an error in the third edition.
will apply more pressure. Groups harmed by regulation will not be harmed as much because of the lower deadweight loss, and so they will apply less pressure against regulation.

In contrast to the Stigler/Peltzman model of regulation, the Becker model provides some justification for NPT. Where there are market failures, there are potential welfare gains from regulation. Some interest groups stand to gain a lot from regulation, whereas other groups stand to lose a little (relative to interest groups in industries not subject to market failure) because of the absence of relatively large deadweight welfare losses. As a result, there is relatively great pressure for regulation of industries subject to market failure. However, the Becker model, in contrast to NPT, does not state that regulation occurs only when there is a market failure. What determines regulatory activity is the relative influence of interest groups, and this influence is determined not only by the welfare effects of regulation but also by the relative efficiency of interest groups in applying pressure to legislators and regulators.

For the interested reader, the appendix to this chapter develops and analyzes a simple mathematical model that identifies several implications for regulatory activity.

**Taxation by Regulation**

One of the many perplexing aspects of economic regulation is the common use of cross-subsidization. Cross-subsidization is the use of revenue from the sale of one product to subsidize the sale of another product. More specifically, the price of one product is set to exceed its average cost, while the price of a second product is set below its average cost. Such pricing behavior is perplexing because it appears to be inconsistent with both profit maximization and welfare maximization.

An explanation for cross-subsidization is provided by Richard Posner.29 He puts forth the thesis that one of the functions of regulation is to assist the government in its role of redistributing resources. In this light, cross-subsidization is interpreted as a means for redistributing wealth from one group of consumers to a second group of consumers. For example, price regulation entails charging a uniform price for providing local telephone service. Thus, a consumer who lives in a city, where the marginal cost of being connected to the system is low, pays the same fee as a consumer who lives in a rural area, where the marginal cost of being connected is considerably greater. Another example is airline pricing, where, under CAB regulation, the fare was often the same for routes of similar length even though average cost was much higher on low-density routes than on high-density routes. Posner’s argument assumes that society desires to redistribute resources from one class of consumers to another class of consumers and concludes that this purpose could be aided through cross-subsidization. In

practice, it would appear that consumers in less densely populated areas tend to be subsidized at the cost of consumers in more densely populated areas.

The analysis of Posner can be interpreted in light of the Becker model. Cross-subsidization might suggest that some consumers (those with price below cost) have relatively more influence on the political process than other consumers (those with price above cost). Although cross-subsidization cannot be explained by either NPT (because it is inconsistent with welfare maximization) or CT (because it is inconsistent with profit maximization), it can be explained as the result of competition among interest groups to influence government policy for the purpose of raising their welfare.

Summary of Results

We have derived four major results using the Stiglerian approach to the theory of regulation. These results characterize the form of regulation and predict which industries will be regulated. First, there is a tendency for regulation to benefit relatively small groups with strong preferences for regulation at the cost of relatively large groups with weak preferences for regulation. In many cases, the implication of this result is that regulation will be pro-producer. Second, even if regulation is pro-producer, policy (in particular, price) will not be set so as to maximize industry profit. Because of the constraining influence of consumer groups, price will be set below the profit-maximizing level. A third result is that regulation is most likely in relatively competitive or relatively monopolistic industries because it is in those industries that regulation will have the biggest impact on some group’s well-being. Finally, the presence of a market failure makes regulation more likely because the gain to some interest groups is large relative to the loss to other interest groups. As a result, the former will have more influence on the legislative process, ceteris paribus.

Critique of ET: Modeling the Regulatory Process

An important assumption in the models of Stigler, Peltzman, and Becker is that interest groups directly influence regulatory policies. However, when one thinks about the process by which regulation is determined, one realizes there are numerous actors. Voters and special interest groups determine who the legislators are, legislators determine the piece of regulatory legislation (in conjunction with the chief executive), and regulators influence the actual policy that is implemented. In order for interest groups to have a significant impact on regulatory policy, it must be true that the process works the right way. First, interest groups must have a strong impact on the outcome of elections. Second, legislators must be sufficiently constrained by the threat of losing interest group support that they implement the policies supported by the interest groups that got them into office (and are presumably needed for reelection). Third, regulators must be sufficiently under the control of legislators if the policy that is implemented is not to deviate from that desired. An important critique of economic theories of re-
gulation is that they ignore some important elements of the regulatory process by assuming that interest groups adequately control legislators and that legislators adequately control regulators.

Legislators obviously care about being reelected (and thus want to appease the interest groups that originally elected them), but they also care about other things. Like voters, legislators have preferences over issues even if they are not directly affected by them. Such preferences have been referred to as an ideology, where “ideologies are more or less consistent sets of normative statements as to best or preferred states of the world.”

Because interest groups cannot perfectly control or monitor the activities of legislators, legislators can be expected to periodically “shirk” their responsibilities to their interest groups and instead pursue their own ideology (which may or may not conflict with the desires of their interest groups).

In addition to legislators not being puppets of their interest groups, regulators need not be puppets of legislators. Regulators are difficult to control because they have access to information not available to legislators and because it is very costly for legislators to draft new legislation to redirect regulatory policy. As a result, regulators can have considerable discretion in implementing policy.

Nevertheless, it has been argued that congressional oversight committees can be quite effective in controlling regulators. With its budgetary powers, Congress can punish regulatory agencies that pursue the wrong policies. In spite of this threat, regulators clearly have a nontrivial amount of freedom from legislators.

Finally, the role of the judiciary has been ignored in ET. The courts have shown that they can be a key player in the regulatory process:

Judicial consent is necessary when a statute must be reinterpreted in order to implement a change. For instance, reinterpretation of the existing statutes was necessary for the deregulation of airline, trucking, telecommunications and several other industries, and the deregulation of various environmental, health and safety standards. Deregulation occurred only in those cases which were approved by the judiciary. Further, where it did occur, the opposition from committees of Congress was irrelevant.


We are aware of ways in which interest groups can pressure the president and Congress, but how can they influence judiciary decisions? What motivates judges? These are important questions that ET has not addressed.

**Testing Theories of Regulation**

**Does the Empirical Evidence Support the Economic Theory of Regulation?**

The central empirical challenge to ET is to explain both the regulation and deregulation of such industries as railroads, trucking, intercity telecommunications, and crude oil. To address this issue, one should pose the question, What changes in the regulatory environment would induce deregulation?

According to NPT, deregulation would occur when there are changes in cost or demand conditions such that a market failure is either eliminated or sufficiently reduced so as to make deregulation socially optimal. Alternatively, ET would predict deregulation when the relative influence of interest groups that are benefited by regulation is reduced. This decline in influence could happen as a result of changes in cost or demand conditions (by affecting such things as the deadweight loss associated with regulation) or changes in the cost of organizing groups; for example, a new mechanism or technology may be discovered that reduces the free-rider problem. In the case of consumer groups, this new technology may be the arrival of a political entrepreneur like Ralph Nader who is proficient in organizing people and forming coalitions.

A casual survey of the recent deregulatory movement suggests that the evidence is mixed. The deregulation of the railroad industry in 1976–1980 would appear to be broadly consistent with ET. The original regulation of the industry is explained by the industry being more influential in the political process. Although regulation originally allowed above-normal profits, it eventually reduced firm profitability for a variety of reasons. In response, one would expect the industry to pressure for deregulation, which is what it did beginning in the mid-1950s. Unexplained, however, is why it took so long for significant deregulation to take place. In contrast, the deregulation of trucking appears quite inconsistent with ET. The trucking industry was earning large rents from regulation at the time of its deregulation. Further, one is hard-pressed to find a reason why consumers of trucking services would have become more influential in the political process relative to trucking firms and the Teamsters Union. Finally, a case that can be argued to be supportive of either ET or NPT is the deregulation of the intercity telecommunications market. As we will see in chapter 15, deregulation can be explained as a response to the industry no longer being a natural monopoly, as NPT would predict.

weakness in this argument is that, originally, the FCC allowed very limited entry and was steadfastly against allowing entry into certain segments of the market (in particular, long-distance telephone service). This policy is difficult to reconcile with NPT. It could be explained by ET, in that technological changes brought forth a new interest group in the form of prospective firms (initially, MCI). This interest group was influential enough to pressure the FCC to allow partial entry, but AT&T was too influential to allow full entry. It was only a U.S. court of appeals that eventually expanded entry.

More systematic and direct tests of ET have been conducted, and this work seeks to determine whether regulation tends to favor interest groups with a low cost of organizing and a high per capita benefit from regulation. This empirical work investigates why states allow reciprocity for dentist licenses, what determines the pricing of nuclear energy, and why some states went from rate-of-return regulation to price caps in regulating the intrastate long-distance rates of AT&T.

A review of the record reveals that while ET is an important advancement in understanding government intervention, there is still much empirical evidence that would seem to be inconsistent with it. It appears that we have a considerable journey ahead of us in understanding why regulation occurs when it does and why it takes the form that it does. A fuller analysis of the political side of regulation will be dealt with when we examine particular industries. Our studies will include an examination of the political economy of regulations on state banking (the following subsection), railroad and trucking (Chapter 17), strip mining (Chapter 19), and pharmaceuticals (Chapter 24).

Deregulation of Bank Branching Restrictions

Regulatory History

The explosion of mergers, acquisitions, and overall expansion in the banking industry in the 1990s is quite distinct from the pattern of industry evolution that preceded it. Prior to the 1970s, bank expansion was severely limited by regulation. At the federal level, the Bank Holding Company Act of 1956 effectively prohibited a bank from having branches in more than one state. At the state level, intrastate branching was restricted, with some states

limiting a bank to having only a single branch (known as “unit banking”). Now, almost all of these regulations have been dismantled, and to a large extent the United States has unrestricted interstate banking and branching.

**When Did Deregulation Occur and Why?**

Given the restrictions on branching under state regulation, one of the corporate forms to emerge was the multibank holding company (MBHC). An MBHC could own and operate multiple bank subsidiaries but could not integrate them. Each bank had to be run independently so that, for example, the holder of an account in one bank held by an MBHC could not have access to that same account at another bank owned by the same MBHC. From the perspective of customers, the subsidiaries of an MBHC were wholly unrelated. An important step in the deregulatory process was to allow MBHCs to convert subsidiary banks into branches of a single bank and to acquire other banks and make them branches as well. Later deregulation would permit banks to open new branches.

A 199 by Randall Kroszner and Philip Strahan explored the determinants of when a state chose to permit MBHCs to convert subsidiaries into branches of the same bank. Though most states did not engage in this form of deregulation until after 1970 (and the authors explore why that is the case), there was considerable variation as to exactly when deregulation occurred, as can be see in figure 10.4. The analysis focuses on the thirty-six states that deregulated over 1970–1992, with the objective of exploring how well various theories of regulation explain the timing of deregulation. In particular, Kroszner and Strahan consider two of the theories of regulation we have discussed, normative analysis as a positive theory (NPT) and the economic theory of regulation (ET).

**Predictions of NPT and ET**

Recall that NPT hypothesizes that regulatory policy is designed to maximize social welfare. It predicts that deregulation will take place earlier in those states experiencing higher social welfare losses from regulation. In considering the effects of deregulation, it is presumed that there are efficiencies from size such that these branching restrictions are effectively limiting the size of banks. Such a presumption is borne out by the postderegulation expansion and merger-acquisition activity. This suggests that prohibiting MBHCs from converting their subsidiaries into branches benefits small less efficient banks by protecting them from encroachment by more efficient large banks. The associated welfare loss should then be higher in those states where small banks have a bigger presence. Therefore, NPT predicts that the time until deregulation (or the delay in deregulation) is shorter in states with a greater presence of small banks. A second factor pertinent to the timing of deregulation is the presence of small firms. Such firms are especially dependent on the local banking sector for credit (in contrast to large firms, which can raise capital through other means such as equity offerings). Given the impact of the efficiency of the banking sector on small firms, the welfare loss from
regulation should be higher in states with a greater presence of small firms. NPT then predicts that the delay in deregulation is shorter when the presence of small firms in the state is greater.

Let us now consider the predictions of ET with respect to those two factors, small banks and small firms. In that both represent interest groups affected by branching restrictions, ET should have something to say about how their presence influences the timing of deregulation. If larger banks were unrestricted and could expand, such expansion would allow the realization of certain efficiencies that would put the small banks at a disadvantage. Hence, small banks benefit when deregulation is delayed. According to ET, we would then expect small banks to exert pressure to delay deregulation so that, contrary to NPT, it predicts that delay is increasing in the presence of small banks. On the other hand, the prediction of ET about the relationship between the presence of small firms and the timing of deregulation is the same as for NPT. In that small firms benefit from deregulation, a bigger presence of
small firms should result in more resources (for example, lobbying) being used to speed up deregulation.

**Performance of ET and NPT**

These predictions were tested by estimating the relationship between the time until deregulation and measures of the presence of small banks and small firms (the authors also take account of several other relevant factors). The presence of small banks in a state is measured by the percentage of banking assets in the state controlled by small banks where a bank is “small” if its assets are below the median level of assets for banks in that state. The presence of small firms in a state is measured by the proportion of all establishments in the state with fewer than twenty employees. Kroszner and Strahan find that delay in deregulation is greater when the presence of small banks is larger and the presence of small firms is smaller. In that the latter prediction is consistent with both ET and NPT while the former is consistent only with ET, ET seems to perform better than NPT in explaining the timing of deregulation of bank branching restrictions.

---

**Summary and Overview of Part II**

As with any sort of economic phenomenon, there are certain empirical regularities associated with economic regulation. It typically entails regulation over price, quantity, and/or the number of active firms. Regulatory activity also has certain time-series properties. We have witnessed periodic bursts of legislation. A large amount of economic regulation took place after the Great Depression, whereas deregulation was hot in the 1980s.

This chapter provided a brief review of the regulatory process, but it could hardly do justice to the complexity of this process. Many economic agents are involved at the time of regulation’s inception, implementation, and, perhaps, its dismantling. To understand why the regulatory environment looks the way it does, one must understand the motives of consumers, firms, unions, legislators, regulatory commissioners, and government bureaucrats. Several theories of why regulation takes the form that it does were discussed. Different variants of the economic theory of regulation appear to be most consistent with the evidence. Nevertheless, there is still much regulation that this theory cannot explain. More research is required before we will have a complete theory of regulation.

In concluding this chapter, let us provide a brief overview of part II. The chapters on economic regulation are divided into two segments: the regulation of natural monopoly and the regulation of potentially competitive markets. Because natural monopoly is perhaps the most important basis for economic regulation, considerable attention is given to understanding what a natural monopoly is, how best to regulate it, and what the effects of regulation are. Chapter 11 provides an introduction to natural monopoly. The standard form of natural
monopoly regulation, along with a discussion of its effects, is provided in chapter 12. In light of its importance, we also consider alternative methods for handling the problem of natural monopoly. Chapter 13 analyzes franchise bidding, using cable television as an application; chapter 14 considers public enterprise, using municipally owned electric utilities as an application. The final chapter on natural monopoly focuses on dynamic issues related to regulation. This task is performed in chapter 15, where the intercity telecommunications market provides an interesting case study. Chapters 16 through 18 assess the effects of regulation in industries that are potentially competitive. A theoretical discussion of these effects and how one might estimate their quantitative size is provided in chapter 16. An analysis of the regulation of transportation is provided in chapter 17 where we focus on the price and entry/exit regulation of the railroad, trucking, and airline industries. Concluding part II, chapter 18 considers price regulation in the crude oil and natural gas industries.

Appendix

A Theory of Interest Group Competition

Consider a market with two interest groups, consumers and firms. For simplicity, suppose that there is either a single firm or that firms act in perfect unison. There are \( N \geq 2 \) consumers who may or may not act in a coordinated fashion. The situation is one in which firms lobby to be regulated (with an associated rise in their profits) and consumers may respond by lobbying against regulation.

Competition for control of government policy takes place in three stages. In stage 1, firms decide how much to spend on lobbying for regulation. Let \( f \) denote the amount of expenditure that they incur. In stage 2, consumers observe \( f \) (and thus know what firms have spent), and each consumer decides whether or not to organize consumer interests in opposition to firms. The organizing cost to an individual consumer is \( z > 0 \). If no consumer incurs that organizing cost, then consumers are unable to mount a response, so regulation is put in place (as long as \( f > 0 \)). If instead at least one consumer incurs that cost, then consumers are effectively organized as a group, and the game moves to stage 3. In stage 3, the consumer group decides how much to spend on lobbying against regulation. Let \( c \) denote their expenditure. Regulation is assumed to transpire if the firms outspend consumers (that is, \( f > c \)), while regulation is prevented otherwise (that is, when \( c \geq f \)). In the event regulation is put in place, each consumer’s welfare declines by an amount \( ((1 + x)T)/N \) (and thus total consumer welfare falls by \((1 + x)T\)) and firms’ profit is raised by \( T \). Hence, the welfare loss is \( xT \) and it is assumed \( x \geq 0 \).

The task is to derive a solution that has each party acting in its best interests, while taking into account how the other party will behave. The method of backward induction is used, which means solving the stage 3 game and then using that solution to solve the stage 2 game and then using that solution to solve the stage 1 game.

Let us then start with stage 3 which supposes that at least one consumer incurred the organizing cost of \( z \). The consumer group is faced with firms having spent \( f \). If firms spent nothing \( (f = 0) \), then consumers will do likewise \( (c = 0) \), and avoid regulation. If instead \( f > 0 \) and regulation occurs, then firms’ payoff is \( T - f \). Hence, it is never optimal for firms to spend more than \( T \), as they would have been better off spending zero. Let us then suppose that \( 0 < f < T \). In that case, each consumer’s payoff is shown below:

\[
\begin{cases} 
  \frac{c}{N} & \text{if } c \geq f \\
  -\frac{(1 + x)T - c}{N} & \text{if } c < f.
\end{cases}
\]

(This does not include the organizing cost, but that cost is sunk and therefore irrelevant for decision making at this stage.) If consumers choose \( c \) so that it is at least as great as \( f \), then regulation is prevented, and each consumer just incurs her share of lobbying expenditure, which is \( c/N \). This is why the payoff is \( -c/N \) when \( c \geq f \). If instead \( c < f \),
then regulation occurs, which means each consumer loses welfare of \((1 + x)T/N\) plus his or her share of lobbying expenditure, which means a consumer payoff of \(-((1 + x)T - c)/N\).

First note that consumers spending \(f\) is clearly preferable to spending anything more than \(f\), since matching firms’ expenditure is sufficient to avoid regulation. Next note that \(c = 0\) is preferable to any \(0 < c < f\), since spending less than firms do does not prevent regulation, so that consumers should spend nothing. The relevant alternatives for consumers are then to spend \(f\) or 0. Matching firms’ expenditure is preferable to not trying to stop regulation when

\[
\frac{f}{N} > \frac{(1 + x)T}{N} \quad \text{or} \quad (1 + x)T > f.
\]

Since \(T > f\) and \(x \geq 0\), then this inequality does indeed hold. We conclude that if consumers organize as a group, their optimal action is to match the expenditure of firms and thereby avoid regulation.

Now let us move to stage 2 and examine whether consumers do indeed choose to organize. Suppose \(f > 0\) so that regulation would take place if consumers fail to organize. Consider the possibility that no consumer chooses to incur the organizing cost. Is this an equilibrium, or will one consumer choose to lead the way? If no one else organizes, the payoff to a consumer from doing likewise is \(-((1 + x)T/N)\), as regulation prevails. If this consumer instead incurs the cost \(z\) then regulation is prevented and that consumer’s payoff is \(-c/N - z\), which includes the organizational cost and the share of lobbying expenditure (recall that we showed that an organized consumer group would match \(f\)). It is an equilibrium for all consumers to choose not to organize only when:

\[
-(1 + x)T > \frac{f}{N} - z \quad \text{or} \quad z > (1 + x)T - zN.
\]

Thus, if firms spend enough (that is, \(f > (1 + x)T - zN\)), consumers will fail to organize and regulation occurs. However, if instead \(f \leq (1 + x)T - zN\), then it is not an equilibrium for consumers to fail to organize and, furthermore, it is an equilibrium for exactly one consumer to incur the organizing cost. (Note that it cannot be an equilibrium for two consumers to incur the organizing cost.) In that case, consumers end up spending \(f\) in stage 3, and regulation is avoided.

Let us now consider the behavior of firms in stage 1. By the preceding analysis, we know that if \(f \leq (1 + x)T - zN\), then consumers will respond by effectively lobbying against regulation. Hence, firms’ payoff is \(-f\) when \(f \leq (1 + x)T - zN\), as consumers organize and succeed. It is then clear that firms prefer \(f = 0\) to any \(f \leq (1 + x)T - zN\), for why lobby if it is for naught? If instead firms spend more than \((1 + x)T - zN\), then consumers will not organize a response, as defeating regulation is too costly. In that case, firms’ payoff is \(T - f\). By spending a penny more than \((1 + x)T - zN\), regulation occurs, which yields an approximate payoff of \(T - (1 + x)T + zN\) or \(zN - xT\). We conclude that firms will spend enough to result in regulation only when

\[zN - xT > 0,
\]

where the left-hand side expression is the payoff from spending \((1 + x)T - zN\) and getting regulation and the right-hand side is the payoff from spending zero and not getting regulation. Regulation occurs when \(zN - xT > 0\) and does not occur otherwise.

This simple theory can now be used to provide some insight as to when regulation is likely to occur.

**Result 1:** Regulation is less likely when the associated welfare loss is greater.

Rearranging the inequality \(zN - xT > 0\), we find that regulation occurs only when \(x < zN/T\). In that \(x\) measures the extent of the welfare loss from regulation, Result 1 follows. When \(x\) is larger, so that the welfare loss is greater, consumers lose more for each dollar that the firms gain. Consumers are then more likely to mount an effective response and firms are less likely to find it worthwhile to spend enough to deter such a response.

**Result 2:** Regulation is more likely when there are more consumers.

Since regulation occurs only when \(N > xT/z\) (once again, this follows from rearranging \(zN - xT > 0\)), more consumers makes it harder to avoid regulation. The cost to any consumer coordinating a response is \(z\). However, the consumer’s share of the net gain to organizing is lower, as there are more consumers with whom to share it. Hence, regulation is more likely when consumers are numerous and thus each has a only small amount at stake. What is important here is that regulation is more likely when the incentive to organize for those agents opposed to regulation is less than the incentive for those in favor of regulation.
Result 3: Regulation is more likely when the cost of consumers coordinating a response is higher. This is true because regulation occurs only when \( z > xT/N \). For example, the presence of a political entrepreneur whose cost from organizing is low or who derives personal satisfaction from doing so will make an industry drive for regulation less likely to succeed.

Result 4: The bigger the effect of regulation, the less likely is it that regulation will occur. The effect of regulation is measured by \( T \) and we know that regulation occurs only when \( T < zN/x \). The bigger is \( T \), the bigger is the welfare loss to consumers, \((1 + x)T\), and thereby the more likely it is that consumers will organize to prevent regulation.

Questions and Problems

1. Do you agree with the *Nebbia v. New York* decision? If not, what do you think would have been a better judicial decision?
2. What are the roles of the legislature, the judiciary, and the regulatory agency in deregulation? How do interest groups affect deregulation? Should they be allowed to affect regulatory policy?
3. Sometimes, former regulatory commissioners are hired by the industry that they previously regulated. What effect do you think this practice has on the relationship between a regulatory agency and the industry? Should it be allowed? Discuss the advantages and disadvantages of prohibiting this practice.
4. Is there a theory that can explain why a competitive industry like taxicabs is regulated and why a monopolistic industry like local telephone is regulated? What about an oligopolistic industry like the airlines?
5. Can one explain why the railroad industry was regulated and then deregulated almost a century later? What about the regulation and deregulation of trucking?
6. What is the empirical evidence for and against the economic theory (ET) of regulation?
7. What would be the effect on regulatory practices if regulatory agencies were composed of seven members where two members represented firms, two members represented workers, and three members represented consumers? More generally, how do you think regulatory commissioners should be chosen?
8. Use the economic theory of regulation to explain the existence of trade barriers like tariffs and quotas.
9. What do you think caused the wave of deregulation that took place during the 1970s and 1980s?
10. In many cities, the number of taxicabs is controlled by regulation, as are rates. In the 1980s there was a loosening of entry restrictions in some cities but not others (see chapter 16 for a description). How could you use the deregulation of taxicab markets to test various theories as to why there is economic regulation?
Instructor Evaluation Copy

Not for Classroom Use
11 Theory of Natural Monopoly

As we discussed in chapter 10, there are a number of market-failure arguments for economic regulation. Perhaps the most important and widely accepted one is natural monopoly. It provides the rationale for regulating electric power and natural gas distribution, local telephone service, water supply, and some common carrier transportation services. We begin this chapter with a discussion of the theory of natural monopoly. Actual regulation of natural monopoly will be the subject of the next two chapters.

We will be taking an economic efficiency view of natural monopoly here. In previous chapters we discussed various explanations for the existence of regulation, including market-failure and capture theory hypotheses. In this chapter we focus exclusively on the natural monopoly market-failure argument and various theoretical and actual solutions.

This chapter is primarily theoretical, but it also serves as an introduction to the next few chapters. Chapter 12 is concerned with the practice of natural monopoly regulation and an evaluation of its benefits and costs. Chapters 13–15 discuss several alternatives to regulation that are introduced only briefly here.

The Natural Monopoly Problem

An industry is a natural monopoly if the production of a particular good or service by a single firm minimizes cost. The typical example is production of a single commodity, where long-run average cost (LRAC) declines for all outputs. Such a case is illustrated in figure 11.1. Because LRAC is declining, long-run marginal cost (LRMC) necessarily lies everywhere below it.

The case shown in figure 11.1 makes clear the public policy dilemma. Simply stated, the problem is how society can benefit from least-cost production—which obviously requires single-firm production—without suffering from monopoly pricing. The idea, of course, is that a single firm would eventually win the entire market by continuing to expand output and lowering its costs. Having won the market, it could then set the monopoly price. Or this high price attracts entry but then resources are being wasted, since cost is minimized with a single firm producing. The appendix to this chapter provides a more complete and detailed examination of how competition fails to perform well in a market that is a natural monopoly.

Shortly, we will turn to an analysis of the variety of solutions to this problem that have been proposed. Before we do so, however, we will examine more carefully the definition and characteristics of natural monopoly.
Permanent and Temporary Natural Monopoly

An important distinction is that between permanent and temporary natural monopoly. Figure 11.1 illustrates the case of permanent natural monopoly. The key is that LRAC falls continuously as output increases. No matter how large market demand is, a single firm can produce it at least cost.

A temporary natural monopoly is shown in Figure 11.2. Observe that LRAC declines up to output $Q^*$ and then becomes constant thereafter. Hence, as demand grows over time, a natural monopoly when demand $DD$ prevails can become a workably competitive market when demand $D1,D1$ holds.

One can argue that such a cost curve can be used to describe intercity telephone service. There are several factors that give rise to sharp unit-cost savings at low volumes of telephone calls, but they play out as volume increases. For example, a microwave telephone system consists of a number of stations, about twenty to forty miles apart, that transmit signals of specific frequencies. Each station requires land, a building, a tower and antennas, electronic

---

1. The term permanent is perhaps misleading, inasmuch as one can never rule out dramatic technological changes that could convert a natural monopoly into a competitively structured industry. For an examination of such a case, see chapter 15.
equipment, and so on. These inputs do not all increase proportionately with the number of circuits, and therefore, as volume increases the fixed costs can be spread over more calls. This spreading effect becomes less and less significant, however, as volume grows.

As an example, long-distance telephone service between New York and Philadelphia required only 800 circuits in the 1940s. At this capacity, unit costs were falling and constituted a natural monopoly situation. In the late 1960s the number of circuits had risen to 79,000 (largely because of the requirements of television), and this volume was such that unit costs were essentially flat (beyond \( Q^* \) in figure 11.2). Hence, by the late 1960s the temporary natural monopoly had disappeared.

This phenomenon is not rare. Railroads possessed significant cost advantages in the late 1800s, and these advantages were considerably eroded with the introduction of trucking in the 1920s. This example introduces a new element, namely, technological change. That is, over long periods of time it is likely that the cost function will shift as new knowledge is incorporated into the production process. Hence, permanent natural monopoly is probably a rare category. Technical change can shift cost functions so as to render competition workable. And, as we will see later, a serious deficiency of regulation seems to be that it often fails to “disappear” when the natural monopoly does.

---

2. Strictly speaking, technological change in lowering costs was also present in the telephone service example.
In the real world, a single-commodity producer is rare. Electric utilities supply high- and low-voltage, peak and off-peak power; telephone companies provide local and long-distance service; and so on. It turns out that multiproduct natural monopoly is not only more realistic, it also creates important theoretical issues that do not exist in the single-product case.

The definition of natural monopoly is that the cost function is subadditive. Subadditivity refers to whether it is cheaper to have one firm produce total industry output or whether additional firms would yield lower total cost. For outputs less than \( Q' \), one firm is the least-cost solution, and therefore cost is subadditive for that range of outputs.

In order to examine the least-cost solution for outputs greater than \( Q' \), we introduce the minimum average cost function for two firms, \( AC_2 \). This curve and the single-firm \( AC \) curve from figure 11.3 are both shown in figure 11.4. The curve \( AC_2 \) is obtained by construction.

Figure 11.3
Economies of Scale up to Output \( Q' \)

Subadditivity and Multiproduct Monopoly

In the real world, a single-commodity producer is rare. Electric utilities supply high- and low-voltage, peak and off-peak power; telephone companies provide local and long-distance service; and so on. It turns out that multiproduct natural monopoly is not only more realistic, it also creates important theoretical issues that do not exist in the single-product case.

The definition of natural monopoly is that the cost function is subadditive. We begin by explaining this concept in the single-product case because it can be illustrated graphically. Consider the average cost curve shown in figure 11.3. Average cost declines until the output \( Q' \) is reached, and then begins to increase. Economies of scale are said to exist at all outputs less than \( Q' \) and diseconomies at all outputs greater than \( Q' \). Subadditivity refers to whether it is cheaper to have one firm produce total industry output or whether additional firms would yield lower total cost. For outputs less than \( Q' \), one firm is the least-cost solution, and therefore cost is subadditive for that range of outputs.

In order to examine the least-cost solution for outputs greater than \( Q' \), we introduce the minimum average cost function for two firms, \( AC_2 \). This curve and the single-firm \( AC \) curve from figure 11.3 are both shown in figure 11.4. The curve \( AC_2 \) is obtained by construction.

---

from $AC$ in the following manner. We know that for least-cost production, each firm must produce at the same output rate and thereby have the same marginal cost. (If that were not true, then one could shift output from high marginal cost firms to low marginal cost firms and thereby reduce total cost.) Hence, for a given point on the $AC$ curve, simply double the output rate to obtain a point on the $AC_2$ curve. For example, at the minimum average cost point $M$ on $AC$, double $Q'$ to get $2Q'$, which corresponds to the minimum point $M'$ on $AC_2$.

The intersection of $AC$ and $AC_2$ at output $Q^*$ defines the range of subadditivity. For all outputs less than $Q^*$, a single firm yields least-cost production. Hence the cost function is subadditive for outputs less than $Q^*$. Notice that subadditivity is the best way to define natural monopoly. Even though diseconomies of scale obtain between $Q'$ and $Q^*$, it would be in society’s interest to have a single firm produce in that range. An important point is that economies of scale (declining average cost) are not necessary for a single-product natural monopoly (although they are sufficient).

When we turn to multiproduct natural monopoly, the distinction between subadditivity and economies of scale becomes even greater. Again, the proper definition of natural monopoly is that the cost function is subadditive. That is, whatever the combination of outputs desired (say, eighty-five cars and sixty-three trucks, or twenty-five cars and seventy-eight trucks), it is cheaper for a single firm to produce that combination if the cost function is subadditive.
In the multiproduct case, it can be shown that economies of scale are neither necessary nor sufficient for costs to be subadditive. Economies of scale would hold, for example, if the total cost of producing, say, a 10 percent greater quantity of each commodity increased by some amount less than 10 percent. The reason that economies of scale are neither necessary nor sufficient for subadditivity is that in the production of multiple products, the interdependence among those outputs is important.

Although various ways have been proposed for measuring these interdependencies, the concept of economies and diseconomies of scope is appealing intuitively. Economies of scope mean that it is cheaper to produce, say, eighty-five cars and sixty-three trucks within a single firm than it is for specialty firms to produce the required outputs. If you think of peak-period electric power and off-peak power as different commodities, then economies of scope are clearly present—the two commodities can share the same power plant and distribution system.

William Sharkey has given an example of a cost function that possesses economies of scale for all outputs but is nowhere subadditive. His example is

\[ C(Q_1, Q_2) = Q_1 + Q_2 + (Q_1 Q_2)^{1/3}. \]  

Notice that the total cost after increasing each output by 10 percent is

\[ C(1.1Q_1, 1.1Q_2) = 1.1Q_1 + 1.1Q_2 + 1.1^{2/3}(Q_1 Q_2)^{1/3}, \]

whereas the total cost increased by 10 percent is

\[ 1.1C(Q_1, Q_2) = 1.1Q_1 + 1.1Q_2 + 1.1(Q_1 Q_2)^{1/3}. \]

Because the former is less than the latter, economies of scale exist. Nevertheless, the function has diseconomies of scope that sufficiently outweigh the economies of scale to make cost nowhere subadditive.

To see this point, note that the third term in the cost function in equation 11.1 adds a positive amount to cost whenever both outputs are produced together. If, for example, all \( Q_1 \) was produced by firm A and all \( Q_2 \) was produced by firm B, then the sum of the total costs of the two firms would be less than if all production was carried out in a single firm, C; specifically,

\[ C_A = Q_1, \quad C_B = Q_2, \quad \text{so} \quad C_A + C_B = Q_1 + Q_2 \]
\[ C_C = Q_1 + Q_2 + (Q_1 Q_2)^{1/3}. \]


Because $C_A + C_B < C_C$, production in the specialty firms, A and B, is cheaper than in a single firm, C. Thus, economies of scale are not sufficient for cost to be subadditive because there can be diseconomies of scope.

In summary, the definition of natural monopoly in the multiproduct case is that the cost function must be subadditive. Subadditivity of the cost function simply means that the production of all combinations of outputs is accomplished at least cost by a single firm. It is a complex matter to specify the necessary and sufficient conditions for costs to be subadditive. We have shown through some simple examples, however, that it generally depends on both economies of scale and economies of scope. If both exist, then subadditivity will likely obtain. Economies of scale alone, however, can be outweighed by diseconomies of scope. Thus, although economies of scale in the single-product case imply natural monopoly, this statement does not hold true for the multiproduct case.

Before turning to the various policy solutions to the natural monopoly problem, we shall briefly explain a related concept known as sustainability. It can be explained best by reference to figure 11.5, which reproduces the cost function for the single-product case from figure 11.4. Recall that the cost function is subadditive for outputs less than $Q^*$. Now consider a

Figure 11.5
Sustainable Natural Monopoly up to Output $Q'$

case in which market demand $DD$ intersects average cost somewhere between $Q'$ and $Q^*$, where $AC$ is rising. If a single firm were to supply all output demanded at a price equal to average cost (at price $P_0$ and output $Q_0$, so that the firm would just cover all its costs), the natural monopoly would be termed unsustainable. That is, under certain assumptions, a potential entrant would have an incentive to enter the market and produce a share of total output even though doing so would increase the cost of producing the total industry output.

The assumptions referred to above are that the entrant expects the incumbent firm to keep its price unchanged for some period of time after entry, and that the incumbent will supply the residual output.\textsuperscript{7} Under these assumptions, the entrant would perceive that it could profit by offering to sell output $Q'$ at some price above its minimum average cost (point $M$) but slightly less than the price $P_0$ being charged by the incumbent.

By contrast, a sustainable natural monopoly would be one where market demand intersects $AC$ in figure 11.5 to the left of $Q'$. In this case an entrant cannot undercut the incumbent and therefore has no incentive to enter. The concept of sustainability is relevant where a regulatory agency must decide whether to allow entry in a particular market of a multiproduct natural monopolist.

\section*{Alternative Policy Solutions}

In this section we examine various alternatives that have been proposed (and, in some cases, implemented) to correct the natural monopoly inefficiency. These alternatives include “doing nothing,” various “ideal” solutions, competition among bidders for the right to the monopoly franchise, and, finally, actual regulation, as practiced in the United States, and public enterprise, as exemplified by the Postal Service.

The first alternative mentioned, doing nothing, might be appropriate if the potential monopoly power is not great. For example, a cable television system might be viewed as a natural monopoly but one with quite limited capacity for earning excess returns, given that substitutes for cable television are not too distant, with direct broadcast satellite and over-the-air broadcasting.

We consider first a collection of “ideal” pricing solutions. The adjective “ideal” is employed to indicate that we are assuming that the firm is to be operated in the public interest and that the only issue is what prices produce economic efficiency.

\textsuperscript{7} A further assumption is that the entrant perceives no entry barriers in the form of sunk costs. That is, the entrant believes that whatever investment is required can be recovered by transferring it elsewhere or by sale. All of these assumptions have been subject to controversy.
Ideal Pricing

The most obvious candidate for the efficient price is, of course, marginal cost. A natural monopolist that charges marginal cost for each product is said to practice linear (or uniform) marginal cost pricing. In other words, a customer’s expenditure for a product is a linear function of price and quantity sold, $PQ$. On the other hand, if the firm charges a fixed fee $F$, regardless of the amount bought, and also a per-unit charge $P$, nonlinear (or nonuniform) pricing would be in effect. Then the customer’s expenditure would be a nonlinear function of the form $F + PQ$.

In our ideal pricing discussion, we begin with the linear marginal cost pricing solution. After considering nonlinear pricing we examine the so-called Ramsey pricing alternative, which applies to multiproduct cases. The section concludes with a discussion of a theoretical proposal that would induce profit-maximizing firms to price efficiently.

Linear Marginal Cost Pricing

Consider a single-product natural monopolist with decreasing average costs over the relevant output range. Figure 11.6 shows such a situation where market demand is $DD$. The marginal cost price would be $P_0$ with output $Q_0$. The price does meet the well-known requirement for efficiency; however, on closer examination, several serious difficulties arise. An obvious difficulty is the loss inflicted on the firm, shown by the shaded rectangle $RP_0ST$. Any enterprise would need a subsidy to continue to operate at this output level, because price is less than average cost. The next question is to ask where the subsidy is to come from and what effect this will have on economic efficiency.

The only “correct” solution is for the government to raise the subsidy through a lump-sum tax, that is, a tax that would not distort other decisions throughout the economy. Such taxes are rarely, if ever, used in practice. Income taxes and sales taxes are unacceptable because they create inefficiencies themselves by introducing wedges between prices and marginal costs. Even this “correct” solution (lump-sum tax to pay subsidy) is subject to some rather persuasive opposing arguments. Three frequently mentioned arguments are:


9. The loss is equal to the difference between price and average cost, multiplied by output.
1. If total costs are not covered by consumer expenditures, it is possible that total consumer benefits (given by the area under the demand curve)\(^{10}\) are less than total costs—which means the good should not be produced at all. Figure 11.7 provides such a case. Total costs $AOQB$ (the area under the $MC$ curve) exceed total benefits $DOQB$. Only if consumers are required to actually cover total costs can we be sure that the good is socially beneficial.

2. Because the enterprise’s management knows losses will be subsidized, the incentive and the capacity to control costs are weakened. Postal Service employees, for example, have an advantage in bargaining with management, inasmuch as both sides know that the enterprise will not fail if revenues are less than costs. The Treasury can always be counted on to subsidize the Postal Service in a pinch. Steel industry labor unions do not have this luxury.

3. On distributional grounds, it can be argued that nonbuyers of the natural monopoly good should not be required to subsidize the marginal cost buyers. That is, why should the taxes paid by individuals without telephone service be used to subsidize individuals who purchase such service at a loss-creating price?

---

10. Throughout this chapter we make the common assumption that the area under the demand curve measures total willingness to pay by consumers. This requires one to assume that the income elasticity of demand is zero (or small enough to make the error unimportant). See Robert D. Willig, “Consumer’s Surplus without Apology,” *American Economic Review* 66 (September 1976): 589–97.
A major point of the preceding analysis is that enterprises should price so that their revenues cover costs. Furthermore, in the United States, because most public utilities are privately owned firms, it is politically unrealistic to imagine government subsidizing the losses of private firms. Hence we conclude that there are compelling reasons to accept the constraint that natural monopolies should operate such that total revenues and total costs are equated.

In the single-product case, linear pricing implies that price must equal average cost if total revenues must equal total costs. This relationship is shown in figure 11.8 as price $P_0$ and output $Q_0$. This departure from marginal cost pricing leads, of course, to the welfare loss given by the shaded triangular area.11

This argument refers to linear pricing, that is, the buyer pays a single price per unit, and therefore the buyer’s total expenditure is proportional to total consumption. An important alternative is nonlinear pricing.

11. For a discussion of welfare loss determination, see chapter 4.
Nonlinear Pricing

A two-part tariff is nonlinear and consists of a fixed payment or fee, regardless of consumption, plus a price per unit. If the price per unit equals marginal cost, then it is possible to have efficient pricing and have total revenues of the firm equal to its total costs.

For example, if the loss under linear marginal cost pricing is estimated to be $K$ (the shaded rectangle in figure 11.6), the fixed fee of the two-part tariff could be set so that the sum over all customers equals $K$. There are various ways for this equality to hold; the simplest is to set the fixed fee equal to $K/N$, where $N$ equals the number of consumers.

There are possible problems with this nondiscriminatory two-part tariff. Because consumers usually vary considerably in terms of their demands for the good, it is possible for some consumers to be driven from the market if $K/N$ exceeds their consumer surpluses at price equal to marginal cost. One might expect this outcome to be more likely for, say, telephone service than for such “necessities” as electricity and water. Hence, efficiency losses will occur if these excluded consumers would have been willing to pay marginal cost. It is also true that in some markets it is not feasible to enforce a fixed fee for the “right-to-buy” at a price per unit. Consumers would have an incentive to have one person purchase for all, thereby paying only one fixed fee. This is not a problem for most public utilities.
The obvious thing to do to avoid excluding consumers is to charge different fixed fees to different consumers or classes of consumers. In short, discriminatory two-part tariffs could tailor the fixed fees to the consumers’ willingness to pay, where the sum of the fixed fees should add up to \( K \). Although this solution is best in terms of efficiency, it may be illegal to so discriminate.

If all consumers must be charged the same fixed fee, it will still be more efficient to use a two-part tariff than to use linear pricing (which in the case of a single product implies average cost pricing). The reason is simply that by using a fixed fee to make a contribution to revenues, the price per unit can be lowered toward marginal cost, thereby reducing deadweight losses. (In principle, one can pick some fixed fee, no matter how small, that will not drive anyone from the market and permit a lowering of the price.)

The next logical question is, What is the optimal two-part tariff? Here, we explain only the economic principle involved.\(^{12}\) Suppose initially that the fee is zero and price equals marginal cost. The result is, of course, a deficit that must be covered by increasing either the fee or the price per unit, or both. In essence, the derivation depends on a balancing of efficiency losses because of exclusion of additional consumers as the fixed fee rises against the increased consumption losses as price per unit increases above marginal cost. Hence the optimal two-part tariff generally will involve a price per unit that exceeds marginal cost and a fixed fee that excludes some consumers from the market.

Multipart tariffs are often used by public utilities. Consider the following example of the type of tariff sometimes used for local telephone service (such tariffs are often referred to as declining-block tariffs).

Fixed fee per month—$5

+10 cents per call for up to 100 calls

+5 cents per call for all calls between 100 and 200

+0 cents per call for all calls above 200

Notice that the marginal price falls as one moves to successively larger calling “blocks”—from 10 cents to 5 cents to 0 cents. This multipart tariff is plotted in figure 11.9 as the bold segmented line \( ABCD \). (The reason for the extensions of these segments in figure 11.9 will become clear shortly.) Hence the figure shows “total consumer expenditure” vertically as a function of total “calls per month” horizontally.

A rationale often given for the declining blocks is that utilities are characterized by economies of scale, and falling marginal prices stimulate consumption, in turn permitting the construction of larger, lower-unit-cost plants. An alternative rationale is to view the

---

Declining-block tariff as a self-selecting set of two-part tariffs, and a set of such tariffs can increase economic efficiency along the lines discussed earlier.

Recall that discriminatory two-part tariffs permit the firm to tailor the tariffs to fit the differences in willingness to pay across consumers. The efficient solution can be achieved if no consumers are excluded from the market and all pay marginal cost per unit. As an approximation to this “ideal,” one can use the multipart tariff in figure 11.9 to cause consumers to self-select a two-part tariff that they prefer, whereby consumers with high willingness to pay, pay high fixed fees in return for lower prices per unit.

The three “self-selecting” two-part tariffs are

<table>
<thead>
<tr>
<th>Fixed Fee</th>
<th>Price/Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$5</td>
<td>10 cents</td>
</tr>
<tr>
<td>$10</td>
<td>5 cents</td>
</tr>
<tr>
<td>$20</td>
<td>0 cents</td>
</tr>
</tbody>
</table>

One can represent a two-part tariff by a vertical intercept (for the fixed fee) and a straight line with slope equal to the price per unit. The three such lines in figure 11.9 represent the...
three two-part tariffs to which we have referred. (Notice that no consumer would wish to consume on portions of the tariffs other than the lower boundary $ABCD$. Hence it does not matter that these “dominated” portions of the two-part tariffs are not actually part of the declining-block tariff.) The point is that the declining-block tariff has the same effect as confronting consumers with two-part tariffs that are tailored to their demands. And, of course, all consumers are free to choose the particular tariff that they prefer, so that there is no discrimination involved that is likely to be disallowed.

Up to this point our discussion of ideal pricing has been limited to a single-product natural monopolist. We now turn to the case of a multiple-product natural monopolist and describe what has become known as Ramsey pricing.

**Ramsey Pricing**

In a famous article published in 1927, Frank Ramsey suggested the following pricing (and taxing) method. It is applicable to a multiproduct natural monopolist that would generate losses if linear marginal cost pricing were used. In essence, Ramsey prices are those linear prices that satisfy the total revenues equal total cost constraint and minimize the deadweight welfare losses. Note that Ramsey prices are linear prices—one for each product—so that we are implicitly ruling out multipart tariffs.

It is useful to illustrate Ramsey pricing with a numerical example. Let the natural monopoly be a two-product firm with total cost

$$C = 1800 + 20X + 20Y.$$  

The market demands for the two goods $X$ and $Y$ are given by

$$X = 100 - P_x,$$
$$Y = 120 - 2P_y.$$  

An important assumption that we will make for our example is that the demands are independent; the demand for $X$ does not depend on the price of $Y$, and vice versa. The more general case of interdependent demands involves much more complex mathematics and is beyond the scope of the discussion here.

It should be obvious that the marginal costs of $X$ and $Y$ are each $20$, and that marginal cost prices would exactly cover the variable costs but not the fixed cost of $1,800$. Because the firm must cover its total costs, it is clear that the prices will necessarily exceed their respective marginal costs. One possibility would be to raise the prices by the same proportion above marginal costs until total costs are covered. This is shown in figure 11.10(a).

The figure shows that prices would need to be raised from $20 to $36.1 to generate sufficient revenues just to cover total costs. In particular, the contribution that product $Y$ makes toward fixed cost equals the rectangle $CEFD$. This is just price minus the constant unit variable cost of $20, multiplied by the output of 47.7. Similarly, the contribution that product $X$ makes equals rectangle $CEJK$. The sum of these two rectangles is $1,800. (The fact that the demands intersect at the price equals marginal cost point for each is not necessary, and was chosen merely to make the graphical exposition simpler.)

Now consider the deadweight losses that this proportionate price increase method causes. The deadweight loss triangle for product $Y$ is triangle $DFH$, and it is $JKH$ for product $X$. The actual numerical values are $260$ and $130$, respectively, or a total of $390$. Hence, one way of summing up this method is to observe that it “costs” $390 in deadweight welfare losses to generate the $1,800 necessary for the firm to break even. The question becomes whether one can find another method for raising prices to generate the $1,800 that entails a lower welfare cost.

A bit of reflection while examining figure 11.10(a) might suggest differential price increases. That is, it is clear that the same price increase produces a smaller contribution to fixed cost from product $Y$ at a higher cost in terms of deadweight loss. This observation is not surprising when one realizes that product $X$ has a more inelastic demand (at point $H$) than

15. Because $P_x$ and $P_y$ must be equal under the assumption that the marginal costs are both $20, the $36.1 value can be found by solving the equation that equates total revenues and total costs.
does product $Y$. This difference suggests that it would be better to raise the price of $X$ more than the price of $Y$.

The Ramsey pricing rule that gives the prices that minimize the deadweight losses is to raise prices in inverse proportion to demand elasticities. Mathematically, the rule\textsuperscript{16} is

$$\frac{P_i - MC_i}{P_i} = \frac{\lambda}{\eta_i},$$

where $P_i$ is the price of good $i$, $MC_i$ is the marginal cost of $i$, $\eta_i$ is the absolute value of the elasticity of demand of good $i$, and $\lambda$ is a non-negative constant. Using this rule, one can derive the actual Ramsey prices. They are shown in figure 11.10(b). Hence the firm would minimize the welfare losses by charging $40 for good $X$ and $30 for good $Y$. At these prices, the demand elasticities are 0.67 and 1.0, respectively. The deadweight loss triangles are $200 for good $X$ (triangle $MTV$) and $100 for good $Y$ (triangle $NTV$), for a total of $300. This is, of course, a lower “cost” in terms of welfare by $97 than the proportionate method of figure 11.10(a).

The Ramsey pricing rule can be viewed as providing theoretical justification for so-called value of service pricing that has been used for years in the railroad industry. It has been common for rail rates for shipping gravel, sand, potatoes, oranges, and grapefruits to be lower relative to shipping costs than for liquor, electronic equipment, cigarettes, and the like. The reason is that the elasticities of demand for shipping products that have low values per pound are higher than for products that have relatively high values per pound. (We are assuming that the actual costs of shipping are proportional to weight.)

In summary, all of the ideal pricing schemes discussed have problems (except for the two-part tariff with price equal to marginal cost and no exclusion of consumers by the fixed fee). It should be kept in mind that we have assumed away the very real difficulty of designing incentive systems that will induce enterprise managers to implement these pricing schemes. In short, managers of private firms are presumably interested in maximizing profits, not total economic surplus. Managers of public enterprises may also have objectives other than economic efficiency.

**Price Regulation of Local Telephone and Long-Distance Service Prior to 1984**

A striking counterexample to Ramsey pricing is the case of long-distance and local telephone service prior to the breakup of the Bell System (and even some time afterward). As a regulated monopolist prior to 1984, AT&T offered both local and long-distance service. Due to scale economies, marginal cost pricing failed to cover total cost. To determine what Ramsey pricing prescribes in this situation, one must consider the elasticities of these two services.

Estimates reveal that the demand for local telephone service is significantly less elastic than that for long-distance service. The (absolute value of the) elasticity of access to basic local telephone service is estimated to lie between 0.05 and 0.20 which means that a 10 percent increase in the service charge would reduce the number of subscribers by one-half, to 2 percent. In contrast, estimates of elasticity for long-distance service range between 0.5 and 2.5, with a common estimate exceeding 1.0. Hence, a 10 percent rise in long-distance rates causes a fall in long-distance demand by more than 10 percent. All this is hardly surprising as most people see a telephone as a necessity, whereas long-distance service is a bit more of a luxury.

The prescription of Ramsey pricing is then clear: the price-cost margin for local telephone service should exceed that for long-distance service. And what was actual regulatory practice? Just the opposite! Long-distance service was priced well above marginal cost and profits were used to subsidize the losses incurred from pricing below cost on local service. One study estimated that the welfare gains from replacing this pricing arrangement with Ramsey pricing was on the order of $30 billion a year! That regulatory policy departed so strikingly from what was socially best could reflect a lack of understanding on the part of regulators or instead alternative goals. It is at least partly the latter, as the goal of universal access—all people having telephones—could rationalize the subsidization of local service by long-distance service notes. Regardless, the breakup of the Bell system in 1984 posed a bit of a dilemma in that local and long-distance service were no longer supplied by the same company. The policy solution was to have long-distance suppliers pay an access fee to the local telephone companies in order to ensure that the local telephone companies earned a normal return. In that this access fee was well above the marginal cost of connecting the long-distance company to the local network, it served to maintain this pricing distortion. Over time this access fee has fallen by regulatory mandate and, consequently, the price of long-distance service has fallen and the subscriber fee for local service has risen. Gradually, the distortions in the pricing of local and long-distance telephone service have then been shrinking.

**Loeb-Magat Proposal**

Of course, if regulators had perfect information as to the monopolist’s costs and demands, the ideal pricing schemes that we have discussed could be put into effect by command. However, such is not the case. Although the monopolist may not have perfect information itself, most people would probably agree that the monopolist has much better knowledge of

its costs than the regulators do. Because the firm’s profits will increase with higher prices, the firm has an incentive to overstate its costs (which is the usual basis that a regulator uses to set prices).

Martin Loeb and Wesley Magat assumed that the monopolist knows costs and demand information perfectly, but that the regulator knows demand only. Hence, given this asymmetry of information and the assumption that the monopolist’s objective is to maximize profit, what might the agency do to induce efficient pricing? The Loeb-Magat (L-M) scheme can be explained easily with the aid of figure 11.11, which shows a single-product natural monopolist.

Figure 11.11
Loeb-Magat Incentive Scheme

The monopolist has declining average cost \((AC)\) and demand curve \((AR)\). For simplicity, we assume the total cost function is \(K + vX\); hence, marginal cost \((MC)\) is constant and equal to \(v\). The L-M proposal is to allow the monopolist to choose its own price—this differs from the usual practice of the regulatory agency setting the price. However, they propose having the agency subsidize the firm by an amount equal to consumer surplus at the selected price.

Suppose that the monopolist selects the price \(P_0\). Its profits will be \(P^*DEB - K\). The firm collects \(0X_0EP_0\) from customers and \(P_0EB\) from the regulatory agency. Its variable cost is \(0X_0DP^*\), leaving a variable profit of \(P^*DEB\). Subtracting the fixed cost of \(K\) leaves the profit just asserted. Observe, however, that the firm can do better by lowering price. For example, if the monopolist selected \(P^*\), it is easy to show that its profits will increase to \(P^*AB - K\). That is, profits increase by the usual deadweight loss triangle \(DAE\). This is, in fact, the profit-maximizing solution for the monopolist! Convince yourself that any other price will reduce profits. (Alternatively, note that the proposal causes the demand curve \(AR\) to become the monopolist’s marginal revenue curve, and setting \(MC\) equal to marginal revenue is the profit-maximizing solution.)

The explanation for this price-equal-to-marginal-cost result is simply that the regulator has changed the firm’s objective function by virtue of the subsidy. Now, in effect, the monopolist is maximizing total surplus—the total area under the demand curve minus costs.

The solution is economically efficient, but most people would find it objectionable on distributional grounds. The monopolist is appropriating the total economic surplus! To rectify this problem, Loeb and Magat suggest that a franchise bidding scheme (or a tax scheme) could recover some of the subsidy for the general treasury. In the case shown in figure 11.11, the regulatory agency would auction off the right to operate the monopoly franchise. The key idea is that above-normal returns (of amount \(P^*AB - K\)) are available to the firm that operates the monopoly and that bidding for the franchise would continue until that amount is bid. Note that the subsidy is not completely recovered; there remains a net subsidy of an amount equal to fixed cost, \(K\).\(^{20}\)

Obviously, the L-M proposal is not the perfect solution to natural monopoly. Informational problems about the demand curve and the existence of a subsidy make it an unlikely substitute for the present regulatory process. It has, however, stimulated research by economists toward the goal of understanding how the regulatory process might be improved with respect to providing better incentive structures for natural monopolists.

In the next section we return to the discussion of alternative policy solutions to the natural monopoly problem. In contrast to the ideal pricing solutions that we have been examining heretofore, we now turn to actual solutions that have been used. The first is franchise bidding.

\(^{20}\) For a variation on the Loeb-Magat proposal that eliminates the net subsidy and the need of the regulator to know demand, see D. A. Graham and J. M. Vernon, "A Note on Decentralized Natural Monopoly Regulation," *Southern Economic Journal* 58 (July 1991): 273–75.
Franchise Bidding

Harold Demsetz has argued that the “theory of natural monopoly is deficient for it fails to reveal the logical steps that carry it from scale economies in production to monopoly price in the market place.”21 His point is that it may be possible to have bidding for the right to supply the entire demand (in effect, bidding for a franchise to serve a certain market). Even though only the single firm submitting the low bid would actually produce, there could be competition among potential suppliers. For example, given the situation shown in figure 11.8, the low bid presumably would be a price of $P_0$ for $Q_0$ units.

Note that $P_0$ is not the efficient price. Nevertheless, $P_0$ would be an improvement over the natural monopoly price (a price above $P_0$). Then $P_0$ would be the lowest price bid for the right to supply the market, inasmuch as any lower price would result in losses. At $P_0$ the winning bidder would just cover costs, including a normal return on investment.

This bidding for the franchise argument has stimulated a great deal of useful thinking about alternatives to natural monopoly regulation. However, the highly abstract example here oversimplifies many of the problems that such bidding would raise. A detailed discussion will be provided in chapter 13.

Actual Solutions

In this section we briefly consider actual solutions that have been implemented in response to the natural monopoly problem. There are basically two distinct solutions: the regulatory agency and public enterprise. Extensive discussions of each will be presented in subsequent chapters; only a short treatment is given here.

Regulation

The typical natural monopoly in the United States is a private firm: Consolidated Edison, Verizon, and so on. The firm is controlled by a regulatory agency that must approve the prices the monopolist can charge. A key goal is that the firm’s revenues just cover its costs.

The measurement of costs is obviously a major task for the agency. Indeed, the attempt by the agency to estimate the proper return on capital investment is perhaps its most time-consuming activity. For example, a typical regulatory hearing involves testimony by numerous experts as to the “true” cost of capital for the firm.

In contrast, relatively little of the agency’s resources are expended on the issue of the correct pricing structure. However, this situation is changing, and agencies are becoming more interested in, for example, marginal cost pricing. In short, regulatory agencies try very hard to ensure that the monopolist’s revenues equal its costs, and historically have been less concerned with the pricing structure used.

As a result, there is no simple way to describe the pricing structures used under regulation. Price discrimination is often employed both across customer groups (industrial, commercial, residential, and so on) and within groups (declining block rates, for instance, 5 cents per unit for the first 300 units, 4 cents per unit for the next 500 units, and so on). Richard Schmalensee has observed,

To the extent that utility regulators in the United States have been concerned with rate structures, they have tended to focus on prices paid by different classes of users. But this focus has typically been motivated and informed by considerations of equity or fairness rather than efficiency.22

Hence, regulatory agencies often try to prohibit undue discrimination across customer groups. They require the firm to allocate its total costs to customer groups and then adjust their prices if the revenues by groups do not correspond to the groups’ “fully distributed costs.”

There is a serious problem implicit in this procedure, however, because a large proportion of a firm’s costs are usually common costs. For example, high-voltage power lines are used in common by all customer groups. And although arbitrary accounting rules can be made up to apportion these costs among groups (for instance, in proportion to their respective annual purchases of the product), none are meaningful in an economic sense as a basis for setting prices.

In summary, an important solution to natural monopoly in the United States is regulation. The regulatory solution is not an attempt to implement the ideal pricing schemes discussed earlier. Regulators do not see as their primary objective achieving economic efficiency. Rather, they appear to seek a set of prices that are not unduly discriminatory but that permit total revenues to cover total costs. However, regulatory agencies have become more interested in pricing schemes that promote economic efficiency, including price caps, where a regulated firm can retain all profit as long as its price does not exceed some level. Also, peak load pricing—charging more when demand presses on capacity, and therefore marginal cost is higher—has been implemented by many electric utilities.

**Public Enterprise**

The second actual solution to natural monopoly is public enterprise, or government ownership and operation of the monopoly. This is not as common in the United States as it is in other countries. The Postal Service is an example in the United States. Other examples include various government-owned electric utilities, such as the Tennessee Valley Authority. Public enterprise would appear to be a sensible alternative. If managers could be directed to maximize economic surplus, there would be no need for regulators to try to channel the decisions

of profit-maximizing firms closer to the public interest. However, implementation is problematic and there have been many identified inefficiencies from public enterprise when compare to regulation. This complex issue will be examined in chapter 14.

Summary

This chapter has been an introduction to natural monopoly. Theoretical issues have been introduced and discussed. First, the definition of natural monopoly was developed in both the single-product and the multiproduct cases. A natural monopoly exists if the cost function is subadditive at the socially optimal quantities; total cost is then minimized by having just one firm produce. Second, alternative policy solutions and their difficulties were discussed. The solutions included “doing nothing,” various efficient pricing solutions, competition among bidders for the right to the monopoly franchise, actual regulation, and public enterprise. Of particular interest is how nonlinear pricing, such as a two-part tariff, and Ramsey pricing—having price-cost margins inversely related to a product’s demand elasticity—can result in higher welfare.

In the next chapter we will elaborate extensively on the regulation alternative, with a focus on actual practice. Chapter 15 will examine further issues in natural monopoly regulation, with an emphasis on telecommunications.

Appendix

The Troublesome Case of a Natural Monopoly

Why doesn’t unfettered competition work well when a market is a natural monopoly? Toward addressing this question, let us consider a market with scale economies. There are many prospective firms, each of which can enter at cost \( k > 0 \) (for example, the cost of building a distribution network) and produce at constant marginal cost \( c \geq 0 \). Thus, average cost is \( (klq) + c \), which is declining for all quantities (as depicted in figure 11.1). All firms produce identical products and the market demand curve is denoted \( D(P) \). Note that this market is a natural monopoly.

Competition occurs in two stages. In stage 1, these prospective firms simultaneously decide whether or not to enter the industry. Entry entails incurring \( k \), which is sunk and therefore unrecoverable if the firm decides to exit the industry. Suppose \( n \) firms enter. Then, in stage 2, these firms compete by simultaneously choosing price, with each firm producing to meet demand. Since firms have identical products, consumers will buy from the firm(s) with the lowest price. If more than one firm charges the lowest price, then we will assume that market demand is equally divided among them. So, for example, if \( n = 3 \) and firms 1 and 2 charge a price of 10 and firm 3 charges a price of 12, then firms 1 and 2 each have demand of \( D(10)/2 \) and firm 3 has zero demand.

To derive an equilibrium for this two-stage game, we first solve for what happens in stage 2. An equilibrium is comprised of a price for each firm such that each firm’s price maximizes its profit given other firms’ prices. In that sense, each firm is content with its price and has no incentive to change it. So, suppose that only one firm entered (that is, \( n = 1 \)). Then, in stage 2, the lone entrant is a monopolist, in which case it will set the monopoly price \( P^m \), which is the price that maximizes \( (P - c)D(P) \). Let \( \pi^e = (P^m - c)D(P^m) \) denote monopoly profit. (Note that we exclude

23. Details on game theory and equilibrium can be found in chapter 5.
the entry cost because it is sunk and thus irrelevant for firm decision making.) This is the stage 2 equilibrium when only one firm entered.

Next suppose there is competition in stage 2 (that is, \( n \geq 2 \)). The equilibrium is very different, as it has all firms price at marginal cost! To see that this is the case, suppose instead that the minimum price is being charged, which we will denote \( P' \), and, for the sake of discussion, suppose firm 1 is charging it. Now consider the profit-maximizing price of firm 2. One possibility is for it to match the lowest price of its competitors by pricing at \( P' \) and earning profit of \((P' - c)(D(P')/m)\), where \( m \) is the total number of firms charging \( P' \). This is clearly better than pricing above \( P' \), as that gives it zero demand and thereby zero profit. Better yet is to undercut and price below \( P' \). If it prices at \( P' - \epsilon \) (where \( \epsilon \) is positive but small), it attracts all customers and receives profit of \((P' - \epsilon - c)(D(P')/m)\) which is strictly larger than \((P' - c)(D(P')/m)\) when \( \epsilon \) is small enough. In other words, a small price cut below \( P' \) allows it to increase its demand \( m \)-fold relative to pricing at \( P' \). Thus, firm 2’s optimal price must be less than \( P' \). Now we want to argue that firm 1 pricing at \( P' \) cannot be optimal for firm 1. Since firm 2 is undercutting it, firm 1’s profit is zero when it could instead have positive demand and profit by matching firm 2’s price. As long as the minimum price in the market is above marginal cost, at least one firm is not maximizing profit, as it can do better by undercutting the lowest price of its rivals. We conclude that it is not an equilibrium for the minimum price in the market to exceed cost.

It is easy to argue that, in equilibrium, the minimum price in the market is not less than cost, because if it was, then those firms with the lowest price would be selling each unit at a loss. But this is not optimal, as a firm can always earn zero profit by pricing at cost.

What we are left with is one remaining possibility: Equilibrium has the lowest price in the market equaling marginal cost. Let us show that one such equilibrium has all firms pricing at cost. Given its competitors are pricing at \( c \), a firm earns zero profit by also pricing at \( c \) because, while it has positive demand of \( D(c)/m \), its per unit profit is zero. Unfortunately, it cannot do any better than that. For if it sets a higher price, then its demand drops to zero—as all of its customers leave to buy from its competitors who are pricing at cost—so that its profit is still zero. If it sets a lower price, it increases its demand but now is incurring losses. We conclude that if two or more firms are active in stage 2, then the equilibrium results in perfect competition, because all firms price at marginal cost.

Now let us consider firms’ decisions about entry. In doing so, we presume that each firm accurately conjectures what will transpire in the next stage, depending on how many other firms entered. Thus, a prospective firm realizes that if it is the only entrant, then it gets to earn monopoly profit, but if it has a competitor, then there is cutthroat competition, with marginal cost pricing.

There are actually three possible solutions to the first stage. One solution is that only one firm enters. Let us convince ourselves that this is an equilibrium. Suppose firm 1 plans to enter and the other firms anticipate its entry. Those other firms will surely not enter because if one does so, they know they will earn zero variable profit, which means their net profit from entry is \(-k\); thus, they fail to earn variable profits to cover the cost of entry. Such firms clearly prefer not to enter and instead earn zero profit (which is the presumed profit from not entering). Given it is the only firm that enters, entry is optimal for firm 1 as long as monopoly profit is at least as great as the entry cost: \( \pi^e - k > 0 \). Assuming that condition holds, firm 1 does indeed want to enter. This is then an equilibrium in that each firm is acting optimally given what the other firms are doing. But unfortunately, from society’s perspective, the outcome is rather undesirable. The socially optimal price is marginal cost, but instead what emerges is the higher monopoly price. The solution is then characterized by allocative inefficiency in that too little of the product is being consumed. On the positive side, productive efficiency is achieved, since the least-cost way in which to produce is to have a single firm, and there is indeed only one entrant.

Another solution is for each of the prospective firms to enter with some probability; it is as if they flip a coin to determine whether or not they enter. If there are \( N \) prospective firms, then the equilibrium value of this probability is \( 1 - (k/\pi^e)^{1/(N-1)} \), which lies between 0 and 1, since \( \pi^e > k \). In that the entry decisions are random, a variety of

24. This model is known as the Bertrand price game. When there are three or more firms, there are multiple (Nash) equilibria as it is an equilibrium as long as (1) all firms price at or above cost, and (2) at least two firms price at cost. We have characterized the unique symmetric equilibrium. Regardless, all equilibria have the property that firms earn zero profit. The comment should be added that the Bertrand price game is rather extreme in its prediction that a perfectly competitive outcome emerges with only two competitors.

25. This probability is determined by the following condition. If each of a firm’s rivals enters with probability \( 1 - (k/\pi^e)^{1/(N-1)} \) then a firm’s expected profit from entry is exactly zero. In that case, a firm is indifferent between entering and not entering, since both yield the same expected profit. Given that it is indifferent, then this firm is also content to randomize in its entry. In game theory, this is known as an equilibrium in mixed strategies.
outcomes can occur. One possibility is that exactly one firm enters, which is the same outcome as just described. But another possibility is that multiple firms enter. In that event, each firm regrets entering, as they end up earning $-k$, though, at the time of entry, it was a reasonable decision (that is, expected profit from entry was zero). Think about the welfare implications of this event. In contrast to the first solution, allocative efficiency occurs as firms price at marginal cost, but productive efficiency is lost as the output of $D(c)$ is being produced at a cost of $mk - cD(c)$ (where $m \geq 2$ firms entered), which exceeds the cost from having a single firm produce.

In conclusion, an unregulated market results in welfare losses. Either there is allocative inefficiency, with one firm pricing too high, or productive inefficiency, with multiple firms producing so that total industry cost is not minimized. This highlights the problems with unfettered competition when there is a natural monopoly in that there is a basic tension between allocative efficiency, which requires sufficiently many firms to achieve an approximately competitive outcome, and productive efficiency, which requires a single firm.

Questions and Problems

1. Consider a single-product natural monopoly situation with the usual U-shaped long-run average cost curve. Is the range of output over which natural monopoly holds from zero to the output corresponding to minimum average cost? If not, explain how to determine the appropriate range. Use the total cost function $C(q) = 1 + q^2$ to answer this question.

2. Assume a natural monopoly with total costs $C = 500 + 20Q$. Market demand is $Q = 100 - P$.
   a. If price is set at marginal cost, what is the monopolist’s profit?
   b. The answer to part a implies that linear (or uniform) marginal cost pricing has a serious problem in natural monopoly situations. Suppose that average cost pricing is employed. Find price, output, and the deadweight loss compared to part a.
   c. Now consider two-part pricing, a type of nonlinear (or nonuniform) pricing. Each consumer must pay a fixed fee regardless of consumption level plus a price per unit. Assume that the market consists of ten consumers with identical demand curves for the product. If the price is set equal to marginal cost, what is the largest fixed fee that a consumer would pay for the right to buy at that price? What fixed fee would permit the monopolist to break even? What is the deadweight loss in this case?

3. Assume the same facts as in question 2, but assume that now there are six “rich” consumers with each having inverse demands: $p = 100 - 6.3q$; also, there are four “poor” consumers, each with demands: $p = 100 - 80q$.
   a. What is the largest fixed fee that a poor consumer would pay for the right to buy at marginal cost?
   b. Because the poor consumers would not be willing to pay the uniform fixed fee of $50 necessary for the monopolist to break even, the rich consumers would have to pay a fixed fee of $83.33. What is the deadweight loss in this case?
   c. Third-degree price discrimination could be a solution. That is, if it is legal, resales are not feasible, and consumers could be identified by the monopolist as being rich or poor, the monopolist could charge different fixed fees to the two consumer types. If the price per unit is still equal to marginal cost, what are two fixed fees that are feasible? In this case, what is the deadweight loss?
4. If third-degree price discrimination is not a feasible alternative in question 3c, consider the optimal two-part tariff. That is, what is desired is the two-part tariff that minimizes deadweight loss or that maximizes total surplus. One way to think about it is to imagine the case of a zero fixed fee and price equal to marginal cost. This causes a loss of $500 that must be covered. Imagine raising both the fixed fee and the price simultaneously. Both actions can cause losses, the fee by excluding poor consumers and the price by causing deadweight consumption losses. One possibility is to exclude poor consumers and go to solution 3b. The other possibility is to keep all consumers in the market; this implies that the fixed fee should equal the consumer surplus of a poor consumer. It is optimal to take all of the poor consumers’ surpluses as a fee. To see why, consider the opposite case, where the poor have some excess of surplus over the fee. Then the price could be lowered, reducing deadweight losses, and the surplus could be used to offset the reduction in revenues without excluding the poor from the market.

a. Find the sum of consumer and producer surplus minus the $500 fixed cost (that is, find total surplus) for case 3b where the poor are excluded.

b. Find total surplus for the case of all consumers retained in the market. Hint: An equation in $P$ can be defined that equates to $500 the total contributions to fixed cost (10 times the fixed fee, equal to the consumer surplus of a poor consumer, plus the revenues net of variable cost generated by consumption). Hence, what is the optimal two-part tariff where all are retained in the market?

c. Compare the efficiency of the tariffs in parts a and b.

5. A multipart tariff can be superior to the optimal two-part tariff found in question 4. A multipart tariff involves a fixed fee plus multiple prices per unit, which depend on predefined blocks of consumption.

a. Show that by making an additional two-part tariff available to the consumers that they can use at their option, the “two” two-part tariffs are Pareto superior to the optimal tariff in question 4 (that is, $F = 38.55, P = 21.50$). Let the optional two-part tariff be $P = 20.50$ and $F = 51$. These two two-part tariffs are equivalent to a multipart tariff that has a fixed fee of $38.55 and a price of $21.50 for the first 12.4 units and a price of $20.50 for all units above 12.4. Show this result by plotting the two tariffs on a graph that has total expenditure on the vertical axis and total units on the horizontal axis. The two straight lines representing the tariffs intersect at 12.4 units. Because consumers will always operate on the lowest line that they can attain to minimize expenditure, the multipart tariff is just the lower boundary (that is, the kinked line defined by $F = 38.55$ and the marginal prices of $21.50 for the first 12.4 units and $20.50 thereafter).

b. Demonstrate that the two two-part tariffs are Pareto superior to the optimal two-part tariff in question 4b. Note that the optional tariff will not change the poor consumers’ behavior at all. Why?

c. As a result, we can focus solely on the rich consumers and the monopoly. If both are made better off by the optional tariff and the poor are kept the same, then the optional tariff results in a Pareto improvement, which is a stronger welfare statement than simply saying one tariff yields a higher total surplus. (That is, if we focus on total surplus comparisons, we ignore the fact that some people may be made worse off even though total surplus is higher.) Find the consumer surplus of a rich consumer under the two-part tariff of question 4b.

d. Find the consumer surplus of a rich consumer under the multipart tariff.
e. Find the change in profit of the monopolist. Hence a movement from two-part tariffs to multipart tariffs clearly has the potential for gains in efficiency. The intuition is that the more the “parts,” the better the tariff can be tailored to the differences in willingness to pay across consumers.

6. Assume that a water distribution monopoly serves two consumer types, industrial and residential. The demands by the two classes are as follows. Industrial: \( Q_I = 30 - P_I \) and Residential: \( Q_R = 24 - P_R \). The company has no costs other than the fixed cost of the pipeline, which is $328. Find the Ramsey prices. Hint: See note 18.

7. Assume a natural monopoly with total cost \( 500 + 20Q \) facing a demand of \( Q = 100 - P \).

a. Find the price that enables the monopolist to break even. (This is the same problem as 2b.) Call this price \( P^* \).

b. Loeb and Magat show that if the monopolist is allowed to choose its own price and to have the regulatory agency subsidize the firm by an amount equal to consumer surplus at the selected price, the monopoly will select price equal to marginal cost. What is the price and amount of government subsidy?

c. Loeb and Magat also note that a bidding process for the monopoly franchise would enable the government to recover some of the subsidy. What is the amount recovered and what is the net subsidy after bidding?

d. An alternative proposal would make use of two-part tariffs. For example, assume that the current regulated price is \( P^* \). Now assume that the regulatory agency offers the firm the right to select any two-part tariff that it wishes as long as the consumer continues to have the option of buying at \( P^* \). (For simplicity, assume a single consumer.) What is the two-part tariff that the monopolist will choose, and what is its profit? What is the deadweight loss?

e. Assume that the government uses a bidding process to eliminate the monopoly profit in part d. The bid is in the form of a single price, like \( P^* \), that the consumer will always have as an option to the two-part tariff. That is, the same rules are in effect as in part d except that now the bidding is for the right to offer a two-part tariff optional to some \( P^* \) that the bidding will determine. What is the low bid?

f. Compare the Loeb and Magat proposal in part c with the proposal in part e. Do both proposals give efficient prices? Are there any substantive differences?
The theory of natural monopoly and alternative policy solutions were the main topics of the last chapter. The most common policy solution in the United States—regulation—is the subject of this chapter. In examining regulation, it is useful to keep in mind the benefits and costs. The benefit is to reduce deadweight losses in efficiency that would exist under unregulated monopoly. The costs are less obvious but include the direct costs of regulatory agencies, as well as unintended side effects of regulation. One such effect is higher costs because of weakened incentives for regulated firms to be efficient.

Our primary application in this chapter is the regulation of the electricity market. The technology for centralized production and distribution of electric power was first put into operation in September 1882 in New York City, where Thomas Edison began producing electricity in the famous Pearl Street plant of the Edison Electric Illuminating Company. During the early years the common method of regulation was by the award of a franchise by the town or city. The city would normally grant a franchise for exclusive operation within the city in return for an agreement by the firm that it would provide a certain quality of service at certain rates. As the technology changed and firms began to serve larger regions and even entire states, community regulation became ineffective. It also became more efficient to have regulatory experts at the state level only, rather than having duplicate experts in every community. Finally, the franchise agreement was not very flexible in dealing with constantly changing economic conditions. All these factors led to the institution of the regulatory commission at the state level. The first such commissions began in 1907 in New York, Wisconsin, and Georgia. All states have regulatory commissions today.

Most of the discussion in this chapter will be concerned with state regulatory commissions, but we should observe that several federal regulatory agencies are also involved in natural monopoly regulation. For example, the Federal Energy Regulatory Commission (FERC) regulates interstate wholesale transactions of electricity and natural gas pipelines.

The regulatory commission is usually appointed by the governor, although in some states the commissioners are elected or appointed by the legislature. A typical commission consists of three to twelve commissioners, who are assisted by a staff trained in accounting, economics, engineering, and law. A commission typically focuses on prices charged by the monopolist. Their mandate is usually somewhat vague, such as requiring that prices be “just and reasonable” and that there be no “undue discrimination.” Historically the procedure is that prices are set in rate cases and are generally fixed until the next case. Rate cases are similar to civil court cases. Expert witnesses on various topics are heard (for instance, economists testify on the cost of capital), with the final decision being made by the commissioners. In certain cases, appeal to higher courts is possible.

In the next section we review rate of return regulation, which is the traditional method for regulating a natural monopoly. It involves specifying an allowed rate of return for the regulated firm and then selecting prices that are expected to generate that rate of return. As to be discussed, this form of regulation creates minimal incentives for the firm to be efficient, since
any gains in profit are soon taken away by the regulatory agency’s lowering the firm’s permitted prices. Regulatory practices have since evolved to provide better incentives and include earnings sharing, price caps, and yardstick regulation. In discussing both traditional and more recent practices, the focus is on how the average price is set. However, there are a number of important issues related to the structure of prices, such as the allocation of common costs across different consumer types and the variation of price with patterns in demand (for example, electric energy usage is higher during the day than at night). These are covered in the sections entitled Rate Structure and Peak-Load Pricing. Finally, the chapter concludes with an investigation of the regulation of electric utilities and the experiences of partial deregulation beginning in the late 1990s. Of particular interest is the fiasco that emerged in the California electric power markets in 2000–2001.

Traditional Rate-of-Return Regulation

The essence of rate-of-return regulation lies in the following accounting equation:

\[
\sum_{i=1}^{n} p_i q_i = \text{Expenses} + sB
\]

(12.1)

where \( p_i \) is the price of the \( i \)th service, \( q_i \) is the quantity of the \( i \)th service, \( n \) is the number of services, \( s \) is the allowed or “fair” rate of return, and \( B \) is the rate base, a measure of the value of the regulated firm’s investment. The underlying idea, of course, is that the company’s revenues must equal its costs, so that economic profit is zero. Notice that economically efficient prices are not required by the equation, only prices that cover total costs.

Examining (12.1), one can see a myriad of tasks for a regulatory agency. First, it must decide on the allowed profit, which means setting an allowed rate of return, \( s \), and determining the capital on which this return can be earned, \( B \). Second, given this profit level, it needs to select prices that result in the firm earning that profit. If the regulated firm offered only one product, then the specification of its profit would typically imply a unique price. For example, if the firm is constrained to earning zero profit, then this would imply average cost pricing and thereby a unique price (see, for example, figure 11.8). However, all regulated monopolists offer a variety of services to different types of customers. As a result, there are many arrays of prices that will satisfy (12.1). In that case, the regulatory agency is faced with the nontrivial problem of choosing among them. In a later section, we will deal with some of the issues related to the structure of prices, but in this and the ensuing section on Incentive Regulation, we focus our attention on the overall rate or price level.
The Rate Case

Unless captured by the firm it is regulating, a regulatory agency will want to set the rate of return at the minimum level that maintains the firm’s financial viability and ensures it can raise funds to finance future investment. In the case of the regulation of interstate long-distance rates in the 1980s, the Federal Communications Commission (FCC) set rates commensurate with AT&T earning a rate of return of 12.2 percent.

Rate hearings are a quasi-judicial process in which evidence is put forth to determine the appropriate rate. While a regulatory commission can, on its own initiative, initiate a hearing, it is more typical for the firm to do so by applying for a rate increase. The process typically entails selecting a “test period,” which is usually the last accounting period. A company submits detailed financial exhibits to show, for example, that at current prices, its rate of return on its rate base for the test period is too low. Adjustments are also made to reflect anticipated changes in input prices or productivity due to the adoption of new technologies. Historically, the dialogue between the company and the agency has had the company arguing that its true cost of capital is such that it needs a higher rate of return—in other words, it should be allowed to raise prices—while the commission staff argues that the company’s requested rate of return is too high. Eventually, after much testimony, a value for $s$ is selected and prices are adjusted to yield the new rate of return. Examining (12.1), note that this requires knowledge of demand elasticities in order to predict the change in quantities from the change in prices.

Rate Base

Though the largest part of a rate case is devoted to the issue of what the proper return to investment should be, determination of the rate base is equally important since it is total profit, $sB$, that ultimately impacts a firm’s financial status and, through the ensuing prices, consumer welfare.

Many commissions have used an original cost method of valuing the rate base. Original cost valuation is simply the amount that the company originally paid for their plant and equipment, less depreciation. There can be little debate about the actual numbers in original cost valuation (aside, of course, from the issue of imprudent investment). Other valuation methods are much more subject to judgment calls. For example, valuing the rate base by reproduction cost means estimating the current cost of reproducing the plant, even though some of the plant may be twenty years old.

One concern about using original cost is that in periods of inflation, the reproduction cost rate base will exceed the original cost rate base. That is, an electric utility might have built much of its capacity at much lower plant prices twenty years ago. To reproduce the plants at current prices might cost five times original cost. And, because the economically correct prices should reflect current marginal costs, it might be thought that original cost leads to a
setting of prices that are too low. This is certainly a possibility if the commission determines price by simply dividing the sum of “Expenses” and $B$ by quantity.

However, in principle the commission is not bound by this method. Ideally, the prices should depend on current marginal costs, and if these prices yield total revenues that are too high or too low in terms of the rate-level “solution,” it is necessary only to adjust fixed fees in an appropriate manner. (Recall from chapter 11 that economic efficiency generally requires only that price per unit equal marginal costs; fixed fees, independent of output, can be adjusted to cover deficits or return additional revenues.)

There are other valuation methods that we should also mention. Replacement cost refers to what it would cost to replace the capacity with plants embodying the newest technology, as opposed to simply reconstructing the older technology plants at today’s prices. Another valuation method is simply to add up the value of all of the company’s outstanding stocks and bonds as given daily in the Wall Street Journal. This method has the defect that it is circular. The purpose of finding a rate base is so that the prices and returns can be determined, but the market-value method just described takes as the rate base a value that depends on prices and returns set by the commission in the past.

Finally, let us emphasize that not all investments are allowed to become part of the rate base. Commissions have taken a tough stance in disallowing those investments deemed irresponsible. Consider, for example, the following excerpt from an annual report of the California Public Utilities Commission:

In May 1987, the PUC Public Staff Division’s Diablo Canyon Team recommended that of the $5.518 billion that PG&E spent before commercial operation of Diablo Canyon Nuclear Power Plant, the utility should only be allowed to collect $1.150 billion in rates. . . . The Public Staff Division alleges that unreasonable management was to blame for a large part of this cost overrun.\(^1\)

Indeed, failed nuclear power plant investments have been a common source of contention. Recognizing that excessive capital investment may not enter into their rate base gives regulated firms at least some incentive to be efficient.

**Regulatory Lag**

Once a rate case is settled and new prices are set, they remain unchanged until the next rate case. Hence the period during which prices remain fixed provides an incentive for the company to be cost-efficient. The company is able to earn higher rates of return than allowed if it can reduce its costs, and, of course, it earns lower rates of return if its costs rise. This incentive for cost efficiency is often referred to as the result of regulatory lag. That is, if the

---

commission were somehow able to continuously adjust prices to keep the company’s rate of return always equal to $s$, there would be no lag and thus no incentive for cost efficiency.

Ironically, regulatory inefficiency in adjusting regulated prices enhances the firm’s productive efficiency. In our later discussion of incentive regulation, we will see how such efficiency-enhancing incentives are intentionally created through regulatory design rather than simply being the side effect of imperfect implementation of rate-of-return regulation.

Regulatory lag can obviously cause harm as well. Consumers forgo surplus when they are forced to wait for the lower prices coming from cost reductions, and firms are harmed when an upward trend in input prices depresses their rate of return. In fact, electric utilities consistently earned a rate of return below that which was allowed during most of the 1970s and 1980s. Regulatory-mandated rate increases consistently lagged the rise in cost associated with the rising price of oil and other energy inputs. This took a major toll on the common stock price of electric utilities.

**Averch-Johnson Effect**

We turn now to an example of how rate-of-return regulation can create perverse incentives. The model we describe is an analysis of rate-of-return regulation published in 1962 by Harvey Averch and Leland Johnson.\(^2\) Their work led to a large outpouring of both theoretical and empirical research. Using what some today regard as very strong assumptions about how regulation constrains the firm, it was shown that firms would choose too much capital relative to other inputs. As a result, output would be produced at an inefficiently high cost. The key idea is that because allowed profit varies directly with the rate base (capital), the firm will tend to substitute too much capital for other inputs. In mathematical terms, the problem is one of maximizing profit subject to a rate-of-return constraint. We will not develop the complete analysis, but it should be instructive to formulate the problem and provide the solution.

The problem is to choose the quantities of labor and capital to maximize profit—that is, revenue minus the costs of the inputs, labor and capital. Maximize

$$\pi = R(K, L) - wL - rK$$

subject to

$$\frac{R(K, L) - wL}{K} = s$$

where $\pi$ is profit, $R$ is revenue function, $K$ is quantity of capital, $L$ is quantity of labor, $w$ is wage rate, $r$ is cost of capital, and $s$ is allowed rate of return.

The rate-of-return constraint, equation 12.3, implies that the firm is continuously restricted
to a rate of return equal to \( s \). The numerator equals total revenue minus the cost of labor and
is divided by capital, which gives the rate of return on capital. This, of course, is not strictly
correct, because the firm’s prices are fixed from one rate case to the next, and therefore the
firm’s rate of return can be greater than or less than \( s \) during these periods of regulatory lag.

Another key assumption is that \( s > r \). In other words, it is assumed that the
regulatory agency permits the firm to earn a higher rate of return on capital than the true cost
of capital. Of course, the opposite case of \( s < r \) would imply that the firm would prefer
to shut down if this were to be a long-term situation. And if \( s = r \), the firm
would be indifferent among the quantities of \( K \) and \( L \), inasmuch as its profit would be zero
for all choices. Hence, Averch and Johnson argued that \( s > r \) is the interesting one.

Using a standard mathematical solution technique (the Lagrangian multiplier method), one
can show that

\[
\frac{MP_k}{MP_l} = \frac{r - \alpha}{w}
\]

where

\[
\alpha = \frac{\lambda(s - r)}{l - \lambda} > 0.
\]

\( MP_k \) is the marginal product of capital, and \( MP_l \) is the marginal product of labor. The \( \alpha \) vari-
able is positive because \( s - r > 0 \) and \( \lambda \), the so-called Lagrangian multiplier, can be shown
to be between 0 and 1. (The economic interpretation of \( \lambda \) is that it measures the increase in
profit of a $1 increase in allowed profit; hence its value between 0 and 1 is sensible.)

This result can be explained by reference to figure 12.1. The figure shows the isoquant for
the level of output chosen by the regulated firm, \( Q^* \). The axes show the quantities of the
inputs, \( K \) and \( L \), that can be used to produce output \( Q^* \). The economic theory of production
requires that to minimize the cost of producing \( Q^* \), it is necessary to equate the slope of the
isoquant, that is, the ratio of marginal products, to the ratio of input prices. Equation 12.4
implies that the Averch-Johnson regulated firm would meet this requirement if \( \alpha = 0 \).
However, \( \alpha > 0 \), and the firm acts as if the cost of capital is cheaper than it actually is. That
is, the firm acts as if its cost of capital is \( r - \alpha \).

In figure 12.1, let the slope of lines \( MM \) and \( NN \) equal \( r/w \) and the slope of line \( PP \) equal
\( (r - \alpha)/w \). Then, cost minimization requires operation at point \( E \), where the slope of \( NN \) equals
the slope of the isoquant. Note, however, that the regulated firm will choose to operate at
point \( F \), which equates the slope of \( PP \) [or \( (r - \alpha)/w \)] with the slope of the isoquant. The
result shows that the regulated firm uses too much capital, \( K^* \), and too little labor, \( L^* \), as
compared to the least-cost solution, \( K' \) and \( L' \). The excess cost can be measured in units of
Figure 12.1
The Averch-Johnson Effect versus Least-Cost Production

labor by the distance $MN$ on the vertical axis. That is, the actual cost of producing $Q^*$ is $OM$ units of labor inasmuch as $MM$ passes through point $F$. However, the least-cost production of $Q$ is $ON$ units of labor.

A less rigorous explanation is as follows. The key point is that the regulated firm perceives that its cost of capital, $r - \alpha$, is less than the true cost $r$. For simplicity, take $s = 10$ percent, $r = 8$ percent, and $r - \alpha = 6$ percent. The regulated firm can earn a “bonus” of 2 percent on each dollar of new capital (costing 8 percent) because it is allowed to earn 10 percent. This “bonus” of 2 percent per dollar can be interpreted roughly as a 2 percent discount, making its perceived cost of capital only 6 percent.

We should mention one possible beneficial effect of the Averch-Johnson bias toward capital intensity. For most regulated industries, technological change takes place through the substitution of capital for other inputs. For example, direct long-distance dialing replaced many human operators with automatic switching equipment. Hence one might argue that the Averch-Johnson effect has possibly stimulated innovation. Other characteristics of regulation, however, can be argued to retard innovation, so the net effect of regulation on innovation is
unclear. For example, profits created through innovation can be expected to be reduced through price decreases at the next rate case.

**Incentive Regulation**

One can think of rate-of-return regulation as cost-plus regulation in that the regulated firm is allowed to earn revenue that is some fixed amount above cost. If its cost changes, then its revenue is adjusted to hold fixed the rate of return. The major problem with this form of regulation is that it provides weak incentives for the firm to reduce cost or, more generally, adopt efficient practices. It is only by virtue of regulatory delay that these incentives are not entirely stifled. Furthermore, the lack of price flexibility between rate cases prevents the firm from efficiently responding to cost and demand shocks.

These various deficiencies to the traditional method of regulating price in natural monopolies have led to a series of innovations in regulatory policy. Referred to as incentive regulation, they are designed to create incentives for the regulated firm to lower cost, innovate, adopt efficient pricing practices, improve quality, and the like. As of 2001, twenty-eight electric utility companies in sixteen states were deploying incentive regulation of some type, and their adoption in telecommunications has been even more widespread.  

The common properties of these regulatory policies are giving the firm a certain degree of discretion in pricing and allowing them to share in profit increases. We will review the three primary forms of incentive regulation: earnings sharing, price caps, and yardstick regulation. Whereas various forms of incentive regulation have been used to regulate electric utilities for decades, they have been more aggressively applied in the telecommunications industry. Yardstick regulation is by far the least used of the three methods. Though it has some nice theoretical properties—in particular, achieving improvements in efficiency with minimal information requirements imposed on regulators—there are some challenges in implementing it. However, yardstick regulation has been used in conjunction with price caps.

**Performance Standards**

Before we launch into some full-fledged alternatives to rate-of-return regulation, we will note the various amendments that have been made to traditional regulatory practices to promote

---


enhanced efficiency. They provide financial rewards to achieving improvements in specific operating measures.

Here is an example that has been widely applied to electric utilities since the 1970s⁵:

These incentive payment programs take many forms and focus on different operating statistics: they reward utilities which experience high levels of base load generating unit utilization and availability, low heat rates (reflecting the efficient transformation of fuel into electricity), and keep fuel and purchased power costs below externally-determined indices.

A firm that has a heat rate below the specified level will be able to retain some or all of the cost savings. In the case of New York, the regulatory agency forecast the rise in fuel costs and allowed the electric utility to pass on to consumers only 60 to 80 percent of any costs above that forecast; in addition, the utility retained a similar fraction of any savings when costs fell below that forecast. This created incentives to keep fuel costs down, incentives that were lacking when there was an automatic fuel cost pass-through program.

Though a regulatory agency may not jettison rate-of-return regulation for the incentive-based methods we are about to describe, this example highlights how incentives can be created in a more focused and less encompassing way alongside traditional regulation. Interestingly, one study found that these more focused methods tended to result in greater gains in efficiency than earnings sharing schemes, which we review next.⁶

### Earnings Sharings

The basic problem with rate-of-return regulation is that the firm does not get to share in any of the cost savings: reductions in cost induce commensurate reductions in price so as to keep the firm’s rate of return fixed. The regulated firm then has no incentive to lower cost (except that which is created due to regulatory lag).

*Earnings sharing* (or sliding scale) regulation is based on the idea that if we want to induce regulated firms to reduce cost and engage in other efficiency-enhancing practices, we need to allow them to retain some of the gains that they create. However, if we let them retain all of the gains, then we are, in essence, leaving them unregulated. Earnings sharing regulation finds middle ground by having the firm and consumers share in any excess earnings.

As an example, consider the earnings sharing scheme that Pacific Bell faced in the 1990s in California. It could keep all profits if its rate of return was less than 13 percent. If the rate of return was between 13 percent and 16.5 percent, it could retain 50 percent of the profits in excess of the 13 percent return, with the remainder being rebated to its customers. Finally,

---


all profits in excess of a 16.5 percent return were to be rebated. This resulted in a maximum rate of return of 14.75 percent.

A general class of earnings sharings can be described by the following formula. Letting \( r \) denote the gross rate of return (that is, before netting out customers’ share of profit), a firm’s net rate of return is:

\[
\begin{align*}
    r & \quad \text{if } r \leq \underline{r} \\
    \underline{r} + \theta(r - \underline{r}) & \quad \text{if } \underline{r} \leq r \leq \bar{r} \\
    \bar{r} + \theta(\bar{r} - r) & \quad \text{if } \bar{r} \leq r
\end{align*}
\]

(12.5)

where \( \underline{r} < \bar{r} \) and \( 0 \leq \theta \leq 1 \). When its gross rate of return lies in the band ranging from \( \underline{r} \) to \( \bar{r} \), the regulated firm retains a fraction \( \theta \) of the excess profit. But its rate of return is capped at \( \underline{r} + \theta(\bar{r} - r) \) as it must rebate all excess profit once a gross rate of return of \( \bar{r} \) is reached. In the Pacific Bell example, \( \underline{r} = 0.13 \), \( \bar{r} = 0.165 \), and \( \theta = 0.5 \).

The higher is \( \theta \), the greater is the incentive for the firm to reduce cost and raise revenue since it is able to retain a higher fraction of the rise in profit. Of course, it also means that it is able to set higher prices, so one would not want to set \( \theta \) and \( \bar{r} \) too high or the entire reason for regulation would be mooted. Notice that this class of regulatory schemes includes some of the previous cases we have discussed. Traditional rate-of-return regulation is when \( \theta = 0 \) and \( \underline{r} \) is the allowed rate of return, while the case of an unregulated monopoly is when \( \theta = 1 \) and \( \bar{r} = \infty \).

**Example: San Diego Gas & Electric**

For a concrete example of incentive regulation, consider San Diego Gas & Electric in 1999.\(^7\) The earnings sharing scheme is depicted in figure 12.2 and has ten steps to it, a bit more complicated than what was described above. The authorized rate of return is 9.05 percent and there is a “deadband” of 0.25 percent so that SDG&E keeps all returns up to 9.30 percent. This is not so much an incentive device as a simplification in light of the unpredictability of profits. For a gross rate of return between 9.30 percent and 9.80 percent, SDG&E keeps 25 percent of the additional profit, and the share that it retains goes up with each step to the point that it keeps all additional profits once the gross rate of return is 12.05 percent or higher. This scheme then provides increasingly high-powered incentives as profits rise.

In addition to earnings sharings, the regulatory body provided rewards and penalties concerning performance with respect to employee safety, customer satisfaction, telephone response time, and system reliability that, in sum, could result in an annual penalty or reward of as much as $14.5 million. For example, the benchmark for telephone response time is that

\(^7\) This discussion is based on Richard Myers and Laura Lei Strain, “Electric and Gas Utility Performance Based Ratemaking Mechanisms” (Energy Division, California Public Utilities Commission, September 2000).
80 percent of calls are answered in 60 seconds and there is a reward (penalty) of $10,000 for each 0.1 percent increase (decrease) above (below) 80 percent, with a maximum reward or penalty of $1.5 million. Table 12.1 shows the results for 1999.

**Price Caps**

*Price caps* were first proposed in the early 1980s by Stephen Littlechild in connection with the regulation of British Telecom and Roy Radner at AT&T’s Bell Labs. They are based on the idea that if the price the firm can charge is independent of any cost reductions, then it knows that any such reductions will go to the firm’s bottom line. This provides powerful incentives to act efficiently and, in addition, gives the firm some flexibility in adjusting price. As opposed to earnings sharing, the constraint is on price rather than profit. The trick to price caps is setting price at an appropriate level, which requires forecasting future productivity gains.

A policy of price caps requires that the regulatory agency specify a maximum price, which is adjusted on a predetermined frequency according to a predetermined formula. This formula

![Graph](image-url)
is comprised of as many as three parts: (1) an inflation factor that controls for general price changes and changes in input prices, (2) an X factor that reflects anticipated increases in productivity (but, most important, does not depend on the actual realized changes in productivity), and (3) a Y factor that allows for pass-through of certain cost factors not controlled by the regulated firm. Given that a regulated firm has an array of services, the price cap is usually an average price over those services. The prices for the individual services are left more to the discretion of the firm, though they may also be subject to some individual price caps as well.

The biggest challenge is setting the X factor. The intent of the regulator is to set the X factor at the rate of productivity growth that would emerge if the firm was subject to competitive pressures. If the X factor is set too low, then prices are too high relative to cost, and this creates the usual deadweight welfare losses. If the X factor is set too high, then prices may be insufficient to cover cost and could induce financial difficulties for the firm. In forecasting future productivity growth, a regulatory agency may be inclined to use the historical rate of productivity growth, but this is problematic because the historical rate occurred under rate-of-return regulation, when the firm had little incentive to enhance productivity. In setting the X factor, regulators have then added a “stretch factor” to the historical rate, which is estimated to be the gain in productivity growth from having price caps.

As an example, the FCC used rate-of-return regulation in the interstate telecommunications market until March 1989, when it approved the use of price caps. These price caps required AT&T to annually reduce its rates by a productivity factor of 2.5 percent and a “consumer dividend” of 0.5 percent (after adjusting for inflation), so the X factor was 3 percent. AT&T’s services were grouped into three baskets—residential toll, 800 service, and all other business services—and AT&T could not increase or decrease the prices of any of the three baskets by more than 5 percent. 8

Price caps have increasingly been used in state telecommunications regulation, as shown in table 12.2. Interestingly, earnings sharing increasingly replaced rate-of-return regulation

---

8. That it was constrained in how much it could lower price was due to the presence of unregulated competitors and the concern that AT&T might price predatorily so as to drive them out of the market. Predatory pricing is covered in chapter 9.
in the early 1990s, only to then be replaced with price caps. A number of studies show that price caps are associated with lower prices. One such study found that intrastate long-distance rates fell more in states that instituted price caps than in those states that persisted with rate-of-return regulation.9

For price caps to create the desired efficiency-enhancing incentives, proper implementation is crucial. If the regulatory agency lowers the price cap every time the company’s profit rises, then price caps are no different than rate-of-return regulation. It is essential that the time path of the price cap be independent of the firm’s actual realized cost, so that efforts by the firm to lower cost do not translate into a commensurately lower price. The tendency for superior performance to be penalized with higher standards is known in the regulatory literature as the “ratchet effect.” For price caps to work, regulators must resist it.

Though we present earnings sharing and price caps as distinct approaches, there have been hybrids as well. For example, in the mid-1990s, the FCC offered the Regional Bell Operating Companies (RBOCs), the local telephone companies created by the breakup of the BELL System in 1984, the following three options in its setting of access rates (which are the prices

---

9. Alan D. Mathios and Robert P. Rogers, “The Impact of Alternative Forms of State Regulation of AT&T on Direct-Dial, Long-Distance Telephone Rates,” *RAND Journal of Economics* 20 (Autumn 1989): 437–53. It is unclear, however, to what extent these studies are measuring the effect of price caps or instead a rise in competition which may have induced the state regulatory agency to institute price caps.
charged to long-distance suppliers for connecting to the local network). The first option was price caps on, with an X factor of 4 percent and limited earning sharings in that the RBOC would receive all earnings up to a maximum of 12.25 percent, then partial earnings thereafter, with a maximum rate of return of 13.25 percent. The second option was price caps with a higher X factor of 4.7 percent but with expanded earnings sharing, so that the RBOC could earn as high as 15.25 percent. The third option was yet more aggressive in that it was pure price caps (hence, no maximum rate of return) but with an X factor of 5.3 percent.

Yardstick Regulation

All of the regulatory policies reviewed thus far require the regulatory agency to have a fair amount of information, whether it is about market demand, cost, or future productivity increases. But suppose the regulatory agency is broadly lacking such information. How can a reasonably efficient regulatory policy be designed and implemented? If there are regulated firms serving distinct markets—for example, electric utilities in different geographic areas—then a regulator can use information on the performance of these other regulated firms to lead to an efficient solution in any particular market, and this can be done for all markets. This is known as yardstick regulation, because the other regulated firms’ prices and performance are used as a yardstick or benchmark to evaluate the performance of an individual regulated firm.10

Suppose a regulator in Illinois selects electric utilities in other states that face production and demand functions comparable to those of firm X in Illinois. The regulator determines the average cost per megawatt-hour for all comparable firms, say AC, and then sets firm X’s price equal to AC. If regulators in other states followed the same procedure, then this approach would eliminate the cost-plus character of regulation. Each firm’s prices would be completely independent of its own costs, and cost reductions would lead to profit increases. Note that if this scheme induces all of the other regulated firms to be aggressive in reducing cost, then a regulated firm that fails to do so will have relatively bad performance and will thereby be penalized.

What has prevented widespread use of yardstick regulation is the difficulty of finding truly comparable utilities. As stated by Paul Joskow and Richard Schmalensee,11

Utilities differ from one another in so many dimensions, not only because of current market conditions but also because of past investment decisions, that we are unlikely to find a large number of truly comparable utilities.

Rate Structure

The rate structure has to do with how prices vary across customer classes and products. In the preceding chapter we described the prices that are economically efficient under various conditions. Of course, economically efficient prices (for instance, prices equal to marginal costs) are often not the prices set under regulation. However, peak-load pricing (a type of marginal cost pricing) has become important in electric power. We will examine this topic in depth later in this chapter.

A common method of pricing used by regulatory commissions is to begin by allocating all of the utility’s costs to various customer classes and services. Most utilities provide a variety of services to different customer groups. They also have many facilities that are used in common by these customers and services. For example, power plants and transmission lines, telephone switching centers, and pipelines all represent common costs that apply to most customer classes or services.

FDC Pricing

To illustrate concretely one such pricing method, fully distributed cost (FDC) pricing, consider a simple two-product natural monopolist that sells electricity to two classes of customers. We denote the electricity sold to residential buyers by $X$ and to industrial customers by $Y$. (Electricity sold to residential customers is usually at a lower voltage than industrial customers require, and therefore there truly are two different products.)

Assume the following cost functions:

To produce $X$ alone: \[ C_x = 700 + 20X. \] (12.5)

To produce $Y$ alone: \[ C_y = 600 + 20Y. \] (12.6)

To produce both: \[ C_{xy} = 1,050 + 20X + 20Y. \] (12.7)

Note that the joint production of $X$ and $Y$ is subadditive. That is, least-cost production requires that $X$ and $Y$ be produced together because fixed cost is $1,050, as compared to a total of $1,300 if produced separately. The $1,050 fixed cost also represents common costs that must be allocated to each product in order to implement FDC pricing.

It has been observed that utilities’ common costs “may be distributed on the basis of some common physical measure of utilization, such as minutes, circuit-miles, message-minute-miles, gross-ton-miles, cubic feet, or kilowatt-hours employed or consumed by each. Or they may be distributed in proportion to the costs that can be directly assigned to the various services.”

The particular method may appear quite reasonable, but the essential point is that it is necessarily arbitrary. And, more important, such cost allocations lead to prices that have no necessary relationship to marginal costs. To take an example, assume that some “reasonable” method leads to an allocation of 75 percent of the common costs to product $X$ and 25 percent to product $Y$. Hence, FDC average costs would be

$$AC_x = \frac{787.5}{X} + 20 \quad \text{and} \quad AC_y = \frac{262.5}{Y} + 20.$$ \hspace{1cm} (12.8)

That is, the average cost of $X$ equals its 75 percent share of the $1,050 common cost, divided by the units of $X$ sold, plus the clearly attributable variable cost of $X$ per unit of $20$.

At this point, the demands for the two products must be specified. Assume that the demand functions are

$$P_x = 100 - X \quad \text{and} \quad P_y = 60 - 0.5Y. \quad \hspace{1cm} (12.9)$$

With the demand information, the actual FDC prices can be found by equating equations 12.9 and 12.10. This method simply sets $P_x = AC_x$ and $P_y = AC_y$, ensuring that total revenues are equal to total costs. The result is

$$P_x = AC_x = 31.5 \quad \text{and} \quad P_y = AC_y = 23.6$$

$$X = 68.5 \quad \text{and} \quad Y = 72.8. \quad \hspace{1cm} (12.10)$$

Hence these FDC prices clearly satisfy the requirement that total revenues equal total costs. Again, however, there is no basis for expecting these prices to be the economically efficient prices. In general, such prices lead to deadweight losses.

It is easy to show that the efficient prices in this case are the Ramsey prices (as explained in chapter 11):

$$P_x = 30 \quad \text{and} \quad P_y = 25$$

$$X = 70 \quad \text{and} \quad Y = 70. \quad \hspace{1cm} (12.11)$$

Figure 12.3 illustrates the Ramsey solution. By definition, the Ramsey prices have the smallest deadweight loss triangles (shaded in figure 12.3) for all possible pairs that yield revenues equal to costs.

Although we have implicitly assumed that two-part pricing is not feasible in the example, notice that if it were feasible, the “first best” solution would be to charge marginal cost prices to each group; that is, each price would be $20. Then the fixed fees would be set to just cover the fixed cost of $1,050. And in this case, in achieving economic efficiency it would make no difference how the $1,050 was allocated between the two groups. (This point was also explained in chapter 11.)

One problem that FDC pricing raises is that “reasonable” allocations of common costs lead to disputes among customer classes. It is natural to expect residential customers to argue that
their share should be lower than 75 percent and for industrial customers to argue that their share should be lower than 25 percent. Commissions have long been concerned with “undue discrimination” across customer classes, and we turn to that subject next.

**Undue Discrimination**

Undue discrimination is really not an efficiency issue; rather, it has to do with the fairness of the existing set of prices in the sense of whether one group may be “subsidizing” another group. It is clearly a controversial issue for commissions, inasmuch as a rate case may find intervenors representing residential customers and industrial customers in total opposition to each other. Both groups may argue that they are paying too large a share of the common costs! This is basically an equity issue, which we have avoided for the most part in this book, but there are some clarifications that economic analysis can offer to the debate.

Economists argue that if one must examine cross-subsidization issues (assuming Ramsey prices are not used), the most logical tests are the so-called stand-alone average cost and the average incremental cost tests—which are, in fact, equivalent.¹³

Consider the stand-alone average costs for product X for an output of X of 70 units. That is, returning to equation 12.5, the average cost of X is $30. The Ramsey price of $30 for the same output therefore does not give an incentive for customers of X to break away and produce X alone. This test therefore would classify the Ramsey price as subsidy-free. Similarly, the Ramsey price of Y is also subsidy-free. (Note that the FDC price of X for 68.5 units is $350.)

---

units is $31.5, which exceeds the stand-alone average cost of $30.2. It therefore fails the subsidy-free test.)

Another test is the average incremental cost test. Here, we compute the average incremental cost, AIC, of producing \( X \) in joint production with \( Y \). Thus, subtract the cost of producing \( Y \) alone from the cost of producing \( X \) and \( Y \) jointly to get the incremental cost of \( X \). The AIC of \( X \) is therefore \( 20 + 450/X \), or $26.4 for \( X = 70 \). Similarly, the AIC for \( Y \) at \( Y = 70 \) is $25. Here, the test for subsidy-free prices is that the prices equal or exceed their respective AICs. The logic is that if each product contributes to total revenue an amount that at least covers the extra costs it causes (when added to the production of the other products), then it should be viewed as a beneficial addition. To the extent that its incremental revenues exceed its incremental costs, the revenues required from the other products are reduced. The Ramsey prices of $30 and $25 also pass this test. In fact, these two tests always give the same answers. (Note that the FDC price of \( Y \) for 72.8 units is $23.6 while the AIC of \( Y \) is $24.8. Hence, the FDC prices fail the subsidy-free test by this test too.) It should be pointed out that FDC prices do not necessarily fail subsidy-free tests. They may pass the subsidy-free tests and still be economically inefficient.

Under certain conditions of subadditivity of cost, it is true that Ramsey prices are subsidy-free in the sense that no outsider would find it profitable to enter. Hence the regulator need not be concerned about whether permitting entry would be socially beneficial. This argument assumes, of course, that the regulator permits the monopolist to charge Ramsey prices (rather than hold it to the FDC prices given previously). However, there are cases in which the cost function is subadditive and yet subsidy-free prices do not exist. This is the case of a natural monopoly that is unsustainable—least-cost production requires a single firm, but there are no prices that can keep all of the monopolist’s products invulnerable to entry.

Here is an example. Three towns are in need of a well for water supply. One deep well could supply all three towns at a cost of $660. This is the least-cost solution and would imply a price of $660/3 = $220 per town. Two towns could go together and dig a shallower well for $400, and each town alone could dig an even shallower well for $300. Clearly, $660 is lower than any of the alternatives. If each town had its own well, the total would be $900, and if two went in together for $400 and the third went alone at $300, the total would be $700.

The problem is that $220 per town would provide an incentive for two of the towns to join forces at $400, or a price of $200 each. One can think of it as (any) two towns, if they go along with the three-town project, subsidizing the third town in the amount of $20 each. Clearly, there is no way to avoid the subsidies if one is going to achieve least-cost production.

In anticipation of the next section on peak-load pricing, we provide a final example of “unfair” subsidization. Consider two groups of electricity customers, day customers and night
customers. A plant costing $K$ dollars is necessary to meet the day customers’ demand, which is larger than the night demand. Because electricity cannot be easily stored, the plant must be large enough to supply power on demand.

The plant, of course, can be used to supply the smaller night demand as well. Under certain assumptions to be discussed in the next section, the economically efficient solution is to charge day customers for the total cost of the plant, $K$. Even though both groups use the plant, the day customers pay for the entire plant cost, plus fuel costs for their output. Night customers pay only for their fuel costs. It is certainly “unfair” in certain senses; however, it is the demand of day customers that necessitates such a large plant, and it would be inefficient if the price they paid did not signal the cost of the larger capacity.

In conclusion, the major point of this section is to make clear that the objective of economic efficiency may sometimes require pricing that conflicts with common notions of fairness. One justification for opting for efficiency, of course, is that the “size of the pie” is larger as a result, and authorities can in principle make everyone better off by appropriate taxes and/or subsidies.

**Peak-Load Pricing**

A major development in electric power was the implementation of peak-load pricing. This term refers to the variation in prices by time of use—for example, in the middle of the day more electricity is demanded than in the middle of the night. The marginal cost of electricity is, as a result, much higher in the middle of the day than it is at night. Setting prices that vary over the day in proportion to the variation in marginal costs is a form of peak-load pricing.

**Costs of Power Production**

It is useful to examine the cost structure of a typical electric power system before continuing with our discussion of peak-load pricing principles. A major point is that it is generally too costly (or impossible) to store electricity, and therefore sufficient capacity must be on hand to supply the demand at all times. This fact implies that capacity is determined by the amount of peak demand.

Demand for electric power typically varies in a reasonably predictable cyclical pattern—daily, weekly, monthly, and seasonally. The demand might follow the pattern in figure 12.4 for a typical weekday, with peak demand occurring at midmorning and demand at midnight only 70 percent of that. Demand over the weekend might equal only 50 percent of the high during the week.

A typical electric power system has a mixture of plant types because it leads to a lower overall cost of supplying the variable pattern of demand. Nuclear plants have relatively low
variable or “running” costs but have relatively high fixed (capital) costs. They are, therefore, suited for running as the “base load” plants—as many hours per year as possible. The extent of nuclear power in the United States is limited, so that coal-fired plants are important for providing the base load; they have higher marginal cost but lower fixed costs than nuclear plants. Combustion turbines, by comparison, have relatively high running costs but low fixed costs. They are used to meet peak demands that last for only a small number of hours per year.

The result is that the short-run marginal cost curve of a power system is similar to the rising curve shown in figure 12.5. The costs given by segment $AB$ might represent the base load nuclear plants’ running costs; $BC$, the costs of coal-fired plants of varying ages and efficiency; and $CD$, the costs of the peaking plants (such as combustion turbines). In this context, it is easy to realize that, since demand varies continuously over time, charging a price equal to short-run marginal cost (SRMC) would require a continuously changing price.
In order to explain the principles of peak-load pricing most clearly, it will help to abstract greatly from the real-world complexity just described. Hence we turn now to a vastly simplified model.\(^{14}\)

### Peak-Load Pricing Model

In figure 12.6 we make the assumption that demand is given by the peak demand curve for exactly half of the day, and by the off-peak demand curve for the other half of the day. For simplicity, it is assumed that the two demands are independent—the price in the peak period, for instance, does not affect the quantity demanded in the off-peak period.\(^{15}\) Also, it is assumed that “running” costs—for example, fuel for electricity production—are constant at the level \(b\) until capacity is reached at \(K\). (As we will explain, we have chosen \(K\) because it is the socially optimal plant size for the demands shown in the figure.) At the output \(K\), no further output is possible, as indicated by the vertical line that is labeled \(SRMC\). Hence we have a so-called “rigid” plant, with the \(SRMC\) curve being equal to \(b\) for outputs less than \(K\), and

---


15. This assumption is clearly too strong, as consumers can shift their demand across time. For example, when AT&T began to charge lower rates for long-distance telephone calls after 5 P.M., they found themselves deluged with calls from people who formerly called during the day. This interdependence of demands can be handled with an increase in mathematical complexity.
then becoming vertical at the plant capacity. One can think of this as an approximation to the smoothly increasing $SRMC$ curve in figure 12.5.

The dashed horizontal line at the level $b + \beta$ is labeled long-run marginal cost ($LRMC$). The assumption here is that $\beta$ represents the cost of an additional unit of capacity, and that it is possible to add capacity in increments of single units if desired. The economically efficient solution is to charge a price equal to $SRMC$ in order to use the existing plant optimally. The $LRMC$ comes into play in order to decide whether the existing plant capacity is optimal. Hence, in figure 12.6 the off-peak price should be $b$ and the peak price should be $b + \beta$. Notice that the peak price is equal to both $SRMC$ and $LRMC$. This fact indicates that the capacity is in fact optimal. The reason is that the price can be interpreted as the “marginal willingness-to-pay,” and $b + \beta$ represents the marginal cost to supply one more unit.

If, for example, the peak price exceeded $b + \beta$, it would pay society to increase capacity. This situation is shown by the demand curve labeled new peak. This new peak demand intersects $SRMC$ at a price higher than $b + \beta$. Therefore, an increase in consumer surplus can be had by expanding capacity out to $K^*$. This increase in consumer surplus is equal to the area

16. The production function underlying this model is one of fixed coefficients. Much of the early literature employed this assumption. While the suitability of this assumption to describe, say, electric power production is an empirical question, the alternative variable-proportions technology is somewhat more difficult to expound.
under the demand curve between $K$ and $K^*$—which represents willingness-to-pay—minus the cost of supplying the additional output, rectangle $EFK^*K$. Subtracting the cost from willingness-to-pay gives the shaded triangle that represents the increase in consumer surplus attributable to the capacity increase. Hence, at the new capacity $K^*$, price is again equal to $SRMC$ and $LRMC$, which indicates that $K^*$ is the optimal capacity.

Now, assume that the electric utility follows the practice of charging a single price that does not vary over the day, say, a price of $P^*$. This situation is shown in figure 12.7. As we mentioned earlier, this would represent the pricing policy generally followed in the United States before peak-load pricing began to be implemented. In order to satisfy demand at the peak at this price, capacity of $K_0$ is required. Because optimal capacity is $K$, where price equals $LRMC$, the single-price policy leads to too much capacity. The deadweight loss associated with this is shown as the shaded triangle $EFG$. It equals the difference between the cost of the excess capacity, rectangle $EFK_0K$, and the willingness-to-pay for that incremental capacity, $EGK_0K$. Intuitively, the peak demanders are not charged enough for the actual costs that they cause.

There is a second deadweight loss triangle in figure 12.7, and it is associated with the nonoptimal use of the plant in the off-peak period. That is, with the price $P^*$ charged in the off-peak period, consumption in the off-peak period is too low—at $Q_0$ rather than at $Q$, where price would equal $SRMC$. 

---

**Figure 12.7**

Deadweight Losses Due to Nonpeak Pricing
We should observe that the economically efficient prices that we have been discussing are what society would prefer, not necessarily what a profit-maximizing regulated utility would choose. In fact, the utility would have total revenues equal to total costs in either case (peak-load prices or single price). What gave the impetus to regulated electric utilities moving toward peak-load pricing was probably a combination of pressures—the energy crisis of the 1970s, high inflation, and other factors that made regulators seek alternatives to the traditional rate structures. In addition, Congress passed a law in 1978, the Public Utilities Regulatory Policies Act, that among other things required state commissions to study peak-load pricing for possible implementation in their state.

The cases of peak-load pricing described so far indicate that peak demanders pay $b + \beta$ whereas off-peak demanders pay only $b$. That is, peak demanders pay all capacity costs and off-peak pay none. This statement is true, however, only for the particular case shown, known as the firm peak case. An alternative case is the shifting-peak case, which has the property that the demands are “closer” together. Figure 12.8 illustrates this case.

To see why this case is known as the shifting-peak case, consider the effect of charging the peak demanders all the capacity costs and the off-peak demanders none. For simplicity,
we assume that $b = 0$ in this case; this assumption makes the figure less cluttered and does not affect the key points. The result is easily seen in figure 12.8. Peak demanders would demand $R$ units of capacity (where price $= \beta$) and off-peak demanders would want $S$ units (where price $= 0$), or a greater capacity than peak demanders! Intuitively, this result suggests that the prices are wrong. The correct set of prices can be found in conjunction with solving for the optimal capacity.

To obtain the optimal capacity, construct the demand for capacity. Conceptually, think of the plant as a public good: it can be used by both peak and off-peak demanders (though at different times). This statement implies that the total willingness-to-pay for the plant is obtained by adding vertically the demand curves for the two groups of demanders. This total willingness-to-pay curve, the kinked curve $ABC$, is the demand for capacity. For example, at the output $K$ the marginal willingness-to-pay is $P_p$ by peakers and $P_o$ by off-peakers, for a total of $\beta$. Because the capacity cost is $\beta$, $K$ is the optimal capacity. The efficient prices for using this capacity are $P_p$ and $P_o$, which, of course, add to $\beta$. Hence, in this case, the two groups share the capacity costs, unlike the firm-peak case. (If we had a nonzero $b$, the prices in each period would also include $b$.)

---

**Regulation and Restructuring of Electric Power**

**Historical, Technological, and Regulatory Background**

There are four primary functions in the provision of electric power: generation, transmission, distribution, and retailing. It starts with the generation of electricity through various production processes based on fossil fuels, nuclear fuel, and falling water. Once produced, transmission occurs, which refers to the high-voltage movement of electricity between generating sites and distribution centers by means of wires, transformers, and substation facilities. The next step is distribution, which is the transportation of electricity to residences and businesses at relatively low voltages using wires and transformers. These three functions are depicted in figure 12.9. Alongside these functions is retailing, or the making of arrangements for supplies of power from generators. It includes metering, billing, and demand management services such as marketing.¹⁷

The traditional organization of the U.S. electric power industry is one of vertical integration in which a single regulated firm provides all four of these functions. Historically, an investor-owned electric utility (IOU) would own generating plants, transmission lines,
the local distribution network and would produce power for delivery to final customers. These IOUs had an exclusive regulatory franchise to provide service to retail customers within their territories. In return, their rates were regulated by state regulatory commissions using rate-of-return regulation, and they were obligated to provide reliable service at those rates to all customers.

Though state public utility commissions have been the regulatory body controlling rates and evaluating proposed investments in generation, transmission, and distribution facilities, a key role is also played by the federal government. The Federal Energy Regulatory Commission (FERC) has responsibilities over wholesale power transactions—which historically meant the sale of power between utilities—and interstate transmission of power (at least that which is not internal to a utility).

Research has generally found that regulation did make a difference in constraining prices. One study used data for 1969 and 1974 to answer the question, to what extent do the prices charged by regulated firms differ from what unregulated profit-maximizing firms would charge?\(^\text{18}\) They found that the regulated prices were significantly lower, but that even lower

---

**Figure 12.9**
A Simple Electric System

---

prices would have been preferred. Similar work found that unregulated prices for electricity are 20 to 50 percent higher than actual regulated prices.19

Overview of Recent Legislation

While the most substantive regulatory restructuring in recent years has occurred at the state level, none of these changes would have been possible had it not been for essential changes in federal regulatory policy. A key restructuring role was played by the Public Utilities Regulatory Policy Act of 1978 (PURPA), which gave “qualifying facilities” the right to sell power to vertically integrated utilities. This led to a sizable increase in the number of non-utility power generators (that is, companies that produce power but do not distribute it to final consumers).

Though PURPA demonstrated the viability of competitive entry in the power generation business, it had two important limitations. First, it did not allow these nonutility power generators to contract directly with customers. Second, it did not permit them to sell outside the service territory of their host utility. The Energy Policy Act of 1992 repealed the second limitation but maintained the first. Since these lines could be an essential bottleneck constraining wholesale competition, it granted FERC the authority to order vertically integrated utilities to transmit power for others over regional transmission line.

The third key step occurred in 1986 with FERC Order 888, which mandated that owners of regional transmission networks act as common carriers of electric power. This meant providing interconnection service between independent power producers and wholesale buyers on the same terms and conditions with which it provides such service to itself. These changes laid the groundwork for restructuring of state regulation starting in the late 1990s.

While details will be offered when we examine the case of California, the common element in new regulatory regimes across states and countries is vertical unbundling. Rather than have one vertically integrated provider of electricity, the four functions are unbundled so that different firms can provide these services. By the customer choice (or retail wheeling) model, a retail customer accesses the wholesale power market directly and purchases unbundled distribution and transmission services from their local utility to deliver power. Such is the approach pursued in England and Wales, as well as California. Key to this approach is the emergence of competitive wholesale power markets, for which federal deregulation proved essential. At the same time, regulation persists over the transmission and distribution of power. This reflects that transmission and distribution continue to have strong economies of scale, while there have not been substantive scale economies in generation for some time.

Restructuring in California\textsuperscript{20}

The debates over the design of these wholesale market institutions in California in 1996 and 1997... were contentious and highly politicized, reflecting perceptions by various interest groups about how different wholesale market institutions would advance or constrain their interest and, in my view, an inadequate amount of humility regarding uncertainties about the performance attributes of different institutional arrangements. The discussion of alternative institutions was polluted by an unfortunate overtone of ideological rhetoric that attempted to characterize the debate about wholesale market institutions as one between “central planners” and “free market” advocates. The market design process in California in 1997 and 1998 also demonstrates how market design by committee, reflecting interest group compromises and mixing and matching pieces of different market models, can lead to a system with the worst attributes of all of them.\textsuperscript{21}

Movements towards restructuring started in such states as California, Massachusetts, and New York, where, not coincidentally, regulated rates were well above the cost of generating power using the best available technology. Restructuring efforts then appeared to occur first where the potential for lower rates was greatest from more efficient generators being allowed to sell directly to consumers.\textsuperscript{22}

After several years of discussion, the California state legislature approved a four-year transition plan that went into effect on March 31, 1998. Under this plan, consumers continued to have the option of buying electricity from their existing utility distribution company (UDC), which would either have been Pacific Gas & Electric, Edison, or San Diego Gas & Electric. For up to four years, the rate was capped at 90 percent of the regulated retail rate during 1996. Initially, this rate was well above wholesale prices for power. The new alternative was that consumers could buy from a nonutility electric service provider (ESP).

A second key element of the plan was imposing certain restrictions on the UDCs. Their role in power generation was reduced, with an increased emphasis on power procurement in the wholesale market for sale to final consumers. Towards this end, they were ordered to divest at least half of their fossil generating capacity, and they ultimately chose to divest it all. They did retain their nuclear plants, hydroelectric plants, and existing long-term power contracts. This divestiture dimension to restructuring varied across states, as some chose to permit UDCs to retain most of their generating assets and move them into separate unregulated wholesale power affiliates. As full divestiture served to contribute to the fiasco of


2000–2001, this latter approach has grown in popularity. Another obligation of UDCs was that they must meet all demand not supplied by the ESPs.

The restructuring created two institutional structures whose purpose was to run the energy markets. The California Power Exchange (CALPX) ran the public wholesale markets for energy that was to be delivered in the next hour or next day. This entailed conducting an auction that determined a wholesale price paid by all purchasers. Independent power producers could either sell their power through the CALPX or directly to consumers, but the UDCs were required to sell their power to the CALPX. Furthermore, during the transition phase, they had to buy all their power from the CALPX in the day-ahead and hour-ahead market for reselling to their customers. These restrictions were put in place to limit the market power of the UDCs, though, ironically, time would show that market power rested with the power generators.

A second institution was the California Independent System Operator (CAISO), whose objective was to provide nondiscriminatory transmission access to all generation facilities and to manage transmission congestion. Though owned by the UDCs, CAISO was independent so as to avoid the UDCs being given preferential treatment in the transmission of power.

To see what this vertical unbundling meant for a retail consumer’s bill, consider the sample bill in table 12.3. It includes the wholesale price of power of $14.40 and the cost of getting the power from the generator to the customer of $23.00 (transmission plus distribution charges). There are a number of other fees, but the one worth noting is the competition transition charge, which was intended to compensate the UDCs for past investments not covered by retail rates.

<p>| Table 12.3  |</p>
<table>
<thead>
<tr>
<th>Sample Electricity Bill (California, 1998)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bill Item</strong></td>
</tr>
<tr>
<td>CALPX (energy charge)</td>
</tr>
<tr>
<td>Transmission charges</td>
</tr>
<tr>
<td>Distribution charges</td>
</tr>
<tr>
<td>Public purpose programs</td>
</tr>
<tr>
<td>State regulatory fee</td>
</tr>
<tr>
<td>Nuclear decommissioning</td>
</tr>
<tr>
<td>Trust transfer amount</td>
</tr>
<tr>
<td>Competition transition charge</td>
</tr>
<tr>
<td>10 percent rate reduction</td>
</tr>
<tr>
<td><strong>Total</strong></td>
</tr>
</tbody>
</table>

California Energy Crisis, 2000–2001

With the exception of some very short-lived price spikes in the wholesale power market, the new regulatory structure initially worked reasonably well. Such price spikes were not uncommon in other wholesale electricity markets, such as the Pennsylvania–New Jersey–Maryland market and the New England market. What proved to be unusual about California was a sustained period of time during which the price remained very high. When combined with certain regulatory design flaws, the end result was a financial disaster that caused a retraction of deregulation.

Problems began to emerge in June 2000, when the wholesale price of power on the CALPX rose sharply and would continue to be high for almost a year (see figure 12.10). At the same time, retail prices were capped at $65 per megawatt-hour (mwh) which resulted in UDCs losing on the order of $50 million per day. FERC, which has the authority to set “just and reasonable” wholesale electricity prices, did not effectively intervene. The UDCs were becoming insolvent and stopped paying for power. Power producers curtailed supplies out of fear they would not be paid, but the U.S. Department of Energy stepped in to mandate their

Figure 12.10

supply lest the “lights be turned off.” The CALPX stopped operating and subsequently went bankrupt. After suspending the retail competition program, the state of California subsumed market functions, including power procurement and regulating retail prices so as to recoup the cost of the power they purchased during the crisis. Retail consumers are now paying about 40 percent more than before restructuring.

What led to the disastrous retreat of deregulation? The first step in addressing this question is to understand why there was such a large rise in the wholesale price of power. Several forces were at work. First, there was the rising cost of producing electric power due to a significant increase in the prices of natural gas and nitrous oxide emission credits (the latter gives power producers the right to pollute). Then there was a large and unanticipated increase in electricity demand in the late 1990s. Third, imports, in particular hydroelectric power from the Pacific Northwest, were significantly less. A fourth factor is the wielding of market power by wholesale power generators, which may have deliberately reduced available capacity so as to produce a price spike.

In recent years, capacity investment failed to keep up with the rise in demand and, to compound matters, an abnormally high amount of existing capacity was offline, about 35 percent, which was twice the historical average. While a power generator has to periodically take capacity offline for maintenance, the evidence suggests that some of this capacity was offline so as to create a shortage.

Figure 12.11 depicts both the capacity restrictions, as reflected in a sharply rising marginal cost after some quantity, and the inelastic demand in the hour-ahead and day-ahead spot markets. When demand is low (see $D'$), price is low and reflects the low marginal cost of another megawatt-hour. But when demand is high (see $D''$), price rises sharply as consumers’ willingness to pay, not marginal cost, is driving price up. One study attributed about one-third of the rise in wholesale prices over July to September 2000 to market power.

From a regulatory design perspective, there are at least two points of concern. First, we may point to the particular features of partial deregulation, which created incongruities in the system. Second is the difficulty of avoiding periodic episodes of market power. The UDCs faced a severe financial squeeze because of the unholy mix of unregulated wholesale power prices and retail price caps.

Not only were the UDCs losing hundreds of dollars per megawatt-hour but, with retail prices fixed at $65/mwh, consumers had no incentive to reduce their demand. Furthermore, by mid-2000, only 12 percent of demand had switched to ESPs, so that the default demand faced by UDCs was considerable. That they had divested themselves of most of their generating capacity exacerbated the problem, for this meant they needed to buy vast supplies in


24. Actually, there were wholesale price caps, but they were many times higher than the retail price cap.
the CALPX. It has also been argued that the UDCs were restricted by regulation in their ability to enter into forward contracts. This prevented them from hedging wholesale spot price risk, which required that they buy their supply in the hour-ahead and day-ahead markets.  

As yet another example of the unintentional perversities of regulation, the ESPs had the incentive to “return” customers to the utilities, which meant more loss-inducing sales. An ESP with a forward contract to provide power to a customer at, say, $40/mwh could “buy out” the contract by paying the customer the difference between the retail price faced when buying a utility at $65 and the contractual price of $40. The ESP could then take that now available megawatt-hour and sell it in the spot market for hundreds of dollars. One then has a partially deregulated market in which restrictions and freedoms are incongruous, which creates distortions.

In sum, a big problem with restructuring in California was the restrictions placed on the UDCs. They were encouraged to divest their power, there was a cap on retail prices, and there were limitations in their ability to enter into forward contracts at the same time that whole-

---

Figure 12.11
Wholesale Electricity (Spot) Market

25. This argument is made in Joskow, “Difficult Transition,” and a counterargument is made in Wolak, “Diagnosing the California Energy Crisis.”
sale prices were largely unregulated. These features are correctable, and indeed other states and countries have avoided some of these problems. The greater challenge lies with the market power of wholesale power producers. There are some unique features of electricity markets that make it challenging to transform regulated electric power monopolies into a structure that relies on competition in wholesale power generation.

One feature is that electricity cannot be easily stored, so that demand must be met with just-in-time production by whatever generating capacity is available. The inability of suppliers to build inventories to meet demand in the event that price is high makes for opportunities for firms to exert market power. Second, network congestion, combined with nonstorability, limits the ability of remote suppliers to meet demand. This is why prices could be high in California yet much lower in other states. Third, short-run electricity demand is highly price inelastic. Spot prices can then be volatile and subject to strategic manipulation by limiting supply when demand is high and alternative sources of power are limited. Fourth, electricity demand varies widely over the day, over the week, and over the year. Given the lack of storability, generating capacity has to be sufficient to meet peak demand, which may occur for only a few hours during the year. It also needs to be mentioned that the performance of these markets depends crucially on the way in which the regulated transmission network is operated: whether there is adequate investment in transmission capacity and how it is priced.

In spite of the debacle in California, it must be remembered that there are many states and countries that have achieved some success with restructuring. As argued, the electric power market is particularly challenging because full deregulation is unlikely to work and it is often hard to get the right balance of regulation and deregulation. When combined with the endemic presence of market power, it is indeed a regulatory challenge. Some proposals put forth include introducing real-time pricing so that short-run demand is more responsive to price, and to expand transmission capacity in order to allow generating capacity from a wider geographic area to supply power to a market. Research and policy continue to evolve.

Summary

This chapter has discussed numerous challenges and problems of rate of return regulation. Regulators are confronted with the difficult task of setting the correct price and, even more challenging, of setting the correct price structure. Peak-load pricing is clearly an important development of the 1980s that has improved efficiency significantly, especially in the electric power industry, where demand fluctuates and capacity is costly.

Various cost distortions of rate of return regulation were discussed. The Averch-Johnson effect is when a regulated firm overinvests in capital because it expands its rate base and
thus the amount of profit that can be earned. A more ubiquitous inefficiency is associated with the cost-plus property of rate-of-return regulation. In that a regulated firm expects to receive profit equal to some markup over cost, cost-reducing activities are discouraged. Firms that cut costs do not receive those cost savings, inasmuch as prices will be reduced to the new level of costs. However, the “accidental” institution of regulatory lag and the practice of disallowing certain expenses or additions to the rate base are factors that tend to offset this problem.

Recent developments in regulatory practice, known as “incentive regulation,” have consciously, not accidentally, created incentives for a regulated firm to reduce cost. Earnings sharing allows a firm to retain a portion of all increases in profit and, while it does not fully eliminate the disincentive to reduce cost, it is an improvement. Within telecommunications, the most widely adopted form of incentive regulation has been price caps. A firm can retain all profits as long as its price does not exceed some specified level. Incentives are then high-powered, but regulators are faced with the challenge of properly setting the price cap. A cap set too high will create the usual monopoly deadweight welfare loss and a cap set too low can impose financial difficulties on a regulated firm.

There have been vast changes in the regulatory structure of wholesale electric power markets beginning in the late 1990s. This has introduced competition into power generation while maintaining regulation over the retail distribution market. One of the early innovators, the state of California suffered from poor regulatory design and unfortunate shocks to the demand and cost of power in 2000–2001. Though this disastrous series of events has caused a retreat back to regulation in California and may have slowed down restructuring in other states, the trend in the industry still appears to be toward injecting more competition into the electric power industry.

Questions and Problems

1. The rate base can be valued in various ways. What, if anything, is wrong with the utility commission valuing the rate base at its value according to the stock market and the bond market? That is, use the market value of the firm rather than, say, the original cost of assets less depreciation.

2. Utility company executives are often quoted as saying that if their allowed rate of return is too low, they will be unable to attract capital to pay for capacity to meet increasing demands. Does this mean that regulators cannot use the allowed return as a device to provide incentives for utilities—raising the rate for good performance and lowering it for poor performance?

3. Consider the Edison Electric Company with a production function \( Q = K^{0.5}L^{0.5} \), where \( Q \) is output, \( K \) is capital, and \( L \) is labor. The market rental rate of capital is \$0.50\) and the wage rate is \$0.50\) also. The utility commission has set the allowed rental rate at \$0.80. (Rental rates of capital are in dollars per unit of capital per year. With zero depreciation they are related to percentage costs
of capital in the following way. Suppose that the utility must invest in a generator at a cost of $5 per kilowatt of capacity, and 10 percent is its cost of capital; then the rental rate per year is 10 percent of the $5 per unit, or $0.50. Similarly, the percentage allowed rate of return would be 16 percent, since 16 percent of $5 is $0.80. Rental rates are therefore comparable to wage rates and other factor costs in applying standard static production theory.)

Edison faces a demand curve with the constant elasticity of demand 2.857, or \( Q = P^{-2.857} \). If Edison were unregulated, it would produce efficiently at a constant average and marginal cost of $1. However, because of Averch-Johnson effects, it uses too much capital under regulation and produces at an average cost of $1.01. Edison charges a price of $1.35 and sells \( Q = 0.42 \).

a. Find the price and quantity if Edison were an unregulated monopoly. Hint: Marginal revenue is \( P(1 - 1/2.857) \).

b. Find the sum of consumer and producer surplus for the case where Edison is regulated and where it is not. Hint: Using calculus, it can be shown that consumer surplus is \((0.54)Q^{0.65}\). Does regulation, even though imperfect because of Averch-Johnson effects, nevertheless result in an improvement over an unregulated monopoly case?

c. Of course, the first-best case of price-equal marginal cost and efficient production is superior to regulation. Find the efficient solution. Draw a figure that shows the two types of losses that regulation causes as compared to the efficient solution.

d. Assume now that the utility commission decides to lower the allowed rental rate from $0.80 closer to the market rate of $0.50. Assume that it picks $0.58. It can be shown that Edison will now choose to sell 0.67 units at a price of $1.15. Its average cost of production rises to $1.04. Compare this Averch-Johnson equilibrium with the earlier one in terms of total economic surplus. This, in fact, is the socially optimal allowed rental rate. Lower rates actually reduce total surplus. For further details, see A. Klevorick, “The Optimal Fair Rate of Return,” Bell Journal of Economics and Management Science (Spring 1971).

4. How does the Averch-Johnson characterization of the regulatory process differ from reality?

5. Edison Electric Company’s president has been arguing that residential electric rates need to be raised relative to industrial rates. His reason is that the rate of return that the company earns on its assets is higher from its industrial customers than from its residential customers. Is this a good reason? Hint: How can Edison determine its assets dedicated to the two classes of customers?

6. In a certain city where all parking is controlled by the city, it is possible to provide parking facilities in the downtown area at a constant marginal capital investment of $10,000 per space. Costs of operation can be neglected. There are three equal periods during the day of eight hours each, and spaces are rented only for complete eight-hour periods. During the peak period of each of 250 days per year, the demand for parking is given by \( P = a - bQ \), where \( P \) is the price per period for a parking space. During the other two off-peak periods of those 250 days, the spaces demanded are half that in the peak period, for each possible price. On other days demand is zero. Assume that the interest rate is 10 percent and the facilities do not depreciate.

a. If \( a = 16 \), \( b = 0.08 \), and existing spaces are 120, what would be the socially optimal prices during the three periods?
b. What is the optimal number of spaces, and what are the corresponding prices?

c. This case is a so-called firm-peak case, with peak demanders paying all capital costs. Now suppose that $a = 5$ and $b = 0.08$. If peak demanders pay all capital costs, what quantity is demanded by peak demanders? If off-peak demanders pay zero, what is their quantity demanded? (Fractions of spaces are legitimate.) This is the shifting-peak case.

d. For the demand curves in part c, find the optimal number of spaces and the corresponding prices.
13 Franchise Bidding and Cable Television

In spite of considerable deregulation in the 1980s, the government has continued to play a significant role in markets that are firmly believed to be natural monopolies. Most economists argue that there is a need for government intervention in some capacity when social optimality requires a single firm to produce some good or service. It is because we foresee a continued role for government intervention that we have dedicated several chapters to analyzing issues related to natural monopoly.

This and the following chapter investigate alternatives to regulation for the case of a natural monopoly. The approach explored in this chapter is the auctioning off of a franchise for the provision of a monopoly service. Referred to as franchise bidding, it entails issuing the franchise to the firm that proposes the lowest price for service while meeting certain criteria concerning quality of service. As will be made clear, it substitutes competition at the bidding stage for regulation. Our case study will concern cable television, where franchise bidding was used extensively.

A second alternative, which is investigated in chapter 14, is having the service provided by a public (or government) enterprise as opposed to regulating a privately owned firm. As an alternative to regulation, public ownership has been used rather widely in the distribution of electricity at the municipal level as well as for water utilities. Municipal electric utilities will be used as a case study in assessing the performance of publicly owned firms.

The main purpose of the two chapters is to explore these two policies—franchise bidding and public enterprise—and to assess whether they represent attractive alternatives to the traditional solution of regulation in solving the natural monopoly problem.

Theory of Franchise Bidding

If an industry is a natural monopoly, then cost efficiency requires a single firm to operate in the industry. Of course, in an unregulated environment with only one supplier, one would expect price to be set at the monopoly level. Because this entails price exceeding marginal cost, deadweight welfare losses would result, a situation that may not sustain itself. Potential entrants would observe one supplier in the industry earning above-normal profits and recognize that there is room for profitable entry by undercutting the monopolist’s price. Entry drives price down, but it results in too many firms operating in the industry. Given scale economies, average cost is higher than if only one firm produced. In an unregulated environment with a natural monopoly, we would expect to observe a price that is too high and/or excessive entry.

This scenario provides the rationale for the regulation of a natural monopoly. To ensure the efficient number of firms in an industry, entry regulation is typically proposed to prevent more than one supplier from operating. However, without the threat of entry (or the threat of regulation), a monopolist can be expected to set price at the monopoly level. To avoid welfare
losses induced by an excessively high price, price regulation is used to keep price at the level that maximizes social welfare.

In an important article in 1968, Harold Demsetz questioned the accepted belief that a natural monopoly must be regulated in order to achieve the social welfare optimum.1 In place of regulation, he proposed that there be franchise bidding. Specifically, the government would award a franchise to one firm for the provision of this service. The franchise would be awarded via competitive bidding, where a bid would take the form of the proposed price that would be charged for service. The prospective firm that offered the lowest bid would be awarded the franchise. If there is sufficient competition at the bidding stage, then price should be bid down to average cost (\( \hat{P} \) in figure 13.1) and the winner would earn normal profits. The role for government would be to act as an auctioneer rather than as a regulator.

---

Competition at the Bidding Stage

An Example of Franchise Bidding—A Modified English Auction

The motivating force of franchise bidding is that ex ante competition at the bidding stage keeps price and profit at the competitive level. Thus, ex ante competition serves the role usually played by active competition in the industry. To appreciate the effectiveness of ex ante competition, let us consider a particular auction form through which franchise bidding could take place.

The English (or oral ascending) auction is probably the most commonly used auction form. When an item like a painting or an oil lease is to be sold, the English auction works as follows. The auctioneer announces a bid, to which the bidders respond by signaling whether they are willing to buy at that price. If there are at least two active bidders (a bidder is active if he signals he is willing to buy at the going bid), the auctioneer raises the bid. He keeps raising it until there is only one active bidder. The last remaining bidder wins the item and pays a price equal to the final bid.

With franchise bidding, the franchise is instead awarded to the bidder who offers the lowest price for service, so that we need to slightly modify the English auction. Assume that the auctioneer starts at a high bid and lowers the bid as long as there are two or more active bidders. As soon as the bid is lowered to the point at which there is only one remaining bidder, the franchise is awarded to that bidder. In exchange for the franchise, the winning bidder is to charge a price for service equal to the final bid.

Suppose there are four firms bidding for the franchise. Let $AC_i(Q)$ represent the average cost function for firm $i = 1, 2, 3, 4$. As depicted in figure 13.2, the firms have different cost functions. The differences could be due to firms having different production technologies as the result of patents or trade secrets. Constrained to linear pricing (that is, a constant per-unit price), the social welfare optimum is to have firm 1 supply the good at a price of $\hat{P}_1$. This entails the most efficient firm pricing at average cost.

If the franchise is awarded via a modified English auction, the first question to ask is, what will be the optimal bidding strategy for a firm? Imagine that prior to the start of the auction, a firm decides on the bids for which it will remain active. It should be clear that firm $i$ will choose to remain an active bidder when the bid $B$ is greater than $\hat{P}_i$. If $B$ exceeds $\hat{P}_i$ and firm $i$ wins, it then earns above-normal profits, as it ends up being able to charge a price that exceeds average cost. If $B$ is less than $\hat{P}_i$, then firm $i$ leaves the bidding, because if it should win then it would have to charge a price less than average cost and thereby incur losses. Thus the optimal bidding strategy for firm $i$ is to remain active as long as $B$ is at least as great as

2. For some historical background on the use of auctions, see R. Cassady, Jr., *Auctions and Auctioneering* (Berkeley: University of California Press, 1967).
$\hat{P}_i$. At a bid of $\hat{P}_i$, firm $i$ would earn normal profits, as winning the franchise would result in average cost pricing.

Now let us determine what the outcome of franchise bidding will be if firms 1, 2, 3, and 4 compete via a modified English auction and use the bidding rule described. Suppose that the auctioneer starts at a bid above $\hat{P}_4$. All four bidders will signal they are active. Because there is more than one bidder active, the auctioneer lowers the bid. Once $B$ falls below $\hat{P}_4$, firm 4 drops out of the bidding. However, the bid will continue to fall because there are still three active bidders. Firm 3 drops out once $B$ is less than $\hat{P}_3$, which leaves firms 1 and 2 competing. As soon as $B$ falls just below $\hat{P}_4$, firm 2 leaves the bidding. Because this leaves firm 1 as the only active bidder, the outcome of the auction is that firm 1 wins the franchise and charges a price slightly less than $\hat{P}_2$ (the final winning bid).

There is good and bad news resulting from the auctioning off of the franchise. The good news is that the firm with the lowest average cost curve is the franchise owner. This will always be true because the firm that is most efficient can always outbid the other firms. The bad news is that price is approximately equal to $\hat{P}_2$, which exceeds the franchise owner’s
average cost. Consumer’s surplus is lower by the shaded area $\hat{P}_2 ab \hat{P}_1$ than when average cost pricing is used.

The reason price is too high is that there is insufficient competition. Firm 1 does not face other firms that are as efficient as it is. Suppose instead that there are two firms with average cost functions $AC^1(Q)$ instead of just one firm. In the auction, when the bid falls below $\hat{P}_2$, there will still be two active bidders (the ones with average cost function $AC^1(Q)$). Thus the bid will fall until it reaches $\hat{P}_1$. At that bid, the two firms are indifferent to winning the franchise. Once the bid falls below $\hat{P}_1$, both drop out. To decide on the winner, the auctioneer will have to choose randomly between the last two bidders. What is important is that the outcome of the auction entails a firm with the lowest cost structure producing and pricing at the socially optimal level of $\hat{P}_1$. Thus the ability of franchise bidding to result in a desirable social outcome is very much dependent on there being sufficient competition in the most efficient average cost range.

**Analysis of Franchise Bidding**

As long as there is sufficient competition at the bidding stage, franchise bidding results in average cost pricing and the most efficient firm operating. Although regulation could, in principle, achieve this outcome, the advantage to franchise bidding is that it imposes no informational requirements on a government agency. With regulation, the regulatory agency must have cost and demand information in order to achieve average cost pricing. Under franchise bidding, no such information is required, as competition, rather than a regulator’s decision, results in average cost pricing. Thus, franchise bidding can achieve the same outcome as regulation but at lower cost, since less information is required and a regulatory agency need not be established. A second advantage is that the inefficiency of rate-of-return regulation is avoided. With franchise bidding, there is no incentive to overcapitalize (the Averch-Johnson effect; see chapter 12). The franchise owner has every incentive to utilize resources efficiently because it retains all profits.

Thus far we have painted an excessively rosy picture of franchise bidding. It is best to think of our analysis up to this point as representing the potential of franchise bidding. In the next section we will be concerned with how franchise bidding performs in a more realistic setting. However, aside from any practical problems with franchise bidding, one criticism of it is that it does not result in marginal cost pricing. Average cost pricing is certainly preferable to price being set at a monopoly level, but we know from chapter 11 that a two-part

---

3. This analysis assumes, of course, that the two bidders do not collude and try to keep the bid above $\hat{P}_1$. For example, one firm could agree to drop out of the bidding at $\hat{P}$, where $\hat{P}_1 < \hat{P} < \hat{P}_2$, in exchange for a cash payment. This collusion would allow both firms to earn higher profits in comparison to when the bid is driven down to $\hat{P}_1$. The inability to collude is then an essential condition if franchise bidding is to achieve a socially desirable outcome.

tariff with per-unit price equal to marginal cost generally yields higher social welfare than average cost pricing.\(^5\)

Although this criticism is valid, it should be noted that franchise bidding can be adapted so as to allow for two-part tariffs. However, we now need to assume that the government knows the market-demand function (though it still does not need cost information, in contrast to the case where regulation is used). Knowledge of the demand function gives the government a measure of how much consumers value a two-part tariff. Instead of awarding the franchise to the bidder who offers the lowest price, the government awards it to the one with the two-part tariff that maximizes social welfare (that is, the two-part tariff that consumers most prefer). Competition results in the franchise’s being awarded to the firm that offers the two-part tariff that maximizes consumers’ surplus, subject to profits being normal.

Consider the demand and cost function in figure 13.3. Let us first show that there is a two-part tariff that beats out average cost pricing, so that a firm offering a price equal to average cost cannot win the franchise. A two-part tariff entails the payment of a fixed fee by a consumer, which gives that consumer the right to buy the good, plus a per-unit price for each unit purchased. For ease of analysis, let us assume the demand curve is perfectly income-inelastic.\(^6\) If a firm bids a price equal to average cost and is awarded the franchise, then consumers’ surplus is triangle \(abc\). Now consider a second firm that offers a price equal to marginal cost and a fixed fee of \(F\). Because marginal cost pricing increases consumers’ surplus by the area \(bcde\), consumers are better off with the two-part tariff as long as \(bcde\) exceeds \(NF\), where \(N\) is the number of customers. That is, in order to have price lowered from average cost to marginal cost, consumers are willing to pay up to \(bcde\). Thus, if it costs them less than \(bcde\), they are better off with the two-part tariff. In order for a firm to earn at least normal profit, it must set \(F\) (given price equals marginal cost) such that \(NF\) exceeds \(defg\) where rectangle \(defg\) represents the losses from pricing below average cost. Because \(bcde\) exceeds \(defg\), there exists a two-part tariff that yields higher welfare for consumers than average cost pricing and allows the firm to earn above-normal profit; that is, \(bcde > NF > defg\). As long as there is sufficient competition, the fixed fee will be bid down to \((defg/N)\) so that the franchise winner earns normal profit. We conclude that if the government has demand information, franchise bidding will result in a two-part tariff with marginal-cost pricing.\(^7\)

---

5. Recall that a two-part tariff has a customer pay a fixed fee for the right to purchase a good and a per-unit price for each unit purchased.

6. The significance of this assumption is that the demand curve will not shift as we change the fixed fee. Increasing the fixed fee is equivalent to reducing consumer income. If the demand curve was not perfectly income-inelastic, it would shift in as we increase the fixed fee under the assumption that the good is normal. In order to simplify the analysis, we assume the demand curve is not responsive to changes in income.

7. One can show that there is always a two-part tariff that is Pareto-superior to average cost pricing. See Robert D. Willig, “Pareto-Superior Nonlinear Outlay Schedules,” \(Bell\ Journal of Economics\) 9 (Spring 1978): 56–69. The idea of franchise bidding with two-part tariffs is analyzed in Daniel A. Graham and John M. Vernon, “A Note on Decentralized Natural Monopoly Regulation,” \(Southern Economic Journal\) 57 (July 1991): 273–75.
As a final remark, it is essential that the franchise be auctioned off to the firm that proposes the tariff with the highest social welfare. In contrast, suppose the franchise was awarded to the firm that was willing to pay the highest fee to the government and it was allowed to price freely. We know that the franchise owner would set price at the monopoly level $P_m$ (see figure 13.4). Because this yields profits of $\pi_m = [P_m - AC(Q_m)]Q_m$, firms are willing to pay up to $\pi_m$ for the franchise. Hence, competition results in the franchise being awarded to the firm that offers $\pi_m$. The winner’s effective average cost is then $[(\pi_m/Q) + AC(Q)]$. As shown in figure 13.4, at the monopoly price $P_m$ the franchise owner earns normal profit. Although normal profit was also earned when the franchise was auctioned off to the firm offering the lowest price for service, a major difference is the market price. With average-cost pricing, the latter franchise bidding scheme results in a price of $\hat{P}$ versus $P_m$. This lower price yields higher welfare. For franchise bidding to lead to a socially desirable solution, the franchise must be auctioned off to the firm offering the lowest price for service and not the highest franchise fee.

Figure 13.3
Franchise Bidding Using Two-Part Tariffs
Quality of Service

If the service is homogeneous, then competition over pricing at the bidding stage leads to the social welfare optimum. Typically, however, a firm chooses the characteristics of the service it offers in terms of such variables as durability and reliability. For example, auto manufacturers do not just set the price of the automobiles they sell. Considerable time and expense are devoted to choosing the product characteristics in terms of design, motor size, and optional features like a rear window defroster and an intermittent wiper. Product performance is also a choice variable in terms of durability (is the body protected against rust?) and reliability (will the car start every winter morning?). Better paint jobs and stronger batteries result in higher quality.

Suppose that product quality is a choice variable and per-unit cost is increasing in the level of quality. If franchise bidding takes place over the price of service, competition will drive
down not only price but also quality. The winning bidder will be the one that offers the lowest quality product at a price equal to average cost. The problem with franchise bidding in this situation is that society may not desire the low-quality, low-price alternative. Consumers may be willing to pay a higher price (at least as great as average cost) for higher-quality service. However, because competition takes place only over price, franchise bidding results in the lowest price, which implies the lowest level of quality.

There are several methods for adapting franchise bidding to cases in which products can be differentiated. One is for the relevant government agency to specify the quality of service that is to be provided by the franchise owner. Franchise bidding can then take place over price without driving down quality. A second approach is to have multidimensional bidding in which firms propose not only a price for service but also the attributes of the service it will offer. These attributes determine the level of quality. The bidding procedure is now considerably more complex, as there is a trade-off between higher quality and a lower price that has to be evaluated by the bidders and the franchising agency.

There are two difficulties inherent in both of these methods. First, they require the franchising agency to have information on consumers’ valuation of quality. This will be needed in order to specify the required characteristics of a service or to analyze the trade-off between quality and price in awarding the franchise. A second difficulty is enforcing the agreement made with the franchise winner regarding quality. Quality is considerably more difficult to monitor than price. For example, assessing the performance of a long-distance telephone call in terms of static interference and connection time is no easy task. The same is true for determining the reliability of service, inasmuch as this requires monitoring service over a period of time. In sum, introducing the quality dimension into the franchise bidding process greatly complicates matters and requires a bigger role for government. In that the main attraction of franchise bidding is the minimal role for government, the advantage of franchise bidding over regulation is reduced as more realistic elements, like product quality, are included in the analysis.

**Inefficiency of Franchise Fees**

To a local government official, a cable system operator is not just a provider of cable services but also an appealing source of tax revenue. Concerned with their prospects for reelection, elected officials may prefer to raise additional revenue by taxing a cable system operator rather than increasing the income and property taxes of their constituents. An implication is that a bidder may be awarded the franchise not because it offered the most attractive package for consumers but partially because it provided more tax revenue.

A common device is to assess a cable system operator a fee that is a percentage of its gross revenues. Let us explore the implications of this fee’s being an element of a firm’s bid. To begin, note that firms will no longer compete price down to the socially optimal price of \( \hat{P} \) (see figure 13.5). In order for them to earn normal profit, a higher price will have to be charged.
in order to cover cost and the franchise fee. For example, suppose that the franchise owner has to pay $100\alpha$ percent of gross revenue to the local government. Firm revenue is then $PD(P)(1 - \alpha)$, where $D(P)$ is the market demand function. An owner’s average revenue is then $P(1 - \alpha)$, which is depicted in figure 13.5. If competition drives average revenue to average cost, the winning bid will be $\hat{P}$, which exceeds $\bar{P}$. Note that if it prices at $\bar{P}$, the firm sells $\bar{Q}$ units and, at that output, its average revenue (after deducting the franchise fee) equals its average cost. Compared to the social optimum, the welfare loss is the sum of triangle A, which is the forgone surplus from producing $\hat{Q} - \bar{Q}$ fewer units, and rectangle B, which is the increase in cost on the $\bar{Q}$ units produced because average cost is higher. Rectangle C is the franchise fee paid by the firm to the local government and is not counted in the welfare loss calculation since it is a transfer.\(^8\)

8. We want to thank James Prieger for correcting an error that was present in previous editions.
Contractual Arrangements for the Postbidding Stage

In our analysis thus far, we have assumed that bidding, whether over price or price and quality, is performed once and for all. Such a situation is suitable if the environment never changes. As we all know too well, the environment does change, and most often in ways that are quite unanticipated. Input prices and technology change over time, causing the average cost curve to shift. The demand function changes as income and preferences change. In order to allow price to adjust in the future, franchise bidding will have to be supplemented with a contract that specifies how changes in cost and demand conditions are to be handled. In an important article in 1976, Oliver Williamson provided a detailed analysis of different types of contracts to handle future, unanticipated events. 9

Recurrent, Short-Term Contracts

One approach is to use recurrent, short-term contracts. This approach avoids having to specify too much in the contract. Periodically, the franchise is put up for auction, at which time a new award is made and a new contract is issued. Thus, cost and demand changes are handled through recontracting. This procedure provides the current franchise owner with the incentive to honor its current contract, especially concerning quality. If it does not, it may be penalized in the next round of bidding through differentially disadvantageous treatment by the government agency.

The key element to the success of recurrent short-term contracts is bidding parity at renewal time. If there is a lack of parity among bidders, in particular if the incumbent firm has an advantage, it may be able to renew the contract at noncompetitive terms. One advantage the current franchise owner does have is that it has already invested in plant and equipment. In contrast, a new firm would have to make this major investment. Thus, while the incumbent firm would be willing to bid down to average variable cost, a new firm would not bid below its average total cost. The result is that a new firm could be more efficient than the current franchise owner but the latter would outbid it. In figure 13.6 the new firm has average cost curve \( AC^N(Q) \), while the incumbent has average cost curve \( AC^I(Q) \) and average variable cost curve \( AVC^I(Q) \). Even though \( AC^N(Q) < AC^I(Q) \), the franchise is renewed by the current owner because \( AVC^I(Q) < AC^N(Q) \). The resulting price would be \( \bar{P} \).

This problem can be rectified by the transfer of capital from the previous franchise owner to the new one. The government could mandate a compulsory transfer of assets in a fair and efficient manner. Thus a new firm need not have to invest from scratch but could instead purchase the existing plant and equipment from the previous franchise owner. In response to this proposal, Williamson argued that this transfer is by no means trivial. Requiring the

incumbent to sell at original cost less depreciation provides room for it to inflate the true value of its assets. In addition, what about human capital? Workers and management learn over time and build up knowledge as to how to best run the company. If a new firm comes in with new personnel, this capital cannot be transferred. Of course, a new firm is likely to hire these people, recognizing their marginal product is the greatest.

It is a valid point that the transfer of capital is indeed not a trivial problem. Nevertheless, if the franchise was won by a new firm, it would be in both firms’ interest to transfer those assets. This is then just a bargaining problem, and one would expect it to be resolved because it is in the best interests of both firms to come to an agreement.

Another source of advantage to the existing franchise owner at renewal time may be imposed by the government agency monitoring the industry. It is often believed that bureaucrats try to maintain the status quo. Change brings additional work, for example, having to assist in the transfer of assets to the new franchise owner. Furthermore, if the change turns out for the worse, bureaucrats are likely to be blamed. In contrast, there is little lost from
maintaining the status quo. Because of all this, the government agency is apt to favor the incumbent firm at renewal time as long as it has performed reasonably well.

**Incomplete, Long-Term Contracts**

The main alternative to recurrent short-term contracts is incomplete, long-term contracts, where long-term is on the order of fifteen to twenty years. The contracts are incomplete in the sense that not all contingencies are provided for but will instead be handled through negotiation and implicit understandings. An advantage to the long-term contract is that it gives the franchise owner the proper incentives to invest in long-lived assets, inasmuch as it is assured of being around to receive the returns from this investment. In contrast, with short-term contracts, a franchise owner may be less inclined to make long-term investments, given that its contract may not be renewed and there is uncertainty over the transfer of assets. This issue is particularly important for natural monopolies, as capital investment is typically a large portion of costs.

A disadvantage of the long-term contract is that it is difficult to write. It must allow for price to be changed in the future in response to changes in cost and demand conditions. Thus a price formula will need to be specified that relates the cost of inputs to the price of service. Monitoring of quality will be essential, and the contract will have to provide for penalties if quality is not kept up to the specified level. With recurrent short-term contracts, penalties are less essential because the incumbent firm can be penalized at renewal time.

**Ex Post Opportunistic Behavior**

As we mentioned earlier, the current franchise owner will typically have certain advantages over prospective franchise owners. These advantages include having already made the necessary capital investment, better knowledge about the technology through learning by doing, better information on market demand, and greater familiarity with the franchising process. These advantages may provide the current franchise owner with the ability to engage in opportunistic holdup by forcing it to make favorable changes in its contract.

Recall that when an incomplete long-term contract is used, the local government and the franchise owner will have to negotiate prices and quality over time as new information arrives concerning the cost of providing the service as well as the demand for the service. If cost turns out to be higher than anticipated or demand turns out to be weaker than anticipated (weaker demand means higher average cost when there are economies of scale), price will have to be raised. One strategy for a prospective firm at the bidding stage is to offer a low price and then, if the franchise is won, to petition for a price increase on the basis that average cost was underestimated or demand was overestimated. As long as the franchise owner has already made some investment in capital and the proposed rate increase is not too great, the government agency is likely to concede to this request rather than have to incur the cost of
performing another round of franchise bidding and having a new franchise owner commit resources.

Several devices are available for reducing the incentive of a franchise owner to engage in opportunistic holdup. Depending on the degree of renewal competition, the current franchise owner can be punished for opportunistic behavior by being relieved of its franchise at renewal time. However, punishment need not be limited to renewal time. A government agency has several devices for disciplining the franchise owner over the life of the contract. First, there are penalty clauses for reneging on one’s proposal that are built into the franchise contract. Second, because the contract is incomplete, there are issues left open that need to be negotiated over the life of the contract. A government agency can punish opportunistic holdup through the way the agency behaves during these negotiations. Another mechanism may exist when a firm operates franchises in several markets. A reputation for opportunistic holdup can harm one’s chances of winning a new franchise.

Finally, it should be mentioned that opportunistic holdup may also work in the other direction. Once a firm has sunk considerable resources, the government agency has power emanating from its monopsonistic position. By controlling the franchise owner’s access to the market, the government agency is acting as an agent for buyers. If fixed costs are sunk, then the franchise owner will operate as long as price is at least as great as average variable cost. Thus a government agency may take advantage of its position by forcing the firm to price at average variable cost, even though the firm would not have made the original investment if it anticipated a price below average cost. A government agency might implement such a policy by requiring that the franchise owner keep price fixed over time in an inflationary environment. If price were initially at average cost, price would fall below average cost over time as long as the inflation rate is positive.

Just as reputational effects are important for controlling the behavior of the franchise owner, they can be effective in deterring opportunistic behavior by the government. If prospective firms believe that price will not be adjusted over time for inflation, this belief will affect their initial bidding behavior. Specifically, they will not bid price down to average cost because they know a price equal to today’s average cost may mean normal profit today but will mean losses in the future as the real price falls below real average cost. Though opportunistic behavior is potentially a problem when agents invest specific resources in their relationship with one another, it is at least partially mitigated by the disciplining influence of reputation.

Assessment of Franchise Bidding

Williamson makes a valid point when he says that uncertainty about the future creates a bigger role for government in the bidding process. In the case of incomplete, long-term contracts, franchise bidding differs from regulation as a matter of degree and not of kind. A government agency must specify quality and monitor the performance of the franchise owner. It
must also negotiate price changes with the supplier. This point is made even more evident once one recognizes that regulation is itself an incomplete, long-term contract in which the firm is guaranteed a fair rate of return and there is an established procedure for making changes.

In an ideal environment, franchise bidding appears quite superior to regulation. It accomplishes the same outcome with at lower cost; there is no need for a regulatory agency and no incentives for the monopolist to act inefficiently. However, as we introduce product quality and uncertainty, franchise bidding begins to look more and more like regulation so that the apparent advantages to franchise bidding become less outstanding. Still, franchise bidding is an interesting alternative to regulation and it has considerable potential. The next section will analyze how franchise bidding has performed with regard to cable television. This application will provide us with some insight as to how franchise bidding has fared in practice in solving the natural monopoly problem.

Cable Television

The first cable systems were constructed in the late 1940s for the purpose of improving the reception of broadcast signals sent by local television stations. By expanding the market of a local television station, cable service was complementary to the service provided by local broadcasters. Then, beginning in the late 1950s and early 1960s, cable systems sought to expand the demand for their services by importing signals from other regions via microwave relay stations. In addition to providing better reception of local stations, cable service also offered a wider selection of broadcasts. The importation of signals then meant that cable services also represented an alternative to local television broadcasts. In this sense, cable service was a substitute for local broadcasting. This dual aspect of cable service—being both a substitute and a complement to local television broadcasting—will be most relevant when it comes to understanding regulatory policy with respect to cable television.10

With a few exceptions, each of the approximately 11,000 cable systems that exist today has a local monopoly over cable service. However, it is important to recognize that there are a number of substitutes for cable service. To begin, almost one-third of the households that could have cable service choose not to have it. Many of these households have chosen to rely on local television broadcasts. If they want the variety and reception of cable service, they have the option of purchasing a satellite dish, for which there are a number of available technologies. Although a cable company may have a monopoly over cable service, it is important to keep in mind that cable service must compete with other forms of video entertainment.

Historical/Regulatory Background

The Communications Act of 1934 created the Federal Communications Commission (FCC) and gave it the power to regulate wire and radio communication. Of particular relevance is that the FCC was given regulatory authority over television broadcasting. During the early period of cable television in the 1950s, cable television was allowed to grow with a minimum of regulatory interference. The FCC even went so far as to refuse to accept regulatory jurisdiction over cable television on the basis that it was neither a broadcasting facility nor a wire common carrier. Its general policy was to allow auxiliary services to television, like cable systems, to develop in an unfettered environment.

It is significant that at the time of the FCC’s decision not to regulate cable television, cable services were very much complementary to television broadcasting in that they improved the reception of signals and thereby expanded a local television station’s market. During the period 1959–1966, cable systems began to actively import signals via microwave relay stations. Importation was a relatively inexpensive way to increase the demand for cable services. Of course, imported channels are substitutes for local broadcasters, so that the latter’s advertising revenue would be expected to fall.

As the economic theory of regulation would predict, television broadcasters pressured the FCC to regulate cable television. The FCC responded in 1962 by forbidding the importation of a distant signal into a market for which the same broadcast was carried by a local television station. Then, in 1966, the FCC asserted full regulation over cable television. It required a cable system to carry all local television stations and prohibited the importation of any additional signals in the top 100 television markets. The FCC ended this freeze on importation in 1972, although, in its stead, it instituted a complex set of rules that still greatly limited the importation of signals. Even with these burdensome restrictions, the number of cable systems and the number of subscribers experienced healthy growth. From 1965 to 1975 the number of subscribers increased more than sevenfold, from 1.2 million to 8.5 million viewers. The subscription ratio, which is the ratio of the number of subscribers to the number of households, had risen to 12.4 percent by 1975 from only 2.3 percent ten years earlier.11

The launch of the Satcom I satellite in 1975 provided cable systems with a relatively inexpensive technology for receiving distant programming. At the same time, the FCC was beginning to loosen its restrictions on the importation of signals. On top of these developments, the Home Box Office decision gave cable systems the right to compete freely with broadcast television. These events allowed a tremendous increase in the services provided by cable television. Prior to 1971, only 6 percent of all cable systems had more than twelve active channels. This figure exceeded 50 percent by 1980, while 90 percent of cable subscribers receive more than twenty channels in 1992.

These developments were followed by tremendous growth in the cable television industry (see table 13.1). In the 1970s less than a third of all households had access to cable. Now, cable service is almost universally available. Commensurate with this rise in availability has been a rise in subscribership, increasing more than fivefold over twenty years, with the number of subscribers exceeding 73 million in 2004.\(^{12}\) Growth in revenue has been even more spectacular, experiencing a tenfold increase over 1976–1995 (in real terms) with revenue exceeding $51 billion in 2003.\(^{13}\)

During this phase in which cable service became ubiquitous, the regulatory issues were very different from those in the early part of cable’s history. Since the early 1980s, the central regulatory issues have concerned cable rates and competition to cable operators. These issues will be discussed in a later section.\(^{14}\)

### Cable Television as a Natural Monopoly

In this section our objective is to analyze the rationale for government intervention in cable television. After reviewing the technology of cable systems and estimates of the degree of scale economies, we will conclude that current evidence is consistent with the hypothesis that cable television is a natural monopoly.


Technological Background

A cable system comprises three key components: (1) the headend, (2) the distribution plant, and (3) the subscriber interface. Figure 13.7 depicts the physical design of a cable system. The purpose of the headend is to receive signals and process them for distribution. A major part of the headend is the antenna, which receives the signals. (It is for this reason that cable systems were originally referred to as community-antenna television, or CATV.) Once received, the signals are then distributed via the distribution plant, which has historically used coaxial cable, though now fiber-optics is the preferred technology. The final element of the cable system is the subscriber interface. It connects a subscriber to the distribution plant.
The capacity of the system, in terms of the number of channels, is determined by the distribution plant. Until the early 1980s most systems had a capacity of forty channels. At that time, the standard became a coaxial cable with fifty-four channels. The advent of fiber-optics introduced the potential of near unlimited capacity. It is quite feasible to have hundreds of cable channels on a system.

A major portion of the cost of a cable system is in the purchase and construction of the distribution plant, in particular, the cost of laying the cable. This can be done aerially through the use of existing utility lines or underground, though the latter is considerably more expensive. If a consumer is in the geographic region covered by the distribution plant, the marginal cost of a subscriber is relatively low, inasmuch as it just entails installing a subscriber interface. Given that marginal cost is relatively low and the cost of the distribution plant and the headend are fixed with respect to the number of subscribers, a cable system is generally considered to experience declining average cost per subscriber.

**Estimates of Scale Economies**

There are two important issues related to the optimal industry structure for cable television. First, for a given geographic market, is industry cost minimized by having cable systems not overlap? In other words, should each cable system have exclusive rights over the geographic area covered by its distribution plant? This is a particular type of scale economy that is referred to as *economies of density*. If there are economies of density, then cost efficiency requires that there be no duplication of cable services; that is, each cable company’s distribution plant should not overlap. However, even if there are economies of density, this fact does not necessarily imply that a metropolitan area should be served by a single cable company. A separate but related issue is whether cost efficiency requires a single firm to supply the entire geographic market or whether instead the market should be subdivided, with a different firm supplying each submarket. We will deal with each of these issues in turn.

To examine the degree of economies of density, we need to estimate the average cost per subscriber for a fixed plant size. There are various measures of plant size, though the most commonly used measures are the number of homes for which cable is available and the number of cable miles. Using a 1979 data set comprising seventeen cable systems in New Jersey, G. Kent Webb estimated an average cost curve for a system size of a thousand miles of cable and a density of one hundred households per mile of cable.\(^\text{15}\) As shown in figure 13.8, average cost per subscriber is clearly declining in the number of subscribers. This cost study provides empirical evidence supportive of the hypothesis that cable television experiences economies of density. For example, as market penetration doubles from 40 percent to 80 percent, average cost declines by over 40 percent, from approximately $14 to $8.

\(^{15}\) Webb, *Economics*, pp. **–**.
Figure 13.8
Average Total Cost for Cable Television Given Fixed Plant Size (1982 Dollars)

System size: 1,000 miles of cable.
Subscriber density: 100 households per mile of cable.

study by Eli Noam used cost data for nearly all 4,200 cable systems for the year 1981.\footnote{16} He found that a 10 percent increase in the number of subscribers reduced unit cost by about 0.5 percent. These studies support the hypothesis that cable systems exhibit economies of density. Cost efficiency then requires that the distribution plants of cable systems not overlap.\footnote{17}

A second and related issue is whether it is more efficient to have one cable system for an entire geographic area or instead to subdivide the area with each cable system serving a sub-market (and no two cable systems overlapping). Cost studies have shown slight economies of scale of this variety. The New Jersey study by Webb estimated average cost per cable mile while holding market penetration constant (see figure 13.9). Economies of scale are present, but they are not particularly strong. The study by Noam estimated the elasticity of cost with respect to the number of homes passed. The elasticity was estimated to be 1.02, which means that a 10 percent increase in the size of the cable system (as measured by the number of homes passed by cable) results in a decrease in unit cost of only 0.2 percent.\footnote{18}

In summarizing this evidence, it appears that there are significant economies to having cable systems not overlap but that there are only slight cost savings from having a single cable system serve a geographic area, as opposed to subdividing the area and having different cable systems serve these different submarkets. Taking into account both types of economies, it was estimated that the overall elasticity of scale was 1.096, so that a 10 percent increase in size (which includes both a 10 percent increase in homes passed and a 10 percent increase in subscribers) would reduce unit cost by almost 1 percent.\footnote{19} Although the evidence is supportive of the hypothesis that cable television is a natural monopoly, it is not overwhelming.\footnote{20}

**Franchising Process**

Given that the available evidence is supportive of cable systems exhibiting economies of density, there exists a rationale for a government providing exclusive rights to a cable company over a geographic area. The evidence we have considered also suggests that there may be little cost inefficiency from allowing several firms to serve a metropolitan area as long as their systems do not overlap. In this section we will briefly describe the process used

\footnote{17. Also see Bruce M. Owen and Peter R. Greenhalgh, “Competitive Considerations in Cable Television Franchising,” *Contemporary Policy Issues* 4 (April 1986): 69–79.}
\footnote{18. Noam, “Economics of Scale.”}
\footnote{19. Ibid.}
The franchising process usually begins with the local government soliciting proposals from prospective cable operators. Often applicants are already cable operators in other markets in the United States. Once proposals are received, a negotiation process between the local government and the prospective cable operators begins. Typically, the local government will select a few applicants and ask them to submit bids. Eventually, a franchise is awarded or, if the local government is unsatisfied with the bids, the process is restarted.21

The typical franchising process took between two and ten years.²² An extreme case is that of Philadelphia. Four rounds of franchise bidding took place starting in 1966; twenty years later Philadelphia still had not issued a franchise. The delay due to this process is certainly not trivial.

The proposal submitted by an applicant to the local government typically provides a technical account of the proposed system, programming services, and prices. A description of the proposed system includes the number of channels, the pay channels that will be available, and any auxiliary services, such as two-way interactive channels. Price information includes the basic monthly price per subscriber and the price of pay channels.

Generally, much more information is required. In the state of Massachusetts, the Cable Television Commission requires the franchise proposal to include:

1. duration of the license (not to exceed 15 years);
2. area(s) to be served;
3. line extension policy;
4. construction schedule;
5. initial rates and charges;
6. amount and type of bond and insurance;
7. plan for local supervision;
8. criteria to be used in assessing applicant qualifications;
9. location of any free installations (e.g., public schools, police and fire stations, and other public buildings); and
10. equal employment opportunity practices.

Suggested items are:

1. capability of the system;
2. plan for access channels and facilities;
3. plan for municipal coordination with the licensee;
4. types and patterns of ownership;
5. coordination with contiguous communities; and
6. subscriber rights of privacy.²³

More generally, one would expect bidders to include any items in a proposal that are likely to give them an edge over competing bids. A discussion of such items will be provided later.

The most common contract used by the franchising authority is the long-term, nonexclusive contract. It is typically fifteen years in length. Nonexclusivity gives the local government the right to put the franchise up for auction if it finds the current franchise owner to be performing in an unsatisfactory manner.

Assessment of Franchise Bidding

In assessing the performance of franchise bidding, there are several issues that need to be addressed. To begin, a necessary condition for franchise bidding to work is that there is sufficient competition at the bidding stage. It is then important that we assess the degree of ex ante competition. However, recall from our discussion of rent-seeking behavior that it is vital that competition be of a certain type. Bidders must compete in terms of price and quality and not in terms of activities that benefit local governments but reduce social welfare. Activities of the latter variety come under the category of nonprice concessions.

The effectiveness of franchise bidding in generating a socially desirable solution depends not only on behavior at the bidding stage but also on how well the franchise owner performs over time. Related to this issue is the degree of opportunistic holdup. We will need to

²² Hazlett, “Private Monopoly.”
²³ Prager, “Franchise Bidding,” p. 118.
investigate whether franchise winners implement their proposals or renege on them by raising price and reducing quality. Finally, a key ingredient in the success of franchise bidding is that competition at the renewal stage be effective. If it is absent, the current franchise owner may be able to renew its franchise at terms that yield an abnormally high price without commensurate quality. In addition, renewal competition can be an important instrument for deterring opportunistic holdup.

**Competition at the Bidding Stage**

Generally, franchise bidding brought forth four or five applicants. A study of 92 franchise bidding cases in Massachusetts over 1973–1981 found the number of applicants per case to range from one to seventeen, with an average of 5.2. Furthermore, the average number of applicants was observed to rise from 2.6 in 1973–1978 (with fourteen cases) to 5.7 over 1979–1981 (with seventy-eight cases). However, more recent evidence shows a declining number of applicants. In twenty-seven of the largest franchises awarded after 1982, the average number of bidders was only 2.7. The franchise bidding process in Baltimore brought forth only a single proposal.

There are four major areas in which competition for a franchise has taken place. One, obviously, is the price of cable service. Table 13.2 lists some of the major franchise awards over 1980–1982, including the range of bids for basic cable service. As one can see, the bids are often rather disparate. To examine the effect of the number of bidders on the winning bid price, Robin Prager surveyed sixty-six cable firms in 1984. She found that for each additional bidder in the franchise auction, the price of basic cable service declined by $0.15 per month per subscriber, and the price of a pay channel declined by $0.16 per month per subscriber. As one would expect, an increase in the degree of ex ante competition reduced the winning bid.

A second dimension of competition is product quality. This dimension includes channel capacity and programming, as well as technical standards concerning signal quality and reliability. Although firms compete in ways other than price and quality, these other dimensions of competition may tend to reduce social welfare. Such activities include franchise fees and nonprice concessions. The local government that is franchising cable service typically demands payment of a fixed fee and/or a percentage of gross revenues. However, competition in terms of financial payments to local governments is limited by federal regulation. The 1972 FCC rules set a maximum of 0 percent of gross revenues, which was then raised to 5 percent in 1984.

24. Prager, “Franchise Bidding.”
25. Ibid.
27. Prager, “Franchise Bidding.”
Because of these restrictions on financial payments, bidders have actively competed in providing nonprice concessions. A common nonprice concession is the provision of channels for public service, educational use, and governmental use. Also included are free hookups, public parks, local origination studios, and excess capacity. Nonprice concessions are estimated to comprise about 26 percent of building costs and 11 percent of operating expenses. Every dollar spent on nonprice concessions was found to raise the monthly rate per subscriber for basic cable service by $0.35. On average, monthly basic rates are lower by $0.49 per subscriber if all nonprice concessions are eliminated. As our analysis of rent-seeking behavior predicts, nonprice concessions tend to raise the price of cable service.

29. Ibid.
30. Ibid.
31. For a study that explores which aspects of a proposal are conducive to winning a franchise, see Phillip A. Beutel, “City Objectives in Monopoly Franchising: The Case of Cable Television,” *Applied Economics* 22 (September 1990): 1237–47.
In concluding, let us discuss a study that provides some rough estimates of some of the costs imposed by the franchising process. First, as we just mentioned, nonprice concessions result in uneconomic investment. As shown in table 13.3, estimates of these costs are between $4.81 and $6.00 per subscriber per month. Second, the delay in the franchising process results in forgone consumer surplus, as consumers await cable service, and forgone profits, as firms await being able to supply cable service. The estimated cost is between $0.97 and $2.72. Third, the cost of lobbying by bidders (for example, in hiring consultants and attorneys) ranges from $0.71 to $1.71. Finally, franchise fees incur a cost of between $1.29 and $1.88. The total cost is between $7.78 and $12.31 per subscriber per month. This works out to about one-third of gross revenues. Even taking into account the roughness of these estimates, the politically imposed costs of franchising appear to be considerable.

### Performance after the Initial Award

Clearly, there is anecdotal evidence of franchise owners reneging on their proposal once the franchise is issued. The winning proposal for the Milwaukee franchise called for a 108-channel system at a monthly basic service price of $4.95. Not long after the award was issued, the franchise winner renegotiated its contract and instead installed a 54-channel system at a monthly basic service price of $11.95.  

The systematic evidence is rather mixed. On the one hand, cable franchises were renegotiated in twenty-one of the thirty largest television markets, including eight before any homes were wired. Of course, this evidence does not describe the degree of renegotiation. Other evidence suggests that opportunistic holdup is not a serious problem. A survey of the trade press over 1980–1986 revealed only sixty cases of cable operators’ reneging on their contract

### Table 13.3

<table>
<thead>
<tr>
<th></th>
<th>Uneconomic Investment</th>
<th>Delay</th>
<th>Lobbying</th>
<th>Franchise Fees</th>
<th>Total Costs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low estimate</td>
<td>4.81</td>
<td>0.97</td>
<td>0.71</td>
<td>1.29</td>
<td>7.78</td>
</tr>
<tr>
<td>Midpoint estimate</td>
<td>5.41</td>
<td>1.84</td>
<td>1.21</td>
<td>1.58</td>
<td>10.04</td>
</tr>
<tr>
<td>High estimate</td>
<td>6.00</td>
<td>2.72</td>
<td>1.71</td>
<td>1.88</td>
<td>12.31</td>
</tr>
</tbody>
</table>


---


from a pool of over 3,000 cable operators.\textsuperscript{34} Evidence from a survey of franchises in Massachusetts over 1973–1984 found that construction schedules and quality levels were generally consistent with the accepted proposals.\textsuperscript{35} More important, the average time between the issuance of the franchise and the first rate increase was almost 33 months, and the average time between the first rate increase and the second rate increase was over 29 months.\textsuperscript{36} On average, cable operators waited a considerable length of time before requesting a rate increase. Furthermore, of sixty-two franchises, forty-nine experienced a decrease in the real price of cable service from the time of the award until June 1984, while only thirteen had a real increase in the price of cable service.\textsuperscript{37}

Another study tested the degree to which opportunistic holdup was prevented by reputational effects.\textsuperscript{38} Recall that a cable operator who plans to compete for cable franchises in other markets may be penalized for opportunistic holdup because of the reputation it creates for not fulfilling its side of the contract. Communities may find the proposals of cable operators with such a reputation to be less credible and thus may be less inclined to award them franchises. This study surveyed 221 communities served by franchised cable operators. Multiple-system operators were found to be less likely than single-system operators to have construction delays and more likely to provide voluntary improvements in the cable system. If we assume that a cable operator who currently serves several markets is more likely than a single system operator to compete for other cable franchises in the future, then this empirical finding is supportive of reputational effects deterring opportunistic holdup.

In terms of competition at the renewal stage, the most significant statistic is that of 3,516 refranchising decisions, only seven resulted in the local government’s removing the current franchise owner.\textsuperscript{39} The question, however, is whether this statistic is indicative of the satisfaction of the local government with the performance of the initial franchise owner or whether it is indicative of the lack of competition at the renewal stage. To investigate this issue, a study by Mark Zupan examined 59 randomly chosen renewal agreements over 1980–1984 and compared the terms of the initial contract with the renewal contract. The results are shown in table 13.4. From the perspective of the franchise owner, the renewal contract tends to be more favorable than the initial contract. Quality is lower at renewal time as, on average, channel capacity is reduced by nine channels, while the number of community channels is

\textsuperscript{34} Zupan, “Efficacy.”
\textsuperscript{35} Prager, “Franchise Bidding.”
\textsuperscript{36} Ibid.
\textsuperscript{37} Ibid.
reduced by 0.8. The monthly basic price per channel is higher by $0.01 per subscriber and
the monthly pay channel price is higher by $1.13. However, the monthly basic system price
is reduced by $0.35 and the franchise fee is increased by 0.2 percent.

Rate Regulation

One of the main advantages of the long-term contract is that it avoids the costly process of
frequent bidding and provides the franchise owner with the necessary long-term incentives for
investment. On the downside, a long-term contract needs to make provisions for adjusting rates
in response to changing cost and demand conditions. Ultimately, local governments chose to
handle this problem through the regulation of rates. The typical procedure was for the cable
operator to propose rate changes to a government authority, as is done with regulated public
utilities like local telephone. By 1979 at least ten states had some form of rate regulation.

By the end of the 1970s it appeared that cable television was on a path to traditional price
regulation, but there proved to be many twists and turns in regulatory policy. The order of
the day in the 1980s was to permit cable system operators to price in an unfettered manner
while using regulatory controls to limit entry—hardly traditional public utility regulation!
Deregulation began with the federally mandated elimination of price controls over pay
channels in 1979 and ended with the deregulation of basic cable service rates by the Cable
Communications Policy Act of 1984. A U-turn occurred with the Cable Television Consumer
Protection and Competition Act of 1992, which reintroduced rate regulation. The Telecommu-
nications Act of 1996 represented the next (and latest) turnabout as it instituted a plan to
achieve full price deregulation by 1999.40

40. For estimates of the effect of deregulation on the value of cable franchises, see Adam B. Jaffee and David M.
Kantner, “Market Power of Local Cable Television Franchises: Evidence from the Effects of Deregulation,” RAND

---

Table 13.4
Deviation between Terms of Initial Contract and Renewal Contract

<table>
<thead>
<tr>
<th>Term of Trade</th>
<th>Average for Initial Estimated Deviation in Renewal Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Channel capacity</td>
<td>46.4</td>
</tr>
<tr>
<td>Franchise fee</td>
<td>2.9%</td>
</tr>
<tr>
<td>Community channels</td>
<td>2.8</td>
</tr>
<tr>
<td>Basic system price (monthly)</td>
<td>$9.35</td>
</tr>
<tr>
<td>Basic price per channel offered</td>
<td>$0.52</td>
</tr>
<tr>
<td>Lead pay channel price (monthly)</td>
<td>$9.51</td>
</tr>
<tr>
<td></td>
<td>9 fewer channels</td>
</tr>
<tr>
<td></td>
<td>0.2% higher</td>
</tr>
<tr>
<td></td>
<td>0.8 fewer channels</td>
</tr>
<tr>
<td></td>
<td>$0.35 lower</td>
</tr>
<tr>
<td></td>
<td>$0.01 higher</td>
</tr>
<tr>
<td></td>
<td>$1.13 higher*</td>
</tr>
</tbody>
</table>

* The difference is statistically significant at the 0.05 level in a one-tailed test.

Rate Deregulation, 1984–1992

The Cable Communications Policy Act of 1984 prohibited federal, state, or local regulation of rates for basic cable services. Continued regulation would be allowed only where effective competition was absent. The FCC considered a cable company to face effective competition if it competed with three or more over-the-air television stations. This meant that only 3 percent of the cable systems in the United States were subject to rate regulation.

In addition to prohibiting local authorities from controlling cable rates, the 1984 act also constrained competition. First, it mandated that all cable systems be franchised by local governments. This provision eliminated the threat of entry from a second cable operator as a downward pressure on cable rates. Second, the 1984 act made it more difficult for a local government to fail to renew a cable company’s franchise. To allow it to do so, the cable company would have had to violate the franchise agreement and, even in that case, the company would have had to be given adequate opportunity to rectify the situation. The threat of not renewing a company’s franchise because rates were too high was then considerably weakened. Finally, the act codified the FCC’s 1970 ban on a local telephone company’s providing cable service in its jurisdiction.

Cable rates became unregulated in December 1986, and thereafter followed a steady rise exceeding the rate of inflation (see figure 13.10). By mid-1991, basic cable rates had risen 36.5 percent in real terms. Before jumping to conclusions that cable companies have taken advantage of their market power, let us next note that the average basic package experienced an increase in channels of almost 30 percent, from twenty-nine to thirty-seven channels. Furthermore, cable operators have increased spending on basic programming from around $300 million to over $1 billion. Since deregulation of cable rates, then, we have observed a rise both in price and in quality.

A study by Robert Rubinovitz, an economist at the Antitrust Division of the U.S. Department of Justice, estimated the effect of the 1984 act on quality-adjusted rates. He compared the price and characteristics of basic cable packages in 1984 (the regulated benchmark) with those in 1990 (the unregulated benchmark). During that time, the real price of cable service rose 42 percent. After taking account of changes in the cost of a cable operator providing service and, most important, in the changes in the quality of service (for example, the number of channels), Rubinovitz concluded that real cable rates were higher by over 20 percent.

41. For an overview of the regulation of the cable industry, see Thomas W. Hazlett and Matthew L. Spitzer, Public Policy toward Cable Television: The Economics of Rate Controls (Cambridge, Mass.: MIT Press and AEI Press, 1997).


A proper welfare analysis requires measuring both the fall in consumer surplus from higher rates and the rise in consumer surplus from more channels. Figure 13.11 depicts this trade-off. Adding additional channels causes the demand curve to shift out from $D(P)$ to $D'(P)$, and this results in consumer surplus rising. At the same time, price rises from $P'$ to $P''$, and this lowers consumer surplus. (As the net effect of raising quality and price on quantity could be positive or negative, we will take a neutral position and assume that quantity is unchanged.) Whereas consumers initially had surplus of triangle $P'bc$, they now have triangle $P''ad$. This change can be broken down into higher expenditure, as measured by rectangle $P'baP''$, and higher value from enhanced quality, as measured by parallelogram $cbad$. Deleting the common area, the net change in consumer well-being is area $A$ minus area $B$. Using data from June 1991 to March 1993, a study found these two measures to be roughly the same in

Figure 13.10
Real Cable Service Rates, 1984–1995

Source: Bureau of Labor Statistics. Figure is from Thomas W. Hazlett and Matthew L. Spitzer, Public Policy toward Cable Television: The Economics of Rate Controls (Cambridge, Mass.: MIT Press and AEI Press, 1997).

A proper welfare analysis requires measuring both the fall in consumer surplus from higher rates and the rise in consumer surplus from more channels. Figure 13.11 depicts this trade-off. Adding additional channels causes the demand curve to shift out from $D(P)$ to $D'(P)$, and this results in consumer surplus rising. At the same time, price rises from $P'$ to $P''$, and this lowers consumer surplus. (As the net effect of raising quality and price on quantity could be positive or negative, we will take a neutral position and assume that quantity is unchanged.) Whereas consumers initially had surplus of triangle $P'bc$, they now have triangle $P''ad$. This change can be broken down into higher expenditure, as measured by rectangle $P'baP''$, and higher value from enhanced quality, as measured by parallelogram $cbad$. Deleting the common area, the net change in consumer well-being is area $A$ minus area $B$. Using data from June 1991 to March 1993, a study found these two measures to be roughly the same in
size, so that the consumer surplus loss from the rise in price was of comparable magnitude to consumers’ willingness to pay for the additional channels.\footnote{44}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{welfare_analysis.png}
\caption{Welfare Analysis of Higher Cable Rates and More Cable Channels}
\end{figure}

**Rate Reregulation, 1992–1994**

Rising cable rates soon brought pressure on regulators and legislators to reinstitute some form of regulation. In June 1991 the FCC commissioners voted 5–0 to reinstate local rate regulation for many cable television operators. The following year Congress passed the Cable Television Consumer Protection and Competition Act of 1992 over the veto of President George Bush. The 1992 act required that rates for basic cable services be regulated by the franchising authority or by the FCC. The FCC reduced rates by 10 percent in September 1993 and by an additional 7 percent in July 1994. Thomas Hazlett, an economist

specializing in the cable television industry, described the progression from the 1984 act to the 1992 act:

While the profitability of the cable industry has never been higher, its political vulnerability has never been more evident—and the two events are connected. By fortuitously steering themselves through the regulatory maze to arrive at the bliss point of a legally protected but unregulated monopolist [as created by the 1984 Act], cable companies have traded friends for wealth in the political game.45

As is apparent in figure 13.10, real cable rates did indeed fall during 1992–1993.46 What is more intriguing is what happened to penetration rates. Contrary to the anticipation of Reed Hundt, chairman of the FCC, the growth in the number of basic cable subscribers fell in spite of the fall in cable rates. What did rise was the penetration rate for premium services, which, unlike basic cable services, were not subject to rate regulation. This evidence is consistent with the idea that cable system operators were, to a certain extent, able to thwart the intentions of the 1992 act by watering down basic cable services.47 These results were anticipated by John Malone, CEO of the largest cable system operator, TCI:

These noxious FCC rules are not going to be able to constrain the economics in entertainment very long. What’s gonna happen is there’ll be a shift from basic to [unregulated] à la carte services . . . . We’ll continue to diversify away from the regulated government-attacked core.48

**Rate Dereregulation, 1994–Present**

The lesson to be learned from the experiences of 1992–1994 is that there are serious limitations to the ability of the FCC, or any government agency, to effectively control quality-adjusted prices in the cable industry. Indeed, the behavior of the FCC beginning in late 1994 is consistent with the agency’s having learned this lesson, for at that time it began to effectively dereregulate cable rates. While cable rates did rise in response, it was also observed that the growth in the number of basic cable subscribers accelerated, as did the penetration rate for premium channels. It is then quite possible that consumers were paying more but were happy to do so in exchange for higher quality.

A clear legislative path to rate deregulation was laid out in the Telecommunications Act of 1996. It mandated an immediate deregulation of cable rates for small cable systems, which served about 20 percent of cable households. For larger systems, a three-year phase-out of rate regulation concluded on March 1, 1999. Furthermore, if a cable system operator was in

46. This discussion is based on Thomas W. Hazlett, “Prices and Outputs under Cable TV Reregulation,” *Journal of Regulatory Economics* 12 (September 1997): 173–95.
direct competition with a local telephone company for cable services, then deregulation could occur immediately. This act also laid the basis for entry, as it eliminated the cross-ownership restriction created by the FCC in 1970 (and codified in the 1984 act), which prohibited a local telephone company from providing cable services in its area of telephone service.

What has transpired since deregulation is consistent with the cable companies exercising market power. A big disappointment has been the lack of entry by local telephone companies, so that satellite TV remains the only serious competitor. From 1999 to 2002, the average price of expanded basic cable service increased from $29.41 per month to $36.47, an annual rate of increase of about 7.4 percent. By comparison, the general price level rose only 2.6 percent annually during that period. While the cable companies attribute this increase in price of basic cable service to a rise in programming fees, one study shows that higher programming fees account for only 42 percent of the increase. We will have to wait and see whether these rate increases will bring forth yet another round of regulation in the future.

Is There a Role for Government Intervention?

Natural monopoly is one of the few economic rationales for government intervention. Though the statistical evidence supports the hypothesis that cable television is a natural monopoly, some economists have argued that recent events put into question the social value of government intervention and, in particular, the use of franchise bidding.

Price Regulation

Recent experience with rate regulation is not promising in that the government does not appear to have control over the instruments to achieve a socially preferred outcome. Cable program services are divided into basic, premium, and pay-per-view services, but rate regulation applies only to basic service. Thus, in response to a government authority like the FCC requiring a reduction in basic cable rates, a cable company can respond by retiering—shifting better channels to unregulated tiers—or moving the better channels to à la carte status. Even if the regulation specifies a reduction in the average price per channel of basic service, the cable company can simply replace the shifted channels with cheaper, lower-quality channels. Soon after the 1992 act, jokes began circulating about “The Fireplace Channel” and the “The Fishbowl Network.” Though Congress tried to take account of these strategic responses in their 1992 act, loopholes remained. These techniques allowed some cable systems’ basic rates to rise after the 10 percent rate rollback mandated by the FCC.

Can the government effectively regulate cable rates? This is the question. The difficulty lies in the fact that the cable companies have numerous unregulated instruments to get around or offset rate regulation. Of particular importance is adjusting the quality of the service. Lower rates do not imply that consumers are better off if cable companies reduce the quality of the service at the same time. Regulating quality is intrinsically difficult, and, in addition, cable regulators are legally prevented from controlling programming, because cable operators have First Amendment rights as “electronic publishers.” Even if cable operators are using their market power to set rates, it is not clear that regulating rates can lead to an improvement for consumers.

Role of Competition

When an industry is a natural monopoly, as appears to be the case for the cable television industry, a primary concern is that, in the absence of price and entry regulation, either of two undesirable scenarios will emerge. First, there may be only one active firm that exercises its market power by keeping price above cost. Second, there could be an unstable pattern of entry and exit resulting in periodic price wars—but not a consistently competitive price—and wasted resources. Can competition work in the cable industry? Interestingly, there is some evidence that an unregulated market can result in a stable industry arrangement of two firms that price in a relatively competitive manner.

Though rare, competition has existed among cable operators in a few communities. Companies’ distribution plants overlap, so that consumers have two cable systems from which to choose. When two systems overlap, the situation is referred to as an overbuild. In 1992, overbuilds comprised just 1 percent of cable households. As shown in table 13.5, cable operators in competitive systems (those with two or more cable operators) had over 14 percent more basic channels, and the price of the basic service was 8 percent lower. Compared to the average monopoly system, the price per channel of basic service was 20 percent lower for the average competitive system. Furthermore, it appears that two cable operators can profitably coexist. For the average cable market with one cable operator, Thomas Hazlett estimated that a second cable operator that reduced rates by 20 percent and increased total penetration by 20 percent could expect to earn a rate of return of 16 percent. 51

The more significant source of competition, however, is from direct broadcast satellite (DBS), which has experienced impressive growth in recent years, its penetration rate having risen from 11 percent to 17 percent over 1999–2002. 52 Still, it falls considerably short of the penetration rate for cable service, which is almost four times as high. To what extent DBS has restrained cable rates down is an interesting unexplored question.

Summary

In principle, franchise bidding is a very attractive alternative to standard public utility regulation. Through the use of potential competition, as opposed to active competition, franchise bidding can, ideally, achieve the social welfare optimum without imposing the costs of a regulatory structure. Competition for the franchise can drive price down to average cost. Because price is fixed and the franchise owner retains all profit, it has every incentive to be efficient. Franchise bidding can result in an efficient market structure, cost minimization, and socially optimal pricing.

In practice, franchise bidding is considerably more complicated and less effective in generating a socially desirable solution. As we observed in the case of cable television, bidding takes place over many more dimensions than just price. Because quality is a firm choice variable, the bid must encompass this information. This complexity makes it a considerably more difficult task for a local government to award a franchise. Furthermore, the process creates non-welfare-improving ways in which prospective franchise owners can compete. In sum, the franchising process generates considerable welfare losses in the case of cable television.

At present, franchising is part of the cable television industry’s past. The future course appears to be one of an unregulated marketplace, though if history is a guide, traditional price regulation may return. The experience of the cable industry raises the question of whether franchise bidding is, at best, a useful short-run to medium-run solution of the natural monopoly problem, but not a long-run solution. If long-term contracts cannot be effectively written or if there is no active competition at renewal time, society may eventually find itself faced with a traditional problem of regulation even if it began its course with franchising. To what extent cable television is representative of how franchise bidding will perform is unknown.

Table 13.5
Competitive versus Monopolistic Cable Systems

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Mean Revenue per Month</th>
<th>Mean Basic Rate</th>
<th>Mean No. of Basic Channels</th>
<th>Price/Channel</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Subs)</td>
<td>(dollars)</td>
<td>(dollars)*</td>
<td></td>
<td>(dollars)</td>
</tr>
<tr>
<td>Monopoly</td>
<td>54.89 million 32.12</td>
<td>19.08</td>
<td>35</td>
<td>0.55</td>
</tr>
<tr>
<td>(11,354 systems)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Competitive</td>
<td>1.13 million 25.08</td>
<td>17.49</td>
<td>40</td>
<td>0.44</td>
</tr>
<tr>
<td>(103 systems)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Price/quality difference</td>
<td>-21.9%</td>
<td>-8.3%</td>
<td>+14.3%</td>
<td>-19.8%</td>
</tr>
</tbody>
</table>

* On most popular tier of basic programming.

Such a question can only be addressed with further experience with this method for handling the problem of natural monopolies.

Questions and Problems

1. Assume a franchise is to be auctioned off where market demand is $100 - P$. Suppose there are just two firms competing for this franchise. One firm’s cost function is $C_1(q) = 100 + q$, while the other firm’s cost function is $C_2(q) = 12q$. The franchise is auctioned off using a modified English auction, where the firm offering the lowest price for service wins the franchise.
   a. Who will win the franchise?
   b. What will the winning bid be?

2. Continuing with question 1, suppose the local government decides to issue the franchise to the firm that offers the biggest fee. An English auction is used, where the firm offering the biggest payment to the local government wins the franchise. The franchise owner is free to charge any price that it likes. (Note: The marginal revenue curve is $100 - 2Q$.)
   a. Who will win the franchise?
   b. What will the winning franchise fee be?
   c. What will price be?
   (Hint: Calculate monopoly profit for each of the two bidders.)

3. Which of the two methods described in questions 1 and 2 should a local government use if it wants to:
   a. Maximize consumer welfare.
   b. Maximize government revenue.
   c. Maximize the number of consumers who buy this service.
   d. For each of these objectives, can you think of a better method than the one described in question 1 or 2?

4. Compare the following three methods of franchise bidding:
   b. Long-term contracts.
   c. Recurrent short-term contracts in which the local government owns the capital.

5. In practice, almost all cable television franchises are renewed.
   a. Do you think this fact is evidence that franchise owners are performing in a satisfactory way, or that competition at renewal time is weak?
   b. How would you go about determining which hypothesis is correct?

6. In 1984, Congress prohibited the regulation of the price of cable services. In response to the passage of this legislation, one study found that the market value of a cable system was (statistic-
tically speaking) unchanged in the top 100 broadcast markets but went up in other markets. Explain.

7. What are the technological sources of cable television being a natural monopoly?

8. Should we allow free entry into cable television? In discussing this question, consider whether cable television is a contestable market.

9. Should we use franchise bidding for local telephone? Can you think of any other markets for which franchise bidding would be a viable alternative to monopoly regulation?

10. Should local telephone companies be allowed to operate a cable system? If so, how should they be regulated?
The ultimate objectives of any government policy that is designed to handle the natural monopoly problem are to have a single firm producing efficiently and pricing at the socially optimal level. The first objective can be achieved relatively easily by issuing a monopoly for the provision of the service to a privately owned firm. In acting to maximize its profit, the firm will produce efficiently. Unfortunately, the objective of profit maximization also leads to a price above the social optimum and thereby creates deadweight welfare losses.

To induce the firm to price in a socially efficient manner, we have thus far analyzed two alternative policies. One policy is to establish a regulatory agency that sets price and constrains the firm to earn a normal rate of return. Some of the problems with this approach are that the regulatory agency lacks the necessary information to make effective pricing decisions and that the profit constraint may lead the regulated firm to produce inefficiently. Two examples of the latter are a reduced incentive to adopt cost-reducing innovations, as additional profits earned cannot be fully retained by the firm, and overcapitalization due to the Averch-Johnson effect. An alternative approach considered in the preceding chapter is to auction off a franchise to the firm that proposes the lowest price for service. Ideally, this scheme solves all problems because the franchise owner has every incentive to produce efficiently and competition at the bidding stage results in socially optimal pricing. Unfortunately, in practice, franchise bidding and regulation differ only by a matter of degree and not of kind.

A common element to both of these policies is to have the service provided by a privately owned firm and then to design an institution that forces the firm to price at the socially optimal level. Regulation achieves this objective through the enforcement powers of a regulatory agency, whereas franchise bidding uses ex ante competition to achieve it. The idea is to change the constraints faced by the profit-maximizing monopolist in order to induce it to set a lower price than it would otherwise.

In this chapter we consider a policy that is intrinsically different from both regulation and franchise bidding. Rather than constrain a privately owned firm, the approach is based on altering the objective of the firm from profit maximization to social welfare maximization. This rather quixotic objective is pursued by having the service provided by a public enterprise. In contrast to the privately owned firm, a public (or government) enterprise is owned and operated by the government. The potential of such an approach lies in the fact that the objectives of a public enterprise may be more likely to be in line with the maximization of social welfare. In contrast to the manager of a private enterprise, who is accountable to shareholders, the manager of a public enterprise is ultimately accountable to voters.

This chapter is concerned with assessing public enterprise as an approach to handling the problem of natural monopoly. Of specific interest is comparing its performance to the regulation of a privately owned firm, which is the most common policy in the United States. In contrast, public enterprises are used quite extensively in Europe. Prior to beginning this task, we should first recognize that public enterprises are used for a variety of tasks, of which
handling the natural monopoly problem is just one. Thus, let us first consider public enterprise from a more general perspective before embarking on our narrower goal.

**General Background**

Public enterprises have a long and tainted history. In the ancient days of Athens, the government owned the mines, and prior to the second century B.C., the Ch’in dynasty in China had government monopolies over the provision of salt and iron. The Roman Empire had a large number of public enterprises that handled such diverse activities as recreation (such as games and circuses) and defense (for example, weaponry). A more modern example is the provision of postal service, which is performed by a public enterprise in almost all countries today. In the case of the United States, the postal system was created by the first Congress in 1789. While originally a department of the federal government, it became a semiautonomous, publicly owned firm in 1971 and was placed under the jurisdiction of the Postal Rate Commission.

Public enterprise is used in place of private enterprise for a variety of reasons. The purpose we are concerned with in this chapter is to handle the natural monopoly problem. Rather than attempt to cover all of the other reasons for their use, let us just discuss some examples so as to highlight the diversity of origins of public enterprise.

In the case of Communist-ruled countries like Cuba, the choice of public enterprise is largely based on ideological reasons. Privately owned firms that act to maximize owners’ wealth are anathema to Marxist systems. It was also for political reasons that France expropriated Renault after World War II. At the time, Renault was one of the three largest automobile manufacturers in France. As punishment for cooperating with the Germans during the occupation of France, the French government took control of Renault. It has performed reasonably well under public ownership.

Less by design and more by accident, the Italian government controls vast parts of the industrial sector, including almost all of shipbuilding. This formation came about during the depression in the 1930s. At that time, the government created the Institute for Industrial Reconstruction to handle the assets of failed banks. The end result was considerable government control of assets in the industrial sector. For similar reasons but more by design, the U.S. Congress created the Consolidated Rail Corporation (Conrail) to help the ailing railroad industry.¹ Created by the Railroad Revitalization and Regulatory Reform Act of 1976, Conrail was formed by merging seven bankrupt railroads, including the Penn Central. Until its bank-

---

1. For a discussion of Conrail as well as some other rather interesting government enterprises, see Lloyd Musolf, *Uncle Sam’s Private, Profit-Seeking Corporations* (Lexington, Mass.: Lexington Books, 1983).
ruptcy in 1970, the Penn Central handled about 20 percent of rail-freight movement in the United States.

Public enterprises may also be used in order to raise revenue. This is certainly one of the reasons for the use of state-owned and -operated lotteries and liquor retailing outlets. Besides often being quite profitable and a source of revenue, the fact that they are vices may also explain why they are state-controlled. It is thought by some that free enterprise and vices like gambling and liquor are not a suitable combination. The fear is that private enterprise will lead to abuses. Whatever the reasons, an increasing number of states are operating lotteries, and sixteen states have handled liquor retailing for many years.

There are widely differing opinions on the proper role of public enterprises. This diversity is evidenced by the variety of international experience. Beginning with the Industrial Revolution, and especially since World War II, Europe has shown a much greater inclination than the United States to use public enterprise. For example, most of the railroad industry and all airlines are privately owned in the United States, whereas both industries are almost wholly government-owned in most European countries.

This difference in approach is observed not just across countries but also across time within the same country. A classic case is that of Great Britain. When in power, the Labour Party is inclined to nationalize certain industries, such as steel. In contrast, when the Conservative Party is in control, it pursues a policy of privatization and restores these firms to the status of private ownership. This was the policy of the Conservative administration under Margaret Thatcher. Both British Airways and British Telcom (BT) were sold to private citizens. In the case of BT, a most interesting strategy was pursued:

All employees on privatisation will be given £70-worth of free shares—and will get two free shares for each one bought up to a limit of £100. Similarly, each telephone subscriber who buys shares of £250 (payable in three calls) will qualify for an £18 rebate on his quarterly telephone bill. The aim is no secret: the more people become shareholders, the more difficult it will be for a Labour government to renationalise BT.2

Positive Theory of Public Enterprise

Everywhere else in this book, we assume that firm behavior is motivated by the desire to maximize profit. Profit maximization seems to be the single most plausible objective of a privately owned firm, as it results in the maximization of shareholders’ wealth. By specifying the objective of a firm (or, more generally, an economic agent), we can make predictions as to how it will respond to government policies. Such predictions are essential in determining

2. The Economist, October 6, 1984, p. 89.
the effects of those policies. For example, the Averch-Johnson effect tells us that rate-of-return regulation results in a profit-maximizing firm overcapitalizing. By assuming that a firm maximizes profit subject to rate-of-return regulation, we are then able to identify an inefficiency resulting from government policy.

To compare a policy of regulation with one of public enterprise, we need to provide a theory as to how a public enterprise behaves. Because it is not privately owned, it is unlikely to maximize profit. It would be nice if a public enterprise was to maximize social welfare, but it would be hasty to jump to the conclusion that it does so. This is the difference between positive and normative economics. From a normative perspective, a public enterprise designed to handle the natural monopoly problem should pursue social welfare maximization. A positive theory tells us, in contrast, how a public enterprise actually does indeed behave. In order to derive such a theory, we will present a more general model of the firm than has been previously discussed.

Managerial Model of a Firm

An element common to almost all large organizations is a separation of ownership and managerial control. The modern private corporation is a classic example. Ownership is held by many shareholders, each with typically a small or negligible percentage of shares. As a result, day-to-day control of the firm rests with the appointed manager, though ultimate control lies with the shareholders. This statement is analogously true for most public enterprises, as ownership rests in the populace of the relevant political jurisdiction but control lies in the manager appointed by an elected official or a government bureaucrat, whichever it may be. In today’s economy it is fair to say that the separation of ownership and control is indicative of almost all large organizations.

Let us construct a model of the firm based upon the recognition that managerial control does not rest with the owners but rather with a paid employee. If owners have as much information as managers and can perfectly monitor the actions of the manager, this separation issue is of no importance. In that case, if the manager does not do exactly what is in the owners’ best interest, she will be fired. In spite of the separation of ownership and control, the results are the same as when the organization is owner-operated.

Of course, it is quite unreasonable to assume that owners can even come close to perfectly monitoring the manager. It is generally not in the best interests of any individual owner to spend the resources required to closely observe the manager’s actions. Furthermore, through her day-to-day intimacy with the operations of the firm, the manager has much better information than the owners. Thus, even if the owners did observe the actions of the manager (for example, in choosing project A over project B), it is difficult for the owners to determine whether the manager is acting in their best interests.

These two conditions—the separation of ownership and control and the imperfect monitoring of managerial behavior—are characteristic of most major private and public
enterprises. The managerial model of the firm has been developed to describe how such an organization behaves. A central assumption of this model is that the manager acts to maximize her own utility subject to the constraints and incentives instituted by the owners. Within this model, the goal of the owners is to construct a set of constraints and incentives so as to induce the manager to act in their best interests. Of course, the goal of the manager is to maximize her own utility. Our goal is to derive predictions as to how such a firm behaves.

In terms of her relationship with the firm, a manager’s utility depends on income, nonpecuniary benefits, and effort. The manager’s utility is greater, the greater is income and the greater are nonpecuniary benefits. The latter may include such items as the prestige of having a large staff or having a company jet at her disposal. In contrast, utility is less, the more effort the manager exerts. We can think of more effort as entailing less leisure, and less leisure generally reduces utility.

Not being perfectly monitored, the manager is likely to choose less effort and more nonpecuniary benefits than are in the owners’ best interests. This is why the owners must design incentive schemes and create constraints in order to induce the manager to work hard and act efficiently (note that choosing to have a large staff for the purpose of prestige can be inefficient because it may not sufficiently increase productivity so as to warrant its cost).

Thus far we have not had to specify whether the firm is a private or a public enterprise because the manager in both organizations will tend to act the same, ceteris paribus. The differences in the two organizations lie in (1) the interests of the owners and (2) the mechanisms available to induce the manager to act in the owners’ interests.

Managerial Model of a Private Enterprise

In a private enterprise, owners want the manager to act so as to maximize profit. A key mechanism to induce such behavior is to make the manager’s financial compensation depend on the profitability of the firm, that is, the use of an incentive scheme in which income increases with firm profit. This often takes the form of bonuses or issuing stock and stock options to the manager so as to make her a shareholder. Of course, this hardly solves the problem, because the manager knows that if she works harder and increases profit, she will get only a small part of that profit increase. Making income depend on firm profit provides the right incentives, but it cannot entirely solve the problem of the separation of ownership and control.

An implicit incentive scheme also arises in the labor market. If a manager’s hard work results in the firm earning a high rate of return, this will increase her reputation as a productive manager. Her value in the labor market will rise, resulting in higher future income. Thus, not only does superior performance increase current income, through bonuses and stock holdings, but future income is increased as well, by enhancing the manager’s reputation.

A final mechanism that induces the manager of a private enterprise to maximize profit is the threat of being fired if the firm performs poorly under her management. If owners observe poor performance, they can pursue actions to have the manager dismissed. Such a dismissal
reduces not only current income but also future earnings, by causing deterioration of her reputation in the labor market. In reality, this is not a very effective constraint on managerial behavior because it is unlikely to be used. Given the wide dispersal of ownership in most privately owned firms, the owners are not usually in a position to credibly threaten such a response.

Nevertheless, the threat of being fired can arise quite effectively via the capital market. A firm that is being run inefficiently presents a profitable opportunity for investors. Because it is operating below its potential, the shares of the firm are selling at a relatively cheap price. Thus, investors (or, in today’s parlance, raiders) can purchase shares at this low price, install new and efficient management, and return the firm to efficiency. The increase in profit earned by the firm is the financial return to the investors. In some cases, the investors immediately resell the firm at the higher share price. Thus, even if the ownership of a firm is widely dispersed, poor firm performance will induce investors to buy shares and concentrate ownership in order to replace ineffective management.

Managerial Model of a Public Enterprise

Is there a similarly strong case to be made for the manager of a public enterprise to act in the owners’ best interests and thus maximize social welfare? For two reasons, this is unlikely to be true. First, welfare, unlike profit, is difficult to measure. The use of imperfect indicators should provide opportunities for the manager to act in his own interests and not those of society. Second, the constraints imposed on the manager of a private enterprise by the capital market are conspicuously absent. This lack of constraints will give the manager of a public enterprise greater discretion in his actions.

Although welfare is difficult to measure, there are substitutes that elected officials and voters can use to gauge the performance of a public enterprise. One obviously is price. A second set is attributes of output, for example, reliability of the service provided by the public enterprise. For the case of an electric utility, frequent blackouts would signal poor performance.

Several theories of public enterprise behavior have been formulated on the premise that managers act to maximize political support and this goal is achieved by producing those attributes that signal good performance. By increasing political support, a manager raises his income and the likelihood of increased job tenure. In “A Theory of Government Enterprise,” Cotton Lindsay notes,

[M]anagers are therefore influenced to divert resources from the production of attributes which will not be monitored to those which will. In so doing they will increase the perceived value of the [public enterprise’s] output.3

For example, suppose that the quality of a service is easily observable but cost is not. The manager will then tend to choose too high a level of quality in order to increase political support. The important point is that the imperfections inherent in using attributes to measure managerial performance are likely to lead to strategic behavior by the manager. The manager will tend to produce those attributes to a degree that is excessive from society’s perspective.

Sam Peltzman proposed a different theory, in which the manager of a public enterprise uses price to maximize political support. This theory yields several interesting results. Let us suppose that political/voter support is greater, the smaller is the subsidy given to the public enterprise. This assumption is reasonable because the higher the subsidy, the more tax revenue must be raised, and voters dislike tax increases. If we assume that the public enterprise must earn a normal rate of return if it is to raise new capital, then the subsidy will equal \(-\pi(P)\), where \(\pi(P)\) is profit for a given price, \(P\). If \(\pi(P) > 0\), then the subsidy is negative, as the public enterprise adds to government revenues. Let us also assume that voter support is greater at lower prices, inasmuch as voters are also consumers.

One prediction of this model is that a public enterprise’s manager will set a price below that which maximizes profit. To show why this is true, consider price being set at the profit-maximizing level, which we denote as \(P^m\). Recall that marginal profit (that is, the change in profit from a small change in price) equals zero at a price of \(P^m\). The reason is that if marginal profit was positive, then profit could be increased by raising price, which contradicts \(P^m\) being the profit-maximizing price. Similarly, if marginal profit was negative at \(P^m\), then profit could be increased by lowering price. Hence, marginal profit must be zero at \(P^m\).

Now consider the public enterprise charging a price slightly less than \(P^m\). There is a direct and an indirect effect on political support. The direct effect is that voters care about price in their role as consumers. The indirect effect is that price affects firm profit, which affects the amount of the government subsidy. Voters care about the amount of the subsidy, as they are also taxpayers. Because marginal profit at \(P^m\) is zero, a very slight reduction in price does not affect firm profit, so that it does not affect the amount of the subsidy. The indirect effect on political support from charging a price slightly less than \(P^m\) is then zero. On the other hand, the direct effect of charging a lower price raises political support, as voters always like lower prices, ceteris paribus. It follows that a public enterprise manager’s political support is increased by setting a price below that which maximizes profit. We conclude that a public enterprise will charge a lower price than a private enterprise. For a closely related analysis, see the discussion surrounding figure 10.2 in chapter 10.

A second result is that price will be lower for voters than nonvoters. Consider a public enterprise located in political jurisdiction A but serving residents of both A and jurisdiction B. The owners of the public firm are just the voters in A. By setting a higher price for

consumers in B, the manager can increase profit and reduce taxes, thereby raising political support. This purpose is achieved without having to increase the price charged to voters (that is, consumers in A). Of course, there is an upper bound to the price that can be charged to residents of B, as they can always decide to form their own public enterprise or have a private enterprise supply them.

A third prediction from the Peltzman model is that a public enterprise will pursue less price discrimination than a private enterprise in order to provide a more uniform treatment of consumers. Differential treatment can reduce political support because it alienates certain voters. Thus, relative to a private enterprise, one would expect less discrimination between industrial and residential users of the service as well as less discrimination over time-of-day usage.

Although managers certainly desire political support, it is an overly narrow perspective to think that they act simply so as to maximize political support. Managers also value non-pecuniary benefits and leisure, not just income. For the case of a private enterprise, we argued that the capital market is an important force in constraining the manager in her desire to divert resources to nonpecuniary benefits and to shirk in performing her duties. The threat of the firm’s being acquired and current management’s being replaced can be an effective deterrent against the manager of a private enterprise’s acting too inefficiently.

Perhaps the most important difference between public and private enterprise is that there is no comparable disciplining force faced by the manager of a public enterprise. The reason is that there is no feasible mechanism by which the ownership of a public enterprise can be transferred. An owner of a public firm can forgo his share of ownership only by moving to another political jurisdiction. Hence, there is no comparable mechanism to concentrate ownership as occurs in the capital market. The fact that owners can shift their ownership to another jurisdiction in response to poor performance is unlikely anyway. The performance of public enterprise is just one factor among many that influence an individual’s decision as to where to live.

The significance of the nontransferability of ownership is that it gives the manager of a public enterprise considerably more discretion than his private enterprise counterpart in pursuing his own personal interests. To make his work environment more pleasant, a manager may choose to minimize labor strife by providing higher wages. Similarly, he may choose to overinvest in capacity to prevent potential shortages and consumer strife. It has also been suggested that a manager will choose simpler pricing schedules and adjust them less frequently than a manager of a private enterprise. All these actions tend to increase nonpecuniary benefits from the job.

**Comparison of Public and Private Enterprise**

Relative to an unregulated private enterprise, our analysis suggests that a public enterprise will price lower, practice less price discrimination, and earn lower profits. There also seems
to be a plausible case for public enterprise to be less efficient. A manager of a public enterprise has a tendency to use more capital and labor in order to reap nonpecuniary benefits, such as fewer consumer complaints and an absence of labor strife.

These differences in behavior are generated by two factors. First, income and job tenure are raised by increasing firm profit for the manager of a private enterprise but are raised by increasing political support for a public enterprise’s manager. This difference results in lower prices for a public enterprise and greater inefficiency, for example, overinvesting in product reliability to gain consumer-voter support. The second factor is that the manager of a private enterprise is threatened by the disciplining force of the capital market. If she performs poorly, there is a mechanism by which to replace her even though ownership may be widely dispersed. Because of the nontransferability of ownership in the case of a public enterprise, there is no similar constraint on the manager of a public enterprise. She then has greater discretion to use resources to maximize her own utility rather than social welfare.

Of greater relevance to our objective, however, is the comparison between a public enterprise and a regulated private enterprise. How this comparison differs from the preceding depends very much on the effectiveness of the regulatory constraint. If regulation is not very binding on the firm, then the preceding analysis applies. However, suppose that the regulatory constraint is binding. We know that regulation will cause price and profit to be lower, just as we found for a public enterprise. Whether price and profit differ much between the two institutions is then an empirical question. This issue will be addressed shortly.

In contrast, there is no reason to expect a regulated private enterprise not to practice price discrimination extensively. This is a potential source of difference from a public enterprise. We know by the Averch-Johnson effect that a private enterprise that is subject to rate-of-return regulation also has a tendency to overcapitalize. However, it still does not have the discretion to act inefficiently, like a public enterprise. Even for a regulated private enterprise, the capital market presents a constraining force on inefficient behavior. Furthermore, any increase in cost between rate hearings will reduce firm profit. Thus, one would expect a regulated private enterprise to be more efficient than a public enterprise.

According to this analysis, both a regulated private enterprise and a public enterprise will experience productive inefficiencies. Rate-of-return regulation results in the private firm overcapitalizing, whereas the public firm has less incentive to act efficiently inasmuch as its manager is interested more in political support than profit. From a productive efficiency standpoint, which approach to the natural monopoly problem is preferred is then an empirical question. In the next section we will examine how regulation and public enterprise have performed in the provision of electricity in municipalities.

5. Recall that when a regulated firm overcapitalizes, it does so in order to maximize profits by increasing its rate base. Although the capital-labor ratio is inefficient from a production perspective, the resulting ratio is the one that maximizes profits.
Municipal Electric Utilities

Historically, there has been significant governmental involvement in the generation and distribution of electricity at the federal, state, and local levels. As much as 25 percent of supply has come from public power systems. Certainly the best-known federal project is the Tennessee Valley Authority (TVA), created in 1933. There are also large power systems at the state level in both New York and Nebraska.

The role of municipal electric utilities is quite different from the federal and state power projects. Municipal utilities tend to be small and often only distribute electricity. Rather than generate it themselves, they frequently purchase it in the wholesale power market. For example, the TVA sells about one-half of its power to distributors, of which about a hundred are municipal utilities. Though small, municipal utilities are numerous and have been responsible for as much as 20 percent of sales of electricity in the United States.6

Pricing Behavior

If the degree of regulatory constraint is not too severe, our analysis predicts that a public enterprise will set lower prices than would a regulated private enterprise. By setting a price below the profit-maximizing level, the manager of a public enterprise increases political support and, as a result, income and job tenure.

The evidence on the effect that regulation has on electricity prices is mixed. In a widely cited study using data from 1912–1937, George Stigler and Claire Friedland analyzed the pricing behavior of regulated and unregulated privately owned electric utilities.7 Controlling for such factors as fuel cost and per capita income, they found that although regulation did tend to reduce prices, the effect was rather small and insignificant. In light of the poor information that regulatory agencies typically have about demand and, especially, cost functions, such a finding is not surprising. Because the firm has a significant information advantage, a regulatory agency may often have little choice but to accept the proposed tariff. Other studies, however, have found electricity prices to be severely constrained by regulation. In some cases it was estimated that the monopoly price was 20 to 50 percent higher than actual regulated prices.8


If regulation does not have much of an effect on pricing behavior, we would expect a public enterprise to set lower prices than a privately owned utility. A study by Thomas Moore strongly supports this hypothesis. Moore estimated the monopoly price for a firm sample that included privately owned regulated utilities and publicly owned utilities. Using data from 1962, he then compared these prices to the actual prices. Moore found the prices of regulated private utilities to be 5 to 6 percent below the monopoly price, indicating a small but negative effect of regulation on price. In contrast, the prices of the publicly owned utilities were 10 to 22 percent below the monopoly prices.

A second hypothesis is that a public enterprise pursues less price discrimination. The reason is that more uniform pricing schedules reduce the alienation of consumer groups and this tends to increase political support. In addition, because they are easier to handle, a more uniform pricing schedule increases nonpecuniary benefits from the job. Sam Peltzman did indeed find that publicly owned utilities used a smaller number of rate schedules on average than private utilities (see table 14.1). Although both types of utilities did practice price discrimination, it was more extensive for privately owned utilities. Also consistent with the belief that public enterprise managers have greater discretion to pursue an easier life, Peltzman found that public enterprises adjusted price less frequently.

### Allocative Efficiency Comparison

Because publicly owned utilities set lower prices, it is tempting to conclude that they perform better than regulated privately owned utilities on an allocative efficiency basis. By charging a price closer to that which maximizes profit, privately owned utilities cause greater deadweight welfare losses. However, even though average price may be higher, there can still be

---


---

**Table 14.1**

<table>
<thead>
<tr>
<th></th>
<th>Average Number of Rate Schedules per City Served by</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Private Utilities (69 Cities)</td>
</tr>
<tr>
<td>Residential service</td>
<td>1.884</td>
</tr>
<tr>
<td>Nonresidential service</td>
<td>6.507</td>
</tr>
<tr>
<td>Total</td>
<td>8.391</td>
</tr>
</tbody>
</table>

greater welfare associated with privately owned utilities because of more extensive price discrimination.

To understand this point, let us consider the two demand curves in figure 14.1. The unit cost of supplying good 1 (2) is denoted $c_1(c_2)$. For the sake of simplicity, suppose that a firm discriminates between these two markets and sets price equal to $P'_1$ in market 1 and $P'_2$ in market 2. Now consider a nondiscriminating price $\bar{P}$ such that the following relationship holds:

$$\bar{P} < \frac{Q'_1}{Q'_1 + Q'_2} P'_1 + \frac{Q'_2}{Q'_1 + Q'_2} P'_2$$

so that average price is lower. Comparing prices $P'_1$ and $\bar{P}$ in market 1, welfare is higher with $P'_1$ as measured by the sum of triangle $abc$ and $(P'_1 - c)(Q'_1 - \bar{Q})$. In addition, a price of $\bar{P}$ in market 2 yields lower welfare than $P'_2$ by triangle $def$ less triangle $dgh$. The discriminatory pricing schedule yields higher welfare even though it results in a higher average price. For this example, setting a uniform price results in price being too high in market 1 (as marginal cost is low) and too low in market 2 (as marginal cost is high). Price discrimination allows consumers to get the right signals about which service is more costly.

Empirical evidence by Peltzman shows that greater price discrimination by privately owned utilities resulted in higher average sales of electricity per customer. In estimating the effects of the type of ownership on average sales per customer, he first controlled for other factors that could affect the difference in average sales. These factors included income, population size, regional differences, and average price. Once holding these factors constant, Peltzman
found that average sales per customer were higher for privately owned utilities than for public utilities.

Given this evidence, it is difficult to determine which institution generates greater allocative efficiency. Although publicly owned utilities do seem to set lower average prices, privately owned utilities appear to practice greater price discrimination.

**Productive Efficiency Comparison**

Our previous analysis suggested that both regulated private utilities and publicly owned utilities tend to overcapitalize. The former will do so in order to increase profit through the expansion of their rate base under rate-of-return regulation. A public utility overinvests in capital in order to produce the attributes that garner political support, for example, increasing capacity in order to reduce the likelihood of blackouts and brownouts. We also expect a public enterprise to act more inefficiently than a private enterprise because of greater managerial discretion.

In his 1970 study, Thomas Moore estimated the ratio of peak demand to total capacity for twenty-seven publicly owned and thirty-six privately owned utilities. He found that ratio to be smaller for publicly owned firms, indicating that they had greater capacity relative to peak demand. Though there is certainly evidence supporting the Averch-Johnson effect (see chapter 12 for details), the Moore study suggests that the incentive to overcapitalize is greater for public enterprises.

To gauge productive efficiency, Thomas DiLorenzo and Ralph Robinson, in their 1982 study, measured average productivity of labor for both public and private utilities. They found, on average, that one laborer generated 15.734 kilowatt-hours of electricity in a publicly owned utility. In contrast, the comparable figure for a private utility was 16.566. A publicly owned utility was found to be less efficient, but this difference was not statistically significant.

Contradictory evidence is provided in a study by Donn Pescatrice and John Trapani. They estimated cost and input demand functions with a data set made up of thirty-three private utilities and twenty-three public utilities. Their empirical results showed that a private utility’s costs were 23.5 percent higher in 1965 and 32.9 percent higher in 1970 than the costs incurred by a public utility.

**Assessment of Private versus Public Utilities**

It is clear that the evidence concerning the relative efficiency of regulated privately owned utilities and publicly owned utilities is mixed. Nevertheless, a survey of comparative studies


of industries ranging from electric utilities to refuse collection to weather forecasting provides general support for the hypothesis that there is greater productive efficiency with private enterprise.\textsuperscript{12} Most of these studies conclude that publicly owned firms are less efficient than privately owned firms.

Perhaps the major advantage of a privately owned firm is that it is subject to the disciplining force of the capital market. A study by Louis DeAlessi supports the hypothesis that the constraints placed on the manager of a public enterprise are not as great.\textsuperscript{13} He assessed relative job tenure over the period 1962–1971 for a sample of a hundred private utilities and a hundred public utilities. Table 14.2 presents some of his results. As is quite apparent, job tenure was considerably greater for the manager of a public enterprise; twice as many public firms as private firms did not change managers over the ten-year period. This evidence is consistent with the hypothesis that the manager of a public enterprise is subject to fewer constraints and devotes more resources to gain political support in order to raise job tenure.

### Airlines

In the case of municipal electric utilities, the empirical evidence is inconclusive concerning the hypothesis that regulated privately owned firms are more efficient than publicly owned firms. In his 1977 study of the Australian airline industry, David Davies derived strong evidence to support the hypothesis that unregulated private firms are more efficient than public firms.\textsuperscript{14}

\begin{table}[h]
\centering
\caption{Frequency of Change in Top Executive of Firms in the Electric Power Industry, 1962–1971}
\begin{tabular}{|c|c|c|c|c|c|c|c|c|}
\hline
Ownership Category & 0 & 1 & 2 & 3 & 4 & 5 & \textbf{Size} \\
\hline
Public & 36 & 42 & 17 & 3 & 1 & 1 & 100 \\
Private & 18 & 59 & 18 & 4 & 1 & 0 & 100 \\
\hline
\end{tabular}
\end{table}


At the time of the study, Australia had two interstate airlines, Trans-Australian Airlines (TAA) and Ansett Australian National Airway. The former is a public enterprise, and the latter is privately owned. Using data from 1958–1974, Davies compared the productive efficiency of these two types of ownership.

Davies chose a good pair of firms to test the relative efficiency hypothesis as, except in terms of ownership, TAA and Ansett are quite comparable. They serve similar routes and have comparable frequency of flights. Both charge identical passenger and freight rates and have a similar mix of services. Given these similarities, it is reasonable to believe that any differences in efficiency are probably due to the type of ownership.

Davies estimated three measures of productivity. They are presented in Table 14.3. For all three measures, the privately owned firm, Ansett, experienced greater efficiency than publicly owned TAA. Over the entire period of 1958–1974, an Ansett employee, on average, carried twice as much freight and mail as an employee of TAA. Similarly, passengers carried and revenue earned per employee were higher for the privately owned firm.

This is a good study with which to conclude our analysis of public enterprise because it presents clear and strong results. Although no single study is conclusive, the Australian airline industry does support the hypothesis that public enterprises are less efficient than privately owned firms.

**Privatization**

Government enterprise as a solution to the natural monopoly problem has become increasingly out of fashion. Though it has never been widely used in the United States, government ownership has been extensively used in other countries such as the United Kingdom. However, the last 30 years have witnessed a reversal, as many governments have pursued privatization, in which a government-owned enterprise is sold to private investors. When the industry is electric distribution or local telephone or some other natural monopoly, this conversion to a private company interested in profits is typically complemented with regulatory oversight, as described in Chapter 12.15

The privatization trend is based on more than a change in society’s preferred solution to the natural monopoly problem. It is a manifestation of a broader skepticism that government control is the best solution to most societal problems. In fact, privatization has been most extensive in former communist countries that had previously prohibited private ownership. The privatization wave can be traced back to 1973, when Chile, after the military takeover

---

Table 14.3
Productivity Measures for the Australian Airline Industry

<table>
<thead>
<tr>
<th>Year</th>
<th>Tons of Freight and Mail Carried per Employee</th>
<th>Passengers Carried per Employee</th>
<th>Revenue Earned per Employee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ansett Airline (Private Firm)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1958–59</td>
<td>10.69</td>
<td>282</td>
<td>$7,172</td>
</tr>
<tr>
<td>1959–60</td>
<td>10.77</td>
<td>309</td>
<td>7,758</td>
</tr>
<tr>
<td>1960–61</td>
<td>10.96</td>
<td>337</td>
<td>8,679</td>
</tr>
<tr>
<td>1961–62</td>
<td>10.84</td>
<td>331</td>
<td>8,425</td>
</tr>
<tr>
<td>1962–63</td>
<td>11.09</td>
<td>316</td>
<td>8,510</td>
</tr>
<tr>
<td>1963–64</td>
<td>11.06</td>
<td>324</td>
<td>9,071</td>
</tr>
<tr>
<td>1964–65</td>
<td>12.14</td>
<td>352</td>
<td>9,705</td>
</tr>
<tr>
<td>1965–66</td>
<td>11.08</td>
<td>354</td>
<td>10,479</td>
</tr>
<tr>
<td>1966–67</td>
<td>10.34</td>
<td>348</td>
<td>10,829</td>
</tr>
<tr>
<td>1967–68</td>
<td>9.57</td>
<td>363</td>
<td>12,080</td>
</tr>
<tr>
<td>1968–69</td>
<td>9.54</td>
<td>392</td>
<td>13,185</td>
</tr>
<tr>
<td>1969–70</td>
<td>9.35</td>
<td>414</td>
<td>14,118</td>
</tr>
<tr>
<td>1970–71</td>
<td>8.75</td>
<td>417</td>
<td>15,558</td>
</tr>
<tr>
<td>1971–72</td>
<td>8.82</td>
<td>437</td>
<td>17,280</td>
</tr>
<tr>
<td>1972–73</td>
<td>9.07</td>
<td>468</td>
<td>17,829</td>
</tr>
<tr>
<td>1973–74</td>
<td>10.02</td>
<td>532</td>
<td>21,461</td>
</tr>
<tr>
<td>Average</td>
<td>10.25</td>
<td>373</td>
<td>12,009</td>
</tr>
<tr>
<td>Trans-Australian Airlines (Public Firm)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1958–59</td>
<td>4.42</td>
<td>217</td>
<td>$6,104</td>
</tr>
<tr>
<td>1959–60</td>
<td>4.57</td>
<td>259</td>
<td>7,016</td>
</tr>
<tr>
<td>1960–61</td>
<td>4.52</td>
<td>228</td>
<td>7,052</td>
</tr>
<tr>
<td>1961–62</td>
<td>4.64</td>
<td>246</td>
<td>7,367</td>
</tr>
<tr>
<td>1962–63</td>
<td>4.69</td>
<td>255</td>
<td>7,726</td>
</tr>
<tr>
<td>1963–64</td>
<td>4.83</td>
<td>274</td>
<td>8,093</td>
</tr>
<tr>
<td>1964–65</td>
<td>5.02</td>
<td>287</td>
<td>8,553</td>
</tr>
<tr>
<td>1965–66</td>
<td>4.88</td>
<td>294</td>
<td>9,072</td>
</tr>
<tr>
<td>1966–67</td>
<td>5.11</td>
<td>316</td>
<td>9,954</td>
</tr>
<tr>
<td>1967–68</td>
<td>5.41</td>
<td>337</td>
<td>11,033</td>
</tr>
<tr>
<td>1968–69</td>
<td>5.34</td>
<td>356</td>
<td>11,734</td>
</tr>
<tr>
<td>1969–70</td>
<td>5.80</td>
<td>390</td>
<td>13,146</td>
</tr>
<tr>
<td>1970–71</td>
<td>5.70</td>
<td>399</td>
<td>14,522</td>
</tr>
<tr>
<td>1971–72</td>
<td>5.63</td>
<td>414</td>
<td>15,644</td>
</tr>
<tr>
<td>1972–73</td>
<td>5.62</td>
<td>449</td>
<td>16,541</td>
</tr>
<tr>
<td>1973–74</td>
<td>6.06</td>
<td>496</td>
<td>19,183</td>
</tr>
<tr>
<td>Average</td>
<td>5.14</td>
<td>326</td>
<td>10,740</td>
</tr>
</tbody>
</table>

of Augusto Pinochet, reversed the nationalization of many industries that had occurred in the previous three years under the socialist government. In the ensuing decades, countries that pursued aggressive privatization included Argentina, Poland and much of Eastern Europe, and Russia (of the former Soviet Union). It has also occurred, however, in capitalist countries such as the United Kingdom. Substantively begun by the Conservative government of Margaret Thatcher, the British government sold off many companies; beginning with British Telecom in 1984 and ending with British Energy in 1996 (see table 14.4).

In the case of a capitalist country, what purposes does privatization serve? If public enterprises are less efficient, one would anticipate social welfare gains from privatization. The evidence we previously reviewed does suggest there are allocative and productive efficiency gains when a government monopoly is replaced with a regulated private monopolist. A second

<table>
<thead>
<tr>
<th>Organization</th>
<th>Year of First Share Sale</th>
<th>Industry</th>
</tr>
</thead>
<tbody>
<tr>
<td>British Petroleum</td>
<td>1979</td>
<td>Oil</td>
</tr>
<tr>
<td>National Enterprise Board Investment</td>
<td>1980</td>
<td>Various</td>
</tr>
<tr>
<td>British Aerospace</td>
<td>1981</td>
<td>Aerospace</td>
</tr>
<tr>
<td>Cable &amp; Wireless</td>
<td>1981</td>
<td>Telecoms</td>
</tr>
<tr>
<td>Amersham International</td>
<td>1982</td>
<td>Scientific goods</td>
</tr>
<tr>
<td>National Freight Corporation</td>
<td>1982</td>
<td>Road transport</td>
</tr>
<tr>
<td>Britoil</td>
<td>1982</td>
<td>Oil</td>
</tr>
<tr>
<td>British Rail Hotels</td>
<td>1983</td>
<td>Hotels</td>
</tr>
<tr>
<td>Associated British Ports</td>
<td>1983</td>
<td>Ports</td>
</tr>
<tr>
<td>British Leyland (Rover)</td>
<td>1984</td>
<td>Automobile producer</td>
</tr>
<tr>
<td>British Telecom (BT)</td>
<td>1984</td>
<td>Telecoms</td>
</tr>
<tr>
<td>Enterprise Oil</td>
<td>1984</td>
<td>Oil</td>
</tr>
<tr>
<td>Sealink</td>
<td>1984</td>
<td>Sea transport</td>
</tr>
<tr>
<td>British Shipbuilders &amp; Naval Dockyards</td>
<td>1985</td>
<td>Ship building</td>
</tr>
<tr>
<td>National Bus Company</td>
<td>1986</td>
<td>Transport</td>
</tr>
<tr>
<td>British Gas</td>
<td>1986</td>
<td>Gas</td>
</tr>
<tr>
<td>Rolls-Royce</td>
<td>1987</td>
<td>Aero-engines</td>
</tr>
<tr>
<td>British Airports Authority</td>
<td>1987</td>
<td>Airports</td>
</tr>
<tr>
<td>British Airways</td>
<td>1987</td>
<td>Airlines</td>
</tr>
<tr>
<td>Royal Ordnance Factories</td>
<td>1987</td>
<td>Armaments</td>
</tr>
<tr>
<td>British Steel</td>
<td>1988</td>
<td>Steel</td>
</tr>
<tr>
<td>Water</td>
<td>1989</td>
<td>Water</td>
</tr>
<tr>
<td>Electricity distribution</td>
<td>1990</td>
<td>Electricity</td>
</tr>
<tr>
<td>Electricity generation</td>
<td>1991</td>
<td>Electricity</td>
</tr>
<tr>
<td>Trust Ports</td>
<td>1992</td>
<td>Ports</td>
</tr>
<tr>
<td>Coal industry</td>
<td>1995</td>
<td>Coal</td>
</tr>
<tr>
<td>Railways</td>
<td>1995–97</td>
<td>Railways</td>
</tr>
<tr>
<td>Nuclear energy</td>
<td>1996</td>
<td>Electricity</td>
</tr>
</tbody>
</table>

objective comes from the revenue that is raised from privatization. Privatization typically entails creating ownership shares in the government enterprise and selling those shares to private investors.

Before one jumps to the conclusion that the revenue from the sale of shares is a windfall to the government, keep in mind that the government has lost the profit that the enterprise previously generated; it now goes to private investors. Indeed, if the sale price of the enterprise equals the expected present value of the profit stream, then the government’s net worth has not changed from privatization. This observation suggests there is no purpose to privatization on revenue grounds. But there are two caveats to offer. First, if the privatized firm is anticipated to be more efficient and thus to generate more profit, then the sale price—which should equal the expected present value of the future profit stream under privatization—should exceed the value of the enterprise had it not been privatized. This difference is a real gain to government coffers. And if this additional revenue allows for a reduction in distortionary taxes such as the income and sales tax, there are further efficiency gains. So, one objective of privatization can be to raise government revenue. Second, a government might gain political support from privatization even when there are no efficiency gains. For suppose the revenue from the sale of assets goes directly into the government’s current budget from the sale of assets. Though future revenues are lower, because the government no longer has the profit generated by the enterprise, the government might experience gains in political support from a stronger current budget. This does rely on improper accounting and lack of voter sophistication, both of which are plausible. On these grounds, privatization can serve the interests of politicians, not society.

An important decision associated with privatization concerns the allocation of the government enterprise’s assets. How many privately owned firms should be created, and how should assets be distributed among them? There decisions can have long-run implications for industry structure and allocative efficiency. In the privatization of electric power in England and Wales, government assets were sold off to create three companies. The fossil-fuel plants were now owned by private companies National Power and PowerGen, with shares of generating capacity of 52 percent and 33 percent, respectively. The government continued to own nuclear power, which made up the remaining 15 percent of capacity. The postprivatization structure was then highly concentrated. When Spain was considering how to privatize the government-controlled electric power company Endesa, the industry was fairly concentrated, as Endesa had 37 percent of generating capacity and the privately owned Iberdrola had 28 percent. It would have been natural to have created multiple companies out of Endesa in order to reduce potential market power problems (see chapter 12 for details on market power in wholesale energy markets). Instead, the government increased concen-

tration! It allowed Endesa to acquire two small generators, which raised its share of capacity to 50 percent. One reason for increasing industry concentration is that private investors anticipate a higher future profit stream, since firms will be able to set higher prices, and this means that private investors are willing to pay more for shares of the newly created company. The government then raises more revenue, but at the cost of higher deadweight welfare losses.

In light of the vast privatization that has occurred, one might wonder about the merit of devoting a chapter to public enterprise as a solution to the natural monopoly problem. The reason for doing so is simple. When many students learn about the natural monopoly problem, a common suggestion is nationalization: “Why doesn’t the government just own the firm and set price equal to marginal cost?” What this chapter has hopefully shown is that such a solution is idealistic and that what happens in practice raises doubts as to whether government control is superior to a regulated privately owned firm.

Summary

In comparing the three approaches to the natural monopoly problem, we can draw some general conclusions. From a theoretical perspective, it seems that franchise bidding is preferred to regulation and regulation is preferred to public enterprise. This ordering is also the same in terms of the degree of competitive constraint forced on the firm. Competition is strongest in the case of franchise bidding, as firms compete for the franchise through quality and price. A regulated privately owned firm does not face such competition but is constrained by the threat of takeover and the replacement of existing management if the firm performs poorly. Finally, a public enterprise does not face either type of constraint, which thus allows it greater discretion to stray from efficiency.

In practice, it is more difficult to provide an ordering of the alternatives. As we have seen in the case of cable television, the role of government in franchise bidding is not as different from that of regulation as one might have believed from the theory. Similarly, although regulated private electric utilities appear to perform more efficiently than publicly owned utilities, the evidence is not strong. In addition, private utilities tend to set prices that are too high. Offsetting the high prices is the fact that they practice greater price discrimination.

It is difficult to draw any definitive conclusions, especially because franchise bidding is a relatively new policy. If one alternative had to be chosen to offer the greatest potential, it would probably be franchise bidding. It provides the greatest role for competitive forces, and it is such forces that we count on for achieving social optimality. If franchise bidding is given

greater experience, we are more likely to be able to determine whether it is the best method for handling natural monopolies. Thus, more definitive conclusions than the ones we have offered may present themselves in the near future.

Questions and Problems

1. Do you think that social welfare is higher with a regulated private enterprise or a public enterprise? Does your answer depend on the industry? What about for the distribution of electricity?

2. Why do public enterprises practice less price discrimination than regulated private enterprises? Does it have anything to do with the Robinson-Patman Act?

3. Privatization is a central economic issue for Eastern Europe. Should all public enterprises be privatized? Of those privatized, which should be regulated?

4. Why do public enterprises tend to have higher costs than private enterprises? Do you think this reason is sufficient to always prefer regulated private enterprises over public enterprises?

5. Compare regulation, franchise bidding, and public enterprise. Discuss the advantages and disadvantages of each. If you had to choose one method for all natural monopolies, which would it be?

6. Compare the change in price for an unregulated private enterprise, a regulated private enterprise, and a public enterprise in response to
   a. The legislature becoming more proconsumer.
   b. An increase in fixed costs.
   c. A fall in marginal cost.

7. Between a regulated private enterprise and a public enterprise, which do you think would be quicker to adopt a cost-reducing innovation? What about a product-improving innovation?

8. In 1990, Britain privatized its electric power industry by converting government enterprises into privately owned enterprises. Interestingly, a recent study found that the average chief executive of Britain’s twelve regional electricity distribution companies experienced nearly a threefold salary increase in the two years following industry privatization. This occurred in spite of the fact that the identities of the chief executives largely remained the same. During these two years, average firm revenue was fairly constant, though the average number of employees was significantly reduced and the value of firm assets rose. What do you think was the reason for this large increase in salary?

Handling rate cases is perhaps the most commonly performed duty of a regulatory agency, but it is by no means its only one. In that regulators are ultimately responsible for maintaining a healthy industry and maximizing social welfare, their tasks can be very far-ranging. This chapter is concerned with examining some of the more important tasks related to the evolution of a regulated industry.

The first task we will investigate is the decision to maintain regulation. At any time, changes in the market can result in an industry no longer being a natural monopoly. An important task for a regulatory agency is to identify when such a transformation has taken place and to then open the industry up to entry and eliminate price controls. The next few sections consider how an industry may be transformed from a natural monopoly into a potentially competitive market and provide an assessment of the welfare effects of different regulatory policies. These concepts will be applied to events that have taken place in the intercity telecommunications market over the last few decades.

A second issue regulators may face is whether to allow a regulated monopolist to enter unregulated markets. Referred to as the separations issue, it has arisen in recent years in such industries as telecommunications, cable television, and electric power. After discussing the benefits and costs of different regulatory policies to handle this situation, we will use these concepts to understand the 1982 decision to break up AT&T.

Transformation of a Natural Monopoly

The formulation of regulatory policy would be a considerably simpler task if the environment within which it was made was static and never changing. Having determined that a particular industry was a natural monopoly, regulators would then need to determine the socially optimal price and require the monopolist to meet all demand at that price. Once regulation was put in place, the only role for regulators would be to enforce regulatory policy.

Unfortunately, the environment in which economic agents act is anything but static. Demand shifts over time for such reasons as changes in consumer preferences and income and the vagaries of business cycle. Innovations are constantly taking place that alter the production technology or introduce unregulated substitute products. In response to these changes, regulators must adapt policy. With regard to monopoly regulation, we can identify two potential dilemmas resulting from changes in the environment. First, the shifting of cost and demand functions changes the socially optimal price. Regulators must adjust the regulated price in response to these events. A second and perhaps more troublesome problem arises when shifts in these functions are so severe as to call into question the need for monopoly regulation. In other words, cost and demand conditions may change to the point that the industry is no longer a natural monopoly.
This section will explore the sources of such a transformation and assess the effects of different regulatory policies when such a transformation has been thought to occur. We will then apply these concepts to events that have taken place in the intercity telecommunications market since World War II. Before undertaking this investigation, let us first briefly review the conditions under which monopoly regulation is socially desirable.

Basis for Natural Monopoly Regulation

A natural monopoly exists at an output rate $Q^0$ if the total cost of producing $Q^0$ is minimized by having a single firm produce. If this condition holds, then the cost function, $C(Q)$, is said to be subadditive at $Q^0$. For example, if a cost function is subadditive at $Q^0$, then it is cheaper for one firm to produce $Q^0$ than to have two firms each produce half of $Q^0$. If social welfare is maximized by supplying an amount $Q^0$ of this good, then the efficient market structure is to have a single firm, as it minimizes the resources used in producing $Q^0$.

In order to assess whether a particular industry is a candidate for regulation, we need to determine whether the firm cost function is subadditive at the socially optimal industry output. This step requires both cost and demand information, where the latter is needed in order to calculate the socially optimal output. There is one exception, however, and this is when $C(Q)$ is subadditive at all quantities. In that case, cost efficiency requires single-firm production regardless of the level of demand. We know from chapter 11 that for the single-product firm, scale economies imply that the cost function is subadditive. Referring back to figure 11.1, a cost function is exhibited for which scale economies are never exhausted (that is, average cost is always declining). Even without knowledge of the market demand curve, this industry is a natural monopoly, and that fact makes it a prime candidate for regulation.

If the firm cost function exhausts economies of scale at some output, then demand information is required in order to determine the appropriateness of regulation. Consider the U-shaped average cost curve in figure 15.1. Scale economies are exhausted at a quantity of $\hat{Q}$. If the market demand curve is $D(P)$, this industry is a natural monopoly because $D(P)$ intersects $AC(Q)$ at $Q^*$ and at that quantity average cost is falling, all of which means that the cost function is subadditive at $Q^*$. If the regulators were restricted to using linear pricing (see chapter 11 for details), the optimal policy would be to allow only one firm to operate and to set a price equal to $P^*$ (that is, average cost pricing). This regulatory policy is socially optimal for two reasons. First, $P^*$ is the price that maximizes social welfare subject to the constraint that firm profit is nonnegative. Second, total cost is minimized at $Q^*$, as only one firm is operating in the industry.

To understand the importance of the role of market demand in determining the proper regulatory policy, consider the same average cost curve in figure 15.1 but now assume the demand curve is $\hat{D}(P)$. The socially optimal price is then $\hat{P}$, which equals minimum average cost of $AC(Q)$. At that price, market demand is $3\hat{Q}$, which is sufficient to support three
firms at the efficient size of $Q^\hat{\phantom{}}$. There would appear to be little basis for regulating such an industry.

The general rule of thumb is that monopoly regulation is appropriate when the minimum efficient size of the firm (that is, the lowest quantity at which average cost is minimized) is approximately equal to or larger than market demand at the socially optimal price. In figure 15.1, when the demand curve is $D(P)$, the efficient firm size, $Q^\hat{\phantom{}}$, is larger than market demand (at a price equal to average cost) of $Q^*$, so that a regulated monopoly is likely to be appropriate for this industry. In contrast, when $\bar{D}(P)$ is the demand curve, market demand at a price of $\bar{P}$ is three times the efficient firm size. In that situation, an unregulated environment is likely to be the appropriate policy.

A peculiar feature of the preceding example is that the socially optimal supply is an integer multiple of minimum efficient scale. Specifically, when demand is $\bar{D}(P)$ market demand at a price equal to minimum average cost is exactly equal to three times minimum efficient scale. Because one would not generally expect such a coincidence to occur, it is important that we show our analysis is robust to it.
Interestingly, engineering estimates suggest that a U-shaped average cost curve is not typical. Instead, average cost declines until minimum efficient scale is achieved, and is then flat for some range of output. An example of such a cost function is depicted in figure 15.2. Average cost is minimized as long as output lies between $Q$ and $\bar{Q}$. Note that $Q$ is then minimum efficient scale. For this average cost function, if the market demand function is $D(P)$, then a natural monopoly exists, but it does not with $\bar{D}(P)$. If, for example, $Q < Q^o/2 < \bar{Q}$, then industry cost is minimized by having two firms, each producing $Q^o/2$. We then find that our earlier result holds more generally if the average cost curve has a range of quantities that minimize average cost. Rather than use such an average cost function in future examples in this chapter, we will use the standard U-shaped average cost curve, though remembering that results are robust if we consider a more realistic average cost function of the type in figure 15.2.
Sources of Natural Monopoly Transformation

Suppose the cost and demand conditions for a particular market call for monopoly regulation. In this subsection, we want to consider changes over time that would make continued regulation unjustified. Alternatively, this issue could be analyzed from the opposite perspective. That is, one could investigate how an industry can be transformed into (rather than out of) a natural monopoly. The sources of the transformation would be the same, only their direction would differ. The decision to analyze the case when an industry is transformed into a potentially competitive market was determined by our later application to the intercity telecommunications market.

Demand Side

The sources of transformation are those events that make efficient firm size smaller relative to market demand at the socially optimal price. From the demand side, an industry can be transformed away from being a natural monopoly if the market demand curve shifts upward sufficiently, since it will raise the socially optimal output. For example, in figure 15.1, if the demand curve shifted from $D(P)$ to $\overline{D}(P)$ over time, the optimal output would increase from $Q^*$ to $3\hat{Q}$. Given an efficient firm size of $\hat{Q}$, the basis for monopoly regulation would no longer exist.

There are, of course, many sources of exogenous increase in demand. These sources include, for example, an exogenous change in consumer preferences toward the good, an increase in consumer income (if the product is a normal good), a reduction in the price of a complement, or even the development of a complementary product that raises the value attached to the good by consumers.

Cost Side

Recall that the firm cost function is determined by the best available production technology and current input prices. It follows that the average and marginal cost curves will change when there is a technological innovation that alters the best available technology or when there is a change in input prices. A change in the cost structure can affect the conditions for the industry to be a natural monopoly in two ways. First, the efficient firm size can change. Second, a change in the cost structure can affect the socially optimal output. The qualitative nature of these two effects is very much dependent on how exactly the cost function is affected by the change in technology or input prices.

To analyze this issue, let us recall that total cost is composed of fixed costs, denoted $FC$, and variable costs, denoted $VC(Q)$:

$$C(Q) = FC + VC(Q).$$

The average cost of a firm is then
$AC(Q) = (FC/Q) + AVC(Q)$

where $AVC(Q)$ is average variable cost. Let us assume that $AVC(Q)$ is increasing in output. On the one hand, as $Q$ increases, fixed costs are spread out over more quantity, so that $FC/Q$ falls. On the other hand, average variable cost is increasing in quantity. When $Q$ is small, fixed costs loom large relative to variable costs, so that the reduction in average fixed cost from increasing $Q$ dominates the rise in average variable cost. Hence, when $Q$ is initially low, average cost declines as $Q$ increases. As output is increased further, the reduction in average fixed cost from increasing output becomes less as output is already spread out over many units. If average variable cost rises sufficiently fast, then eventually average cost will stop falling. Minimum efficient scale is achieved when the slope of the average cost function is zero, so that average cost is minimized. As output rises above minimum efficient scale, average cost rises. This argument is what lies behind the standard U-shaped average cost curve shown in figure 15.3.

![Figure 15.3](image1)

The Effect of a Change in Fixed Costs on the Efficient Market Structure
What happens when there is a reduction in fixed costs? Clearly, there is a reduction in average cost. Since cost is lower, the socially optimal output (under average cost pricing) is higher. In addition, since fixed costs make up a smaller part of total cost, minimum efficient scale falls. This relationship is depicted in figure 15.3. A fall in fixed costs shifts the average cost curve from $AC(Q)$ to $\overline{AC}(Q)$. Minimum efficient scale falls from $\hat{Q}$ to $\overline{Q}$. Prior to the fall in fixed costs, the socially optimal price was defined where price equals average cost, $P^*$. With average cost at $\overline{AC}(Q)$, the socially optimal price equals minimum average cost, $\overline{P}$. Since demand is $2\overline{Q}$, the efficient market structure is to have two firms, not one. Thus an innovation that reduces fixed costs makes it less likely that an industry is a natural monopoly. For the case in figure 15.3, the reduction in fixed costs moves the industry from being a natural monopoly to one in which two firms can profitably exist at the socially optimal price.

Although a fall in fixed costs makes it less likely that an industry is a natural monopoly, a change in variable costs has an ambiguous effect. For example, suppose that a rise in input prices caused average variable cost to rise. A likely effect on the average cost curve is depicted in figure 15.4, where the rise in variable costs shifts the average cost curve to $\overline{AC}(Q)$. Obviously, average cost is higher. Because fixed costs now make up a smaller proportion of total cost, minimum efficient scale has fallen from $\hat{Q}$ to $\overline{Q}$. A smaller minimum efficient scale makes the industry less likely to remain a natural monopoly. However, higher average cost reduces the socially optimal output, which works in the opposite direction. The net effect depends on the elasticity of market demand. If the demand curve is highly inelastic, like $D(P)$, then the socially optimal output does not fall very much. In this case, it declines from $\hat{Q}$ to $2\overline{Q}$. Because two efficient-sized firms can profitably exist at the socially optimal price, monopoly regulation is no longer appropriate. However, if the demand curve is instead $\overline{D}(p)$, which is relatively elastic, the optimal output falls a lot, from $\hat{Q}$ to $Q'$. Hence a natural monopoly still exists, even though minimum efficient scale has fallen.

To summarize these results, we find that technological innovations and changes in input prices affect the cost and demand conditions underlying the rationale for monopoly regulation. Reductions in the fixed-cost component of the cost function clearly make the industry less suited for single-firm production by reducing efficient firm size and increasing the socially optimal industry quantity. Changes in the variable-cost function have less clear effects, as their impact depends on the elasticity of market demand.

**Regulatory Response**

Suppose that a regulated monopoly has experienced the type of change just described. Perhaps an innovation occurred that reduced fixed costs or a complementary product was developed that shifted out the market demand curve. In either case, events have taken place that called into question the social desirability of monopoly regulation. The objective of this section is to consider and briefly analyze the policy alternatives available to regulators.
Policy Alternatives

When faced with the situation just described, regulators have basically three alternative policies that they can pursue. The first alternative is to continue price and entry regulation. This is obviously the appropriate policy if they think it is very likely that the industry is still a natural monopoly. If a natural monopoly no longer exists, the pursuance of such a policy generates welfare losses from preventing competition, which would lower price and produce a more efficient allocation of production among firms. A second policy is full deregulation: allowing free entry and removing price controls. A necessary condition for this policy to be appropriate is that the industry no longer be a natural monopoly or, if it is still a natural monopoly, that scale economies be relatively small so that any resulting productive inefficiencies from having more than one firm are likely to be offset by the general benefits of competition. However, the loss of natural monopoly status may not be sufficient to justify full deregulation. It may be preferable to pursue a gradual transition from a regulated environment to a deregulated one. Not only could such a policy be justified on grounds of effi-
ciency and equity, but it also may be politically necessary, as the severe transitional pains that could arise from full deregulation may induce intense lobbying effort to prevent it. Partial deregulation may be acceptable to all concerned groups. Ultimately, of course, full deregulation would be desirable if the industry is no longer a natural monopoly.

The third policy option is to pursue a course of partial deregulation. This would entail loosening entry restrictions by either allowing free entry or instituting relatively lenient standards for entry. The key feature of partial deregulation is that some substantive controls are kept on price. One possibility is for a regulatory agency to specify minimum and maximum limits on price and allow firms considerable flexibility within those bounds. Alternatively, an agency could continue to require regulatory approval for all rate changes but pursue a policy of rate freedom except when they believe the intentions of a firm are anticompetitive, whether to drive out competitors or achieve collusion. Price regulation could either be applied equally to all firms or differentiated between the established firm and entrants. One form of such an asymmetric regulatory policy is to allow considerable price freedom to new firms while continuing to subject the established firm to the rate approval process. In the absence of continued price regulation, it is possible that the dominant position of the established firm, along with possibly greater financial resources than new firms, might allow it to effectively engage in a predatory policy of pricing below cost so as to drive out new firms. Partial deregulation can be a useful intermediate policy when there is considerable uncertainty faced by regulators over cost and demand conditions. Alternatively, it could be used as a transitional policy to ease the adjustment to full deregulation.

**Asymmetric Regulation and Creamskimming**

There are a number of distortions that could arise from a regulatory policy that treats the established firm and new firms differently. One such distortion is that it is possible for an industry to still be a natural monopoly and the established firm to be pricing so that it earns normal profit, yet a new firm could find entry profitable if it was allowed to freely set its price. The profitability of entry in that context is generated by distortions created by a regulatory policy that restricts the pricing of the established firm but not entrants.

This possibility is generated by a regulatory pricing practice known as *cross-subsidization*. Cross-subsidization occurs when the price of one product is set so as to generate additional revenues that are used to subsidize the sales of a second product offered by the regulated firm. This practice has been observed in many regulated industries including telephone, airlines, and railroads. It usually entails setting the price on a low-cost product too high and the price on a high-cost product too low. Revenue earned from the former is used to help cover the cost of supplying the latter.¹

To examine the implications of cross-subsidization, suppose that the regulated monopolist offers two products, X and Y. $Q_X$ is defined to be the number of units of product X produced, and $Q_Y$ is defined analogously. The total cost of producing both products is represented by $C(Q_X, Q_Y)$. We want to suppose that a multiproduct natural monopoly exists. For this reason, it is assumed that the cost function has *economies of scope*. Economies of scope mean that it is cheaper for one two-product firm to produce $Q_X$ and $Q_Y$ than to have one firm produce $Q_X$ and a second firm produce $Q_Y$. In other words:

$$C(Q_X, Q_Y) < C(Q_X, 0) + C(0, Q_Y).$$

The left-hand side is the cost of a two-product firm supplying both products X and Y. The right-hand side is the total cost of having $Q_X$ produced by one firm and $Q_Y$ produced by another firm.

In addition to assuming that there are economies of scope, we will assume that there are *product-specific economies of scale*. For the single-product case, economies of scale mean that average cost is declining. Product-specific economies of scale are a similar concept but defined instead for the multiproduct case. The incremental cost of producing $Q_X$ is defined to be the added cost from producing $Q_X$, given that $Q_Y$ of product Y is already being produced:

$$IC(Q_X) = C(Q_X, Q_Y) - C(0, Q_Y).$$

Average incremental cost of X is then

$$AIC(Q_X) = \frac{C(Q_X, Q_Y) - C(0, Q_Y)}{Q_X}.$$

If product X has product-specific economies of scale, then the average incremental cost of producing X is declining. We assume that the regulated monopolist has product-specific economies of scale with respect to both products X and Y.

We now want to argue that if a multiproduct cost function has both economies of scope and product-specific economies of scale, then it is a multiproduct natural monopoly. That means the cost of producing $Q_X$ and $Q_Y$ is minimized by having a single firm in the industry. Suppose that there are $m$ firms producing product X and $m$ firms producing product Y. As there are economies of scope, total industry cost is reduced by having these $2m$ firms combine into $m$ two-product firms. That is, the average cost of a single firm producing $X$, $AC(Q_X)$, exceeds the average incremental cost of a two-product firm producing $Q_X$, $AIC(Q_X)$ (see figure 15.5(a)). Given product-specific economies of scale, average cost is lower, the more a firm produces. Thus total cost is reduced by having the $m$ two-product firms combine into a single firm. This conclusion means, of course, that the industry is a natural monopoly.
Let us return now to considering cross-subsidization. Suppose that the demand curves for $X$ and $Y$ are as shown in figure 15.5. One cross-subsidization scheme entails pricing product $X$ at $P^o_X$ and product $Y$ at $P^o_Y$ so that the regulated firm earns normal profit. We have assumed that the price of product $Y$ is less than its average incremental cost (figure 15.5(b)). In order for losses from the sale of $Y$ to be covered, the price of product $X$ must then be relatively high (figure 15.5(a)).

What we want to show now is that there is room for wasteful entry under partial deregulation even though the monopolist is earning normal profit and it is a natural monopoly. Depicted in figure 15.5a is the average cost curve of a single-product firm. A new firm can profitably enter by pricing slightly less than $P^o_X$. Note that entry will not take place into market $Y$ because it is unprofitable. This practice of entering only the more profitable markets is known as cream skimming.\(^2\) Entrants are taking the “cream” from the established firm and leaving the “milk,” that is, the less profitable markets.

There are two important facts to note about cream skimming. First, it entails wasteful entry, because single-firm production is cost-minimizing at every output pair. As shown in figure

\(^2\) Optimal regulatory policy to prevent creamskimming by less efficient firms is analyzed in William Brock and David Evans, “Creamskimming,” in David Evans, ed., Breaking Up Bell (New York: North Holland, 1983).
15.5(a), the average cost of a single-product firm exceeds the average (incremental) cost of the regulated two-product firm for product \( X \). The second remark is that cross-subsidization is a socially inefficient way to price. By setting price too high for product \( X \), the wrong signals are being sent to potential entrants. If the regulators instead pursued Ramsey pricing (see chapter 11), entry would not be profitable for less efficient firms.\(^3\)

**Intercity Telecommunications Market**

Events in the intercity telecommunications market (ITM) provide a unique opportunity to apply the analysis presented in the preceding section.\(^4\) The service provided in this market is the sending of a message to a particular place at a particular time. This transference occurs between geographic regions rather than on a local or intracity basis. There are basically three different types of services offered in the ITM. They are message-toll service (MTS), which is what we usually think of as long-distance telephone service, wide-area telephone service (WATS), and private-line service (PLS). The last of these is a circuit that connects two or more points to meet the communication needs of specific users for full-time access. For example, a PLS may be used by a manufacturer to provide point-to-point communications between two factories. The users of PLS are typically medium- or large-sized firms or government organizations.

**Regulatory Background**

Federal regulation of the ITM finds its roots in the Mann-Elkins Act of 1910. This piece of legislation gave the Interstate Commerce Commission (ICC) the power to regulate interstate telephone service. Its ability to control entry resulted in American Telephone & Telegraph (AT&T) achieving a de jure monopoly in long-distance voice transmission. The ICC also had the power to set maximum and minimum rates for services. The Communications Act of 1934 transferred power over the ITM to the newly created Federal Communications Commission (FCC). The FCC had control over most aspects of competition through its control of price, entry, and interconnection. Interconnection is the linking of long-distance lines with local telephone lines. Until the late 1950s, the ITM was a classic case of a regulated monopolist. AT&T was the sole supplier of MTS, WATS, and PLS. The FCC controlled price and prevented any other firm from competing with AT&T.

The basis for regulation of the ITM was that it was considered to be a natural monopoly. For many years the best available production technology was the open-wire line system,

---


which involves stringing wires between poles in order to send messages across geographic regions. Because of the very high fixed costs of such a system and the relatively low marginal cost of adding another customer, economies of scale were believed to exist at the relevant portion of the demand curve.

In the 1930s, AT&T developed coaxial cable, which replaced the open-wire line system as the best available technology. Coaxial cable is able to carry a much greater number of long-distance communications lines simultaneously. However, while a technological improvement, coaxial cable also entails sufficiently large fixed costs, so that a natural monopoly existed even for the largest intercity telecommunication routes. Thus, cost efficiency demanded single-firm production, and that firm was AT&T.

Transformation of a Natural Monopoly

Though microwave transmission existed prior to World War II, it was not commercially viable until certain technological breakthroughs were achieved by the research and development program funded by the U.S. government as part of the war effort. The economic significance of microwave transmission rests in its ability to inexpensively transmit large amounts of information via radio beams. In contrast to open-wire line or coaxial cable, which requires a physical connection between two points, microwave transmission is achieved through a series of microwave relay stations every twenty to thirty miles. Each station receives the microwave signal, amplifies it, and transmits it to the next station. The first microwave radio relay system in the United States for telephone service was installed between Boston and New York in 1947.5

By obviating the need for a physical connection between two points in a communications network, microwave radio technology greatly reduced the fixed cost of providing telecommunication services. As we know from our earlier analysis, a reduction in the fixed-cost component of the cost function results in a smaller efficient firm size. This fact was evidenced in the 1950s, as many private firms and government organizations petitioned the FCC to allow them to build their own private-line systems. On the cost side, we can conclude that the advent of microwave transmission reduced minimum efficient firm size, and thus the ITM was more amenable to having several suppliers.

At the same time, the demand for telecommunication services was shifting out for several reasons. First, income had been trending upward since the late 1940s. From 1949 to 1984, real per capita disposable personal income rose at an annual rate of almost 2.2 percent.6 In his review of the empirical literature, Lester Taylor finds that the estimates of long-run income elasticity of demand for long-distance telephone service range from 0.038 up to 2.76, which

clearly indicates it is a normal good. Given this sensitivity of the demand for long-distance telephone service to income, the upward trend in personal income caused the demand curve in the ITM to rise over time.

A second factor affecting market demand was that the advent of microwave technology altered some of the characteristics of the product itself. This new technology allows communications to be used for a wider range of purposes, which means that market demand would be expected to shift up. An FCC report stated some of the duties that a microwave system could perform for a manufacturer:

Thus the central station in a microwave system can start, stop, slow or speed unattended equipment; open and close valves; record pressure, temperature, engine speed, rate of processing and other data; telemeter voltage, current and power; locate line faults, and perform other supervisory functions. In contrast, the open-wire line system was quite limited in its ability to perform tasks outside of standard telephone service.

Finally, the demand for intercity telecommunications exogenously increased with the development of computers in the 1950s and their later widespread use. Computers represent a product that is complementary to telecommunication service, inasmuch as the latter provides data processing and transmission service. All three of the factors described caused the demand curve in the ITM to shift out.

In analyzing both the cost and the demand side, it is clear that events since the late 1940s called into question the natural monopoly status of the ITM. The minimum efficient firm size had fallen and market demand had increased. At issue, however, is whether the size of these changes was sufficient for the industry to be transformed into one in which monopoly regulation is no longer desirable. As can be seen in figure 15.6, there are two possible scenarios. Under one scenario, the average cost curve shifts from $AC(Q)$ to $\bar{AC}(Q)$, while market demand rises from $D(P)$ to $\bar{D}(P)$. In this case, a natural monopoly still exists, as the cost function is subadditive at $Q^*$. Under the other scenario, the quantitative effects are greater, so that the average cost curve shifts from $AC(Q)$ to $\hat{AC}(Q)$ and the demand curve shifts from $D(P)$ to $\hat{D}$. At the socially optimal price of $\hat{P}$, the market can support four firms at the efficient size. It is clear from the evidence that changes have occurred in the direction shown. The difficulty lies in determining which scenario better describes what has transpired in the ITM.

The empirical evidence on this issue is mixed. As a starting point, consider the cost studies done by AT&T and Motorola in 1962 for an FCC hearing on the Telpak tariff proposed by AT&T. The estimated average cost curves for a 200-mile microwave system are depicted in


FCC Annual Report (June 30, 1956), p. 34.

figure 15.7. There are clearly economies of scale, though the average cost curve seems to be flattening out around 240 circuits. If economies of scale are exhausted at around that size, then a natural monopoly would have existed for many telecommunication routes in the late 1940s. However, with the increase in demand since the 1950s, only the very low-density routes would still be a natural monopoly at 240 circuits. The larger intercity routes have demand of several thousand circuits, suggesting that many efficient-sized firms could have profitably existed. It should be noted that these conclusions are restricted by the fact that they are true only for 1962 and that our conjecture as to what the average cost curve looks like after 240 circuits is correct.

In his 1975 study, Leonard Waverman found for the mid-1960s that scale economies were exhausted at around 1,000 to 1,200 circuits. In the late 1960s the New York–Philadelphia route had demand of around 79,000 circuits. In light of his estimate of minimum efficient scale of 1,000 to 1,200 circuits, he concluded that several suppliers could have existed on many intercity routes. Econometric work by David Evans and James Heckman in 1982 also

10. Waverman, “Regulation.”
supports this conclusion. They estimated a multiproduct cost function for the Bell System based on data during the period 1958–1977. Their empirical estimates show that the cost function was not subadditive at any output configuration for AT&T over that period.

In contrast, other studies have found quite significant economies of scale for AT&T over the period 1947–1976. The estimates of (Average cost/Marginal cost) ranged from 1.58 up to 2.12. If average cost exceeds marginal cost, then average cost must be falling, implying


that there are economies of scale. It is important to note that these studies considered only product-specific economies of scale, and not economies (or diseconomies) of scope. In contrast, the Evans and Heckman study allowed for both. It should be noted, however, that Evans and Heckman comment on the poor quality of data available for performing this type of analysis.

**Regulatory Policy in the Microwave Era**

With the advent of microwave technology, many private firms and government organizations began petitioning the FCC in the early 1950s to allow them to build and operate their own point-to-point communication networks. These demands led to the Above 890 Mc decision in 1959. In this decision, the FCC stated that frequencies above 890 megacycles would be shared by AT&T (the common carrier) with private users. What was significant in this decision is that a system built by a noncommon carrier could be used only for private demand; a firm could not sell telecommunication services. In response to the Above 890 Mc decision, AT&T entered the Telpak tariff, which called for volume discounts on PLS. Presumably, AT&T’s objective was to make it profitable for businesses to buy AT&T’s services rather than build their own system. Ultimately, the FCC disallowed the Telpak tariff because it was not justified by the cost estimates.

Microwave Communications Incorporated (MCI) petitioned the FCC in 1963 to allow it to enter the St. Louis–Chicago PLS market as a common carrier. MCI desired to supply PLSs and act as a competitor to AT&T in this market. After six years of hearings and $10 million of expenses incurred by MCI, its application was approved in 1969. Once the MCI decision was made, the FCC found itself inundated with requests by other firms desiring to enter the PLS market. Furthermore, MCI, as well as the other new carriers, had to petition to enter each individual route. The MCI decision in 1969 only gave MCI permission to service the St. Louis–Chicago route. In response to this demand, the FCC made the Specialized Common Carrier (SCC) decision in 1971, which allowed free entry into the PLS market. A firm needed only to submit an application, and lengthy regulatory hearings were avoided.

Entry was extended to the message toll service (MTS) market with MCI’s introduction of Execunet service in 1975. The MTS market represented considerably larger revenues than...
the PLS market. Later that year the FCC ruled in Execunet I that the SCC decision opened up entry only into the PLS market, so that MCI must discontinue their Execunet service. With the U.S. court of appeals overruling this decision in 1978, entry was extended from the PLS market to the entire ITM.

Starting in 1969 with the MCI decision, the FCC pursued a policy of partial deregulation. While allowing entry into the PLS market and then, by court order, into the MTS market, the FCC continued to regulate rates. In particular, they maintained the practice of cross-subsidization. This occurred at two levels. First, there was the use of long-distance rates to subsidize local service rates charged by AT&T. Within the ITM, AT&T was also forced to use rates on high-density, long-distance routes to subsidize rates on low-density routes. This subsidy specifically took the form of charging the same rate independent of the amount of traffic, even though average cost is lower, the greater the amount of traffic. A major reason for continuing price regulation was the fear that AT&T would set artificially low rates in order to drive competitors out of the ITM. Presumably, below-cost rates would be funded by AT&T pricing above cost in its monopoly markets in local telephone service.

Given these facts, it is not surprising that entry initially took place in the high-density markets. Because demand was largest there, these markets could best support several firms. Furthermore, because AT&T’s prices were above cost, there was room for profitable entry. It is also not surprising that AT&T complained that entrants like MCI were less efficient and found entry profitable only because of cream skimming. The allegation of being less efficient is very difficult to determine empirically. Nevertheless, because of cross-subsidization, the FCC certainly opened up the possibility that less efficient firms could find entry into the high-density markets profitable. Evidence against AT&T’s claim, however, is the fact that MCI eventually entered all of the MTS markets.

As a result of a seven-year antitrust case against AT&T by the U.S. Department of Justice (DOJ), AT&T agreed on January 8, 1982, to sever its connections with its twenty-two telephone operating companies. These twenty-two companies were made into seven holding companies, the regional Bell operating companies (RBOCs). Local Bell System operations were subdivided into 161 “local exchange and transport areas” (LATAs), with each LATA being assigned to one of the RBOCs. The key restriction on RBOCs is that they were not allowed to provide interLATA services. In exchange for spinning off its telephone operating companies, AT&T was permitted to retain Western Electric (its manufacturing division), Bell Labs (its research and development division), and Long Lines, which supplies intercity telecommunication services. In addition, the 1956 consent decree, which prevented AT&T from entering any unregulated markets, would be erased. The result of the breakup, which took place on January 1, 1984, is that AT&T is no longer involved in any monopoly markets.

13. As a matter of fact, in response to the SCC decision, AT&T tried to reduce cross-subsidization practices with their HiLo tariff in 1973. This tariff proposed lower rates on the high-density routes. The FCC did not approve it.
Regulated Monopoly to Regulated Competition

Though the breakup of the Bell System resulted in AT&T operating exclusively in markets with competitors, regulation persisted. AT&T was required to serve all customers, file tariffs whenever it offered a new service, and average rates across broad customer segments. By comparison, competitors like MCI and Sprint were left unconstrained in their pricing, provision of new services, and movement into and out of markets. To add to this unequal treatment, almost every new tariff by AT&T was challenged by its competitors, with the typical basis being that it was not cost-justified and thereby was predatory in intent. The exact method by which AT&T’s rates were regulated evolved considerably over the postbreakup period. In the mid-1980s, the FCC deployed standard rate-of-return regulation that allowed AT&T to set rates commensurate with a return of 12.2 percent. Then, in March 1989, the FCC approved the use of price caps.

What was the rationale for this asymmetric regulatory policy? In one word, “dominance.” The FCC feared that an unconstrained AT&T could use its dominant position to engage in anticompetitive practices. Though there was quick and growing encroachment on AT&T’s position in the market both before and after the breakup, it is clear from figure 15.8 that AT&T still held a lion’s share. During most of the decade after the breakup, the industry was characterized by one dominant firm in the form of AT&T, two reasonably sized competitors with MCI and Sprint, and hundreds of resellers who, lacking their own physical network, leased lines from the three largest long-distance companies. By most measures, this was still a highly concentrated industry.

Given such a market structure, two hypothetical scenarios raised doubts about fully deregulating AT&T. The first scenario was that an unregulated AT&T would substantially raise its rates. Whether it would be profitable for it to do so, as well as the impact it would have on the market, depended on the ease with which customers could switch to other providers and on the ability of competing providers to meet a large increase in demand. If customers found that the services of MCI and Sprint and other suppliers were comparable to those of AT&T and, furthermore, that those providers had adequate excess capacity, then a sharp rise

16. Vietor, Contrived Competition.
17. For a discussion of these price caps, see chapter 12.
18. As an aside, the courts eventually required all carriers to file tariffs.
in AT&T’s rates would have proved unprofitable, as it would have induced many of AT&T’s customers to shift to other suppliers.

In retrospect, the evidence supports that conjectured outcome. Using data for 1984–1993, the price elasticity of AT&T’s demand curve was estimated to be around $-4.2$. This estimate means that its demand would be expected to fall by 4 percent for each 1 percent rise in its price. With such a highly elastic demand curve, a substantial price rise would prove highly unprofitable. This empirical finding seems plausible, for two reasons. First, at least for the residential market, the long-distance telephone service provided by AT&T and its competitors appears quite homogeneous, and therefore consumers would be inclined to buy from the lowest-priced firms. Second, the development of fiber-optic systems created considerable excess capacity, so that competitors would be able to handle a large increase in the demand for their services.

The second scenario that argued against full deregulation was that AT&T would respond by engaging in a predatory policy of pricing below cost. As the story goes, this policy would have imposed losses on competitors and, upon their exit from the industry, allowed AT&T to raise its rates. This course of events seems unlikely as well. The large market share of AT&T would have implied sizable losses if they had pursued a below-cost pricing strategy. The size of these losses would have been further enhanced by the required length of the price war in light of the size and financial strength of MCI and Sprint. Also, given that the cost of their long-distance networks, whether microwave or fiber-optics, is sunk, the marginal cost of continued operation was relatively low. It would have then taken a price war of biblical proportions for AT&T to drive out MCI and Sprint. In sum, a retrospective analysis suggests that continued regulation of AT&T after the breakup of the Bell System was probably unwarranted.

Regulated Competition to Unregulated Competition

With increasing evidence that these two scenarios were unlikely to materialize in an unregulated environment, the 1990s witnessed a gradual elimination of FCC control of AT&T’s rates. Price caps were removed from business services in October 1991 and then from 800 services in May 1993. AT&T was judged to be nondominant in domestic residential services as of November 1995. Effectively, the domestic ITM is now fully deregulated, with one disappearing caveat.

The 1982 consent decree that broke up the Bell System prevented an RBOC from entering the ITM. This decree has since been obviated by the Telecommunications Act of 1996, which will be discussed in greater detail later. With this legislative act, an RBOC is permitted to offer long-distance service to its local service customers if an adequate degree of competition has been achieved in the local service market. Though the FCC initially turned down most RBOC applications, many applications have been approved since 2000. Indeed, as of 2002, RBOCs had 15.8 percent of the residential long-distance market, with Verizon’s market share already exceeding that of Sprint.

AT&T has continued to experience deterioration in their dominant position, as shown in figure 15.8. Whereas AT&T’s market share was more than triple that of MCI in 1994, it was only about 50 percent higher just seven years later. This reflects AT&T’s market share having fallen to 37 percent but also the merger of MCI and WorldCom in 1998. Though the industry is far from its former status of having a dominant firm and a collection of small firms, by most standards it was still highly concentrated at the end of the 1990s, which makes it surprising that the DOJ permitted the second and fourth largest firms to merge to form MCI

21. An RBOC can, however, offer intercity telecommunication services to customers outside of its region.
WorldCom. Regardless, deconcentration continues to occur, and the entry of the RBOCs has further added to this trend.

In evaluating the competitiveness of the long-distance industry in the last two decades, there is no denying that long-distance rates have declined considerably from an average rate of 55 cents per minute in 1984 to 10 cents per minute in 2001 (in 2001 dollars).\(^{23}\) This fact is undisputed. What is contentious is how this decline is to be attributed to three different sources: intensified price competition, enhanced efficiency, and the regulatory shifting of costs from long distance to local telephone suppliers.\(^{24}\) After the breakup, the FCC introduced a fixed monthly “subscriber line charge” for local telephone service. This served to reduce long-distance rates by lowering the access charges paid by long-distance telephone suppliers to the local exchange operator for connecting to their network. The access charge has steadily declined from almost 30 cents per minute in 1984 to 1.7 cents per minute in 2001 (in 2001 dollars).\(^{25}\) Clearly, a big part of the decline in long-distance rates is the lower cost of accessing the local network.

So how much competition was there among long-distance suppliers during the 1990s?\(^{26}\) Some economists characterized the industry as collusive, with AT&T being the price leader and the other firms following with slightly lower rates (but not so low that AT&T would fail to tolerate them).\(^{27}\) Other economists described the market as competitive, as reflected in the steady erosion of AT&T’s market share, the introduction of discount programs like MCI’s “Family and Friends,” and heavy advertising expenditures. But that is now ancient history, for, as this chapter is being written, the ITM is in the midst of fierce competition by virtue of large investments in fiber-optic networks, which expanded capacity far in excess of demand.

### Telecommunications Act of 1996

Historically, cable television, local telephone, and long-distance telephone services have been distinct markets. The future of the industry involves these distinctions melting away as we witness the convergence of communications, computers, and television. Firms are beginning


\(^{24}\) The regulatory practices associated with the subsidization of local telephone rates by long distance revenues are discussed in chapter 12.


\(^{26}\) For a review of the evidence, see David L. Kaserman and John W. Mayo, “Competition in the Long-Distance Market,” in Handbook of Telecommunication Economics, vol. 1, Structure, Regulation and Competition, Martin E. Cave, Sumit K. Majumdar, and Ingo Vogelsang, eds. (Amsterdam: Elsevier, 2002).

\(^{27}\) See, for example, Peter W. Huber, Regulation, 1993, No. 2; and Paul W. MacAvoy, “Tacit Collusion under Regulation in the Pricing of Interstate Long-Distance Telephone Services,” Journal of Economics and Management Strategy 4 (Summer 1995): 147–85.
to integrate these various services, taking a direction quite contrary to that laid out with the breakup of the Bell System in 1984. While the source of this convergence is technological, the speed at which it is achieved will be influenced by regulatory policy.

The primary technological developments are fiber-optics and digital electronics. Fiber-optic technology provides tremendous capacity and reliability compared with coaxial cable and microwave relay systems. Joined with digital electronic technology, the resulting network has the ability to transmit vast amounts of data and video at high speeds. Voice messages now move in digital form over the network, so that the same physical network can move data, video, and voice. Augmenting these advances are innovations in signal compression that allow more information to be transported. A third important factor in the future development of this industry is the advent and advancement of wireless communications.

By the 1990s, innovations had increasingly made the existing regulatory structure confusing and out of sync. A major overhaul of telecommunications legislation had not taken place since the Telecommunications Act of 1934 created the FCC. More recently, the Modification of Final Judgment, which resulted in the separation of local and long-distance telephone service, was written in 1982. With the pace of innovation, that decision had become prehistoric. Although cable legislation was passed as recently as 1992, it had always been done in isolation from what was happening in other telecommunications markets.

This fragmented regulatory policy increasingly conflicted with technological trends. The culmination of these forces was the Telecommunications Act of 1996. Enacted by Congress on February 1, 1996, and signed by President Bill Clinton a week later, it went into effect immediately. It has radically altered the regulatory landscape.

The 1996 act preempts all state laws that limit competition in the markets for local and long-distance telephone services. It obviates the 1982 consent decrees prohibiting the RBOCs and GTE from supplying long-distance telephone service and repeals key features of the Cable Act of 1992. It requires the RBOCs to provide equal access to their systems by interexchange carriers (that is, long-distance telephone systems) and permits them to offer long-distance telephone service to their local customers upon FCC approval. Approval is to occur when there is deemed to be adequate competition in local telephony. RBOCs may immediately offer long-distance telephone service outside of the region for which they offer local telephone service. While a local telephone company may own cable systems outside its region, it is limited to a 10 percent ownership to a cable provider in its market.

The 1996 act not only significantly altered the regulations faced by telecommunications firms but also made the FCC the key arbiter of competition. In contrast, the MFJ effectively made regulators out of the federal district court and the DOJ. However, neither body is well-suited for this task. Judges lack the necessary support in terms of administrative personnel and technical expertise with regards to technology and pricing. Though the DOJ has many qualified attorneys and economists, they do not typically have industry expertise (as they, along with the Federal Trade Commission, are responsible for enforcing antitrust laws in all
industries) and do not have the rulemaking capacity that a body needs if it is to effectively regulate an industry.

As it should be, the FCC is now the regulatory body in charge, and the 1996 act assigned it two challenging tasks. First, it must ensure that the local telephone companies provide equal access of other firms to their networks. This means not only requiring that they offer interconnection but also that it be done at an appropriate price. The second task is to implement Section 271, which requires evaluating RBOC applications to offer long-distance service in areas for which they offer local telephone service. Here, the DOJ remains as an advisor, which is more consistent with its usual role of evaluating the competitiveness of an industry. This they do every day when, for example, they investigate proposed mergers.

As stated in its preamble, the purpose of the 1996 act is to promote competition and reduce regulation in order to secure lower prices and higher quality services for American telecommunications consumers and encourage the rapid deployment of new telecommunications technologies.

Much of these potential gains were to be realized by tearing down regulatory entry barriers with the expectation that telecommunication firms would enter into each other’s markets. While RBOCs have indeed entered the long-distance telephone market as fast as the FCC will allow, this has not been the case in many other markets. Though in the mid-1990s the RBOCs appeared ready to upgrade their systems to fiber-optics so as to enter the cable television market and the cable companies appeared to be gearing up to offer local residential telephone service, it failed to materialize. Cable television suppliers continue to have a near monopoly, with the only source of competition being satellite TV. As reviewed in chapter 13, cable rates have gone up significantly since the 1996 act. While wireless competition has become increasingly important in local telephony, the RBOCs continue to dominate the residential market.

---

**Separation of Regulated Monopolies and Competitive Markets**

For more than half a century, AT&T was a regulated monopolist in almost every respect. Its telephone-operating companies monopolized local service for a very large share of the U.S. population. Western Electric, the manufacturing arm of the Bell System, was the sole supplier of terminal equipment for those customers serviced by the Bell operating companies. Its dominance was evidenced by the fact that the Bell System owned approximately 80 percent of the telephones in the United States in 1956. Finally, the Long Lines division had

---

28. Terminal equipment is that equipment used to terminate a telephone line at the final user’s premises. It includes the ordinary telephone.
an undisputed monopoly of intercity telecommunications, which included long-distance telephone service. By the time of the breakup in 1984, however, it was clear that a regulated monopoly no longer properly characterized the Bell System. Instead, AT&T was operating in a mixture of regulated and competitive markets.

The change in the structure of AT&T’s markets began taking place in the late 1960s. In its 1968 Carterphone decision, the FCC decided that the final users of telephone services could purchase their own terminal equipment. Before that time customers were restricted to connecting Western Electric equipment to Bell telephone lines. The Carterphone decision allowed them to buy equipment from alternative suppliers, subject to quality standards set by AT&T. This action opened up the terminal equipment market to competition. Then, in 1969, the FCC began a policy of free entry into the PLS market with the MCI decision and later with its Specialized Common Carrier decision in 1971. Competition was extended to the remainder of the intercity telecommunications market in 1975 with MCI’s Execunet service.

By the late 1970s, AT&T was involved in markets that ranged from regulated monopoly (local telephone service) to regulated competition (long-distance telephone service) to unregulated competition (terminal equipment). The decision to break up AT&T was a statement that such a mixed situation was untenable. AT&T was forced to divest itself of its regulated monopolies, that is, the twenty-two Bell telephone-operating companies. This divestiture left AT&T with Long Lines, Western Electric, and Bell Labs, all of which were operating in markets that were unregulated or being deregulated. In line with AT&T being transformed into an unregulated firm, the Justice Department erased the 1956 consent decree, which prevented AT&T from entering unregulated markets. Thus the breakup of the Bell System had resulted in an AT&T that was stripped of its regulated monopolies in local telephone service and was prepared to compete freely in the remainder of the communications industry.

The objective of this section is to analyze the general issue underlying the decision to break up AT&T. This issue is concerned with whether regulated monopolies should be allowed to compete in unregulated markets or should instead be separated from them. Referred to as the separation issue, it has arisen not only in the telephone industry but also in the cable television and electric power industries. The next few subsections consider the benefits and costs of separation and then apply this analysis to understanding the decision to break up the Bell System.

Benefits and Costs of Separation

The issue before us is whether a firm that is subject to monopoly regulation in one market should be allowed to enter and compete freely in markets that are unregulated and potentially competitive. On the one hand, we will observe that there are certain anticompetitive effects that can arise when a market consists of a regulated monopolist and unregulated firms. On the other hand, separation may result in certain cost inefficiencies.
Policy Options

One policy that is available to handle this type of situation is that of separation, which prohibits the regulated firm from participating in unregulated markets. There are actually two approaches to pursuing a policy of separation. Suppose that market $X$ is the regulated market and market $Y$ is the unregulated market. Separation can be achieved by restricting the regulated firm to participating only in market $X$. This was the approach implicit in the 1956 consent decree. A second method is to make both markets $X$ and $Y$ subject to monopoly regulation. When market $Y$ is potentially competitive, such a policy would generally be undesirable. Thus our attention will be directed to the first approach. An alternative policy option to that of separation is to allow the regulated monopolist to compete with other firms in market $Y$. Market $Y$ could be left totally unregulated, or the regulatory agency could limit the degree of competition through price controls (setting maximum and minimum prices) or some other form of regulation.

Benefits of Separation

We want to consider the benefits of a policy that prohibits a regulated monopolist from competing in unregulated markets. The benefits of such a policy are defined in relation to what would occur if a policy of no separation were pursued. In particular, there are certain anticompetitive practices that might result from allowing a regulated monopolist to compete against unregulated firms. The benefits of separation rest in preventing such practices from taking place.

Suppose that the regulated monopolist was allowed to compete freely in an unregulated market. One possible undesirable effect is that the monopolist would use profits earned from its regulated markets to fund a policy of predatory pricing in the unregulated markets. By setting a very low price, it could impose losses on its competitors and perhaps induce them to exit the industry. The advantage of the regulated firm is having regulated markets as a source of revenue to fund such activities.

A second, and perhaps more serious, anticompetitive practice may arise if the products in the regulated and unregulated markets are related—in particular, if product $X$ (regulated) is an input in the production of product $Y$ (unregulated). The regulated monopolist can then control the supply of competitors through its supply of product $X$. Alternatively, it may choose to give competitors a lower-quality product. Thus, consumers may find the regulated monopolist’s product preferable to that of its competitors. The exact effects of this type of anti-

29. In the case of local telephone service (market $X$) and long-distance telephone service (market $Y$), some economists argued that the optimal policy was to place both under monopoly regulation, even if the ITM was no longer a natural monopoly. The reason is that Ramsey pricing requires that prices be regulated in both industries in order to maximize social welfare.
competitive practice are very much dependent on the relationship between products $X$ and $Y$. A more detailed analysis will be provided later when we look at applications.

Finally, the regulated monopolist may be able to charge a higher price in its regulated market under a policy of no separation. Because there is likely to be some capital that is shared by products $X$ and $Y$, the regulated firm has some discretion in assigning costs. Because it is required to set price in market $X$ so that revenue equals cost, it will find it optimal to assign all joint costs to product $X$. Doing so will increase the price it can set in the regulated market and thus increase profit. This practice is not anticompetitive for market $Y$, but it does lead to a price in the regulated market that is above the socially optimal level.

The benefits of separation rest in preventing these types of practices. Under a policy of separation, we may observe a lower price both in the unregulated market and in the regulated market.

**Costs of Separation**

The primary cost of separation is the elimination of a potential competitor (the regulated monopolist) from the market. If there is free entry into the market and the regulated monopolist offers no advantages over other firms, then there is little lost from preventing it from competing. Suppose, however, that the regulated monopolist does possess an advantage. One plausible source of advantage is that it can produce the unregulated product at lower cost because of economies of scope between the regulated and unregulated products. By prohibiting the regulated monopolist from supplying the unregulated product, the most efficient firm is then eliminated from the market. The value of the increased resources used to supply the unregulated product represents a cost to society from separation.

If the two products are inputs in the production of the same good or service, there may be gains from vertical integration. These can be due to lower transaction costs from coordinating activities, but they can also be due to the reduction of a pricing distortion known as *double marginalization*. In pricing an input internally, a profit-maximizing firm will price it at marginal cost so as to result in an efficient input mix and thereby cost minimization. However, unless regulation forces price to marginal cost, a firm will price the input above cost when selling it to other suppliers in order to reap profit. However, this causes other firms to have an inefficient input mix, which means wasted resources from society’s perspective. A policy of separation would add to the welfare losses from double marginalization.\(^{30}\)

In summary, the main benefits from a policy of separation are the prevention of anticompetitive practices in the unregulated market and shifting of costs from the unregulated to the regulated market. This should result in a lower price than would be the case without

\(^{30}\) For details on the welfare effects of vertical integration, including double marginalization, the reader is referred to chapter 8.
separation. The potential cost to such a policy is the wasted resources due to preventing the most efficient firm from competing. Such inefficiencies would tend to raise price. The social optimality of separation then depends on the ability of the regulated monopolist to pursue anticompetitive practices in an unregulated market and the degree of economies of scope that may exist between the regulated and unregulated products.

Breakup of AT&T

On November 20, 1974, the DOJ filed an antitrust suit against AT&T for violation of Section 2 of the Sherman Act. AT&T was accused of attempting to monopolize the telecommunications industry by using its dominant position in three segments of that industry: local exchange, long distance, and terminal equipment. The case went to court on January 15, 1981, and on January 7, 1982, two weeks before AT&T was scheduled to complete its case, AT&T and the DOJ reached a settlement. As described earlier, AT&T agreed to divest itself of its local telephone operating companies in exchange for repeal of the 1956 consent decree and retention of Western Electric, Bell Labs, and Long Lines. Judge Harold Greene, presiding over the trial, approved the settlement on August 11, 1982, and, on August 24, 1982, entered what is now known as the Modification of Final Judgment. The divestiture was consummated on January 1, 1984.

The 1982 consent decree that broke up AT&T was a decision that a policy of separation was appropriate. At the time of the breakup, AT&T had a regulated monopoly in local telephone service but was competing with other firms in the partially deregulated ITM. By requiring AT&T to divest itself of its twenty-two telephone operating companies in exchange for repeal of the 1956 consent decree and retention of Western Electric, Bell Labs, and Long Lines. Judge Harold Greene, presiding over the trial, approved the settlement on August 11, 1982, and, on August 24, 1982, entered what is now known as the Modification of Final Judgment. The divestiture was consummated on January 1, 1984.

The 1982 consent decree that broke up AT&T was a decision that a policy of separation was appropriate. At the time of the breakup, AT&T had a regulated monopoly in local telephone service but was competing with other firms in the partially deregulated ITM. By requiring AT&T to divest itself of its twenty-two telephone operating companies in exchange for repeal of the 1956 consent decree and retention of Western Electric, Bell Labs, and Long Lines. Judge Harold Greene, presiding over the trial, approved the settlement on August 11, 1982, and, on August 24, 1982, entered what is now known as the Modification of Final Judgment. The divestiture was consummated on January 1, 1984.

As mentioned previously, the major benefit of separation is to make the unregulated market better suited for equal and unconstrained competition among firms. Because price regulation was still in effect in the long-distance market, separation was not required to prevent predatory pricing. It is interesting to note, however, that on several occasions the FCC prevented AT&T from drastically reducing rates in the ITM. In the Telpak decision in 1964, the FCC disallowed a tariff offering by AT&T that reduced rates on private line services. The reason was that it was not cost-justified because the estimated rate of return would have been under 1 percent. It would appear that the FCC was concerned that the tariff was predatory. Similarly, the FCC disallowed the HiLo tariff proposed by AT&T, which sought lower rates on high-density routes being entered by MCI and other specialized common carriers. On the St. Louis–Chicago route, MCI was charging a full-time rate per channel of $481.65 at the time of the tariff in 1974. AT&T proposed to charge $341.85. However, in fairness to AT&T, it was probably not so much interested in pricing predatorily as in simply trying to eliminate cross-subsidization so to be able to compete.

The real problem arose in that the two products were interrelated. To provide long-distance service, a competitor like MCI had to interconnect with the local service lines provided by
AT&T. The initial response of AT&T to entry in 1969 by MCI was simply to refuse to interconnect with them. In the SCC decision in 1971, the FCC said AT&T should interconnect with their competitors, but the terms were left open to AT&T. This decision did not improve the situation, because AT&T placed considerable restrictions on the specialized common carriers. Only in 1974 did the FCC order interconnection in its Bell System Tariff Offering decision. When MCI expanded entry into message toll service, the same problem arose. Their entry was approved by a U.S. court of appeals in 1975, but not until 1978 was AT&T was forced to interconnect with MCI’s Execunet service.

Only in 1978 were firms like MCI allowed to interconnect with the local operating company as Long Lines. Even after achieving this right, AT&T’s competitors in the ITM were still not treated equally. For example, customers had to dial twenty digits to make a long-distance call with MCI, but only eleven with AT&T. If consumers saw AT&T as offering a higher-quality product, this would require competitors to offer a discount to compete. It was this type of behavior that led to the original antitrust suit against AT&T.

Under the current separation policy, the local operating companies no longer have an incentive to offer discriminating service, as they have been divested from AT&T. Under current FCC policy, the ITM has moved to a position where AT&T, MCI, Sprint, and other common carriers are on an equal basis from which to compete. This is clearly the major benefit from separation.

AT&T did point out that there are also costs from a policy of separation through forgone economies of scope. It was believed that there were cost savings from having one firm supply both local service and intercity service. Although there is certainly a need for coordination, it is unclear that it cannot be effectively achieved through the market rather than within the firm. Furthermore, the cost study by Evans and Heckman in 1983 did not find any such economies of scope. Thus there is little evidence to suggest that separation incurred much of a reduction in cost efficiency.

Summary

The moral of this chapter is that there is more to regulating a natural monopoly than simply setting price and preventing entry. The purpose of regulation is to raise social welfare relative to what it would have been in the absence of regulation. With the decision to regulate comes many responsibilities, two of which we have examined in detail in this chapter. One task of a regulatory agency is to decide which markets a regulated firm can serve. On the one

31. For example, AT&T only provided interconnection with private-line service in which one phone was connected to just one other phone. Thus, more sophisticated private-line services were not able to interconnect.

32. In Execunet II (1975), the FCC stated that AT&T did not have to interconnect with MCI’s Execunet service. This decision was overturned, like Execunet I, by a U.S. court of appeals in 1978.
hand, when some of those markets are competitive, there may be anticompetitive effects from allowing a regulated firm to enter. On the other hand, a regulated firm might be able to provide the product or service at a lower cost than other firms because of economies of scope. A regulatory agency must consider these benefits and costs in deciding this issue.

Perhaps the most important responsibility that society subsumes when it regulates an industry is that it must know when such regulation is no longer necessary. In our technologically progressive world, there should be no presumption that an industry that is a natural monopoly today will be a natural monopoly tomorrow. Just as much as regulating a natural monopoly can be welfare-improving, regulating an industry that is no longer a natural monopoly can be welfare-reducing. Although it is ultimately the responsibility of legislators to decide when regulation is no longer appropriate, the first line of change in regulatory policy rests with the regulatory agency. It can choose to allow entry and to loosen controls on price. Ideally, the regulatory agency should be society’s agent in representing our best interests with respect to the industry it regulates.

Unfortunately, there are obstacles inherent in the bureaucratic structure of a regulatory agency that can impede deregulation even when it is required. Historically, regulatory agencies appear to be resistant to major changes in the industries they control. Change requires bureaucratic resources and brings forth political risks if the change happens to result in higher prices or an unhealthy industry. Perhaps a more significant impediment is that deregulation means a curtailment of the duties of a regulatory agency and perhaps even its ultimate demise. This means reduced power, prestige, and income for the regulators. Though deregulation may be optimal from society’s perspective, it may not be optimal from a regulator’s perspective. Although the FCC, along with MCI, was an important force in opening the intercity telecommunications market to entry, it is also clear that the FCC delayed entry and sought to reduce the extent of entry.

There are two basic lessons to be learned from the past chapters on the regulation of natural monopoly. The first is that regulating an industry is a difficult task. Even in a static setting, a regulatory agency must attempt to set the socially optimal price in spite of having very limited information about cost and demand conditions. The problems become even more difficult when the environment changes in significant ways over time. A regulatory agency must make decisions about which markets a regulated firm should be allowed to serve and whether regulatory controls should be loosened in response to changes in cost and demand conditions. The second lesson is that although regulation has the potential to raise welfare, there are many side effects to regulation that are welfare-reducing. When considering a policy of regulation, even of a natural monopoly, one must evaluate these potential side effects. All this is not to say that regulation is a useless policy tool, but rather only to point out that we must be cautious in the use of such a blunt and powerful instrument as the economic regulation of an industry.
Questions and Problems

1. Suppose that the firm cost function is $C(Q) = 100 + 10Q + Q^2$. This cost function generates a U-shaped average cost curve with minimum efficient scale of 10. Determine whether this industry is a natural monopoly when the market demand function is
   a. $D(P) = 100 - 3P$
   b. $D(P) = 90 - 3P$
   c. $D(P) = 100 - 2P$

2. For the past twenty years, the intercity telecommunications market has been open to entry while the FCC has continued to regulate prices. Should the FCC fully deregulate this market? Should the FCC have fully deregulated it twenty years ago? What are the relevant issues in determining the appropriate regulatory policy?

3. When MCI originally entered the intercity telecommunications market, AT&T argued that MCI was cream skimming. What is cream skimming? How would one go about assessing the validity of AT&T’s claim?

4. Why did the FCC allow entry into the private-line service market but prevent entry into the message toll service market?

5. Is there a multiproduct natural monopoly with respect to products $X$ and $Y$ if the firm cost function is $C(Q_X, Q_Y) = 100 + 20Q_X + 10Q_Y - Q_X Q_Y$? Assume that $Q_X \leq 10$ and $Q_Y \leq 10$.

6. Do you think there are economies of scope between local telephone service and long-distance service? How about between local telephone service and cable television service? What difference does the existence of economies of scope make for the optimal regulatory policy?

7. Should there be a general policy that regulated monopolies cannot provide products or services in unregulated markets? What are the benefits and costs of such a policy?

8. The Telecommunications Act of 1996 prohibits a regional Bell operating company from offering long-distance telephone service to its local telephone customers until the local telephone market is deemed competitive by the FCC.
   a. How would you propose determining the degree of competition in the local telephone market?
   b. What is the rationale for these entry restrictions?
   c. Though limiting entry by RBOCs into the long-distance market, the 1996 act allows a long-distance telephone supplier to provide local telephone service; for example, AT&T is the largest supplier of cellular telephone service, as well as the largest supplier of long-distance service. Is AT&T’s size in these markets contradictory to the restrictions placed on RBOCs? Why or why not?

9. Does the local cable television company have a monopoly? Should its rates be regulated?

10. Should the market for local telephone services be deregulated? In answering this question, you will first need to define what society cares about. Also, think about the different types of customers for local telephone service and the active or potential firms that provide local telephone service.
The Regulation of Potentially Competitive Markets: Theory and Estimation Methods

When a market is a natural monopoly, government intervention is warranted on the basis that competition may not work very well. Because a natural monopoly is characterized by declining average cost, production efficiency requires that there be a single firm producing in the industry. However, in order to achieve allocative efficiency, typically there must be several active firms competing so as to drive price down toward marginal cost. Due to this tension between productive efficiency and allocative efficiency, a natural monopoly is unlikely to attain a socially desirable outcome in the absence of government intervention. While far from an ideal solution, government regulation can be a preferable alternative to unfettered competition.

Though natural monopoly is one of the few convincing arguments for the economic regulation of an industry, it is nevertheless true that economic history is full of episodes in which potentially competitive markets have been subject to vast forms of regulation. This and the following chapters in part II are concerned with analyzing the regulation of potentially competitive markets.

Our analysis will address two important issues. First, if there is no market failure, why then is there regulation? Because unregulated competition is thought to be the most effective way in which to achieve a social welfare optimum, the regulation of such markets suggests that it may be in place for private, and not social, gain. An issue with greater public policy relevance is to assess the effects of regulating potentially competitive markets. Our interest lies in understanding how and to what degree regulation impacts price, service, market structure, productivity, and other relevant economic variables. Of concern is the effect of regulation not only on static welfare but also on dynamic efficiency. While the latter is typically harder to predict and quantify, a plausible argument can be made that dynamic welfare losses from regulation greatly exceed their static counterpart.

This chapter provides an introductory theoretical analysis of the implications of price and entry/exit regulation for firm behavior and social welfare. Although this theory is relevant to most forms of economic regulation, it is of particular relevance to the regulation of the transportation industry. As an application of this theory, we analyze the regulation of the railroad, trucking, and airline industries in chapter 17. In chapter 18 the regulation of crude oil and natural gas is investigated and further theoretical analysis is provided that is of particular relevance in understanding the welfare implications of regulation in those industries. In addition to providing an introduction to the theory of economic regulation, this chapter reviews and applies the different methods for estimating the quantitative effects of regulation.

1. As we discussed in chapter 13, active competition can sometimes be replaced with potential competition through the use of franchise bidding.

2. It should be noted that there is also a role for regulation in responding to other types of market failures like externalities.
The major task before us is to understand how price regulation, along with entry/exit regulation, can directly and indirectly affect the decisions of firms and thereby influence social welfare. Because price regulation is common to most forms of economic regulation, this analysis should be applicable to most regulated industries. The particular form of price regulation that we consider is the specification of the price at which firms must sell their product or service. Modeling regulation in this manner is clearly an abstraction, inasmuch as price regulation can take the form of a regulatory agency setting a maximum and/or minimum price that can be charged. Assuming that instead the agency sets a specific price considerably simplifies the analysis and is often an adequate approximation for actual regulatory practices. As is true of most regulated industries, it is assumed that firms are required to meet all demand at the prices set by the regulatory agency. When price exceeds marginal cost, it is clearly optimal for firms to do so.

Two rather general cases of price regulation will be considered. One case is when price is set above cost and entry is prohibited. The second case is when price is set below cost and exit is prohibited. In the transportation industry, the latter is characteristic of past regulatory policies with respect to railroads, whereas the former is characteristic of past regulatory policies with respect to the trucking and airline industries.

It is important to keep in mind that the objective of this section is to offer some initial insight into how price and entry/exit regulation affect firm behavior and to provide some theoretical foundation for investigating the regulation of the transportation industry. This section does not attempt to cover all the effects of regulation. Price and entry/exit regulation are indeed common to most regulated industries, but each industry also has its own idiosyncratic set of rules. These rules often depend on the particular product being regulated, the history of the industry, the ideologies of the regulatory agency’s commissioners, and other industry-specific factors. It is important that these industry-specific rules be considered in assessing the welfare effects of regulation, as they can have a significant influence on firm behavior. Although the analysis of this section will not be concerned with the effects of idiosyncratic rules of regulatory agencies, our case studies of the transportation and energy industries will be.

**Direct Effects of Price and Entry/Exit Regulation: The Competitive Model**

The welfare effects of regulation are derived by comparing the industry equilibrium under regulation with the equilibrium that would have occurred in an unregulated environment. Performing this task requires making conjectures about the properties of the industry equilibrium in the absence of regulation. It is often natural to suppose that an unregulated industry would achieve a competitive equilibrium. In order to derive some clear and concise results,
this is the assumption we initially make. Although this assumption may be appropriate for some markets, in others it is not. Industries in which the minimum efficient size of a firm is not small relative to market demand may entail an equilibrium market structure of only a few firms. In that situation, a competitive equilibrium is unlikely to be achieved unless there is intense pressure from potential entrants. Recognizing that the competitive solution is not always a good approximation for the equilibrium of an industry, we consider the effects of price regulation in an imperfectly competitive model in a later section.

First-Best Effects

In determining the welfare effects of regulation, the first point to note is that a competitive equilibrium achieves a social welfare optimum. Thus, if price regulation causes price to deviate from marginal cost in an economy that is currently at a competitive equilibrium, then regulation must result in a suboptimal allocation of resources. If price is set in excess of marginal cost, then there is too little of the regulated product produced and consumed. If instead price is set artificially low, then either too much of the product is consumed (if firms are required to meet demand) or too little is consumed and shortages prevail (if firms are left unregulated in their supply decisions). Generally, the farther price is set from the competitive level, the greater is the welfare loss to society.

To consider the effects of entry restrictions along with the regulation of price, suppose the market demand curve and the firm average cost curve are as depicted in figure 16.1. The competitive equilibrium price, denoted \( P^* \), is where price equals minimum average cost. According to figure 16.1, the competitive equilibrium entails twenty firms, each producing at minimum efficient scale of \( \hat{q} \), where \( Q^* = 20\hat{q} \). Recall from chapter 6 that minimum efficient scale is the smallest quantity for a firm such that average cost is minimized. Social welfare is maximized at the competitive equilibrium because price equals marginal cost (so that allocative efficiency is achieved), and the total cost of producing \( Q^* \) is minimized because the efficient market structure is in place (so that productive efficiency is achieved).

Now put in place regulation that specifies that firms must set a price of \( \bar{P} \), which exceeds \( P^* \). The reduction in consumer surplus from the rise in price is measured by trapezoid \( \overline{PabP^*} \). Of course, part of this loss in consumer surplus is transferred to firms in the form of additional profits. Suppose the regulatory agency prohibits any entry into the industry. In that case, each of the twenty firms will produce \( Q/20 \), so that average cost is \( \overline{AC} \), and firm profits will equal \( (\bar{P} - AC)(Q/20) \), that is, rectangle \( \overline{Pad\overline{AC}} \). Subtracting total industry profits from the loss in consumer surplus, the welfare loss from regulation is then the shaded area in figure 16.1.

There are two distinct sources of welfare loss. First, there is the reduction in welfare resulting from the reduction of output from \( Q^* \) to \( \bar{Q} \). This loss is measured by triangle \( abc \). The second source of welfare loss is due to the inefficient market structure maintained under regulation. Each firm is producing at \( Q/20 \). Since this quantity falls below minimum efficient
scale, each firm’s average cost is higher under regulation. The rectangle $A C d c P^*$ measures the value of additional resources used to produce $Q$ relative to the preregulation equilibrium.

It is interesting to note that given price regulation, the imposition of entry regulation raises social welfare. As the regulated price $\bar{P}$ exceeds average cost, there is an incentive for firms to enter the industry. For example, suppose one firm entered. The new equilibrium would still have a price of $\bar{P}$, as that is mandated by the regulatory agency, but each firm would now be producing slightly less, specifically, the amount $Q/21$. Because profits for the new firm equal $(\bar{P} - A C')(Q/21)$, which are positive, entry would occur. However, note that firm average cost has increased from $AC$ to $AC'$ in response to entry. Thus, the total cost of providing industry supply of $Q$ has been increased by the amount $(AC' - AC)Q$. To avoid this additional welfare loss, it is best that the regulatory agency prohibit entry given that it is regulating price. Actually, it may even be best for the regulatory agency to go a step further and actually reduce the number of firms in order to achieve a more efficient market structure. Decreasing the number of firms from twenty to nineteen raises firm output from $Q/20$ to
If $\bar{Q}/19$ is still less than minimum efficient scale, then firm average cost is lower at a firm output rate of $\bar{Q}/19$ than at an output rate of $\bar{Q}/20$. In that case, the total cost of supplying $\bar{Q}$ is reduced by eliminating one of the firms. Of course, these conclusions concerning the optimality of restricting entry are conditional on price regulation already being in place. Regulation of entry may raise welfare given price regulation, but it is clear that price and entry regulation together reduce welfare.

**Second-Best Effects**

If all markets in an economy are at a competitive equilibrium, then the regulation of one market, so that price deviates from the competitive equilibrium price, must reduce social welfare. Now suppose the economy is not initially at a competitive equilibrium because of some preexisting distortion like imperfect competition or regulation or taxes. If a further distortion is imposed through regulation, does social welfare decline? The theorem of second best says, “not necessarily.” If there already exist distortions in the economy such that price is not equal to marginal cost in some markets, further distortions could either increase or decrease social welfare. This issue is of particular relevance for the case of economic regulation in the United States, as one historical pattern that has emerged is for one product to be regulated and then for regulation to be extended to cover substitutes for that product. The theorem of second best tells us that the spread of regulation need not be welfare-reducing.

To see this point, let us examine the case of two products. In the market for product $A$, there are two types of suppliers. Type 1 firms specialize in product $A$ and are able to produce it at a constant marginal cost, $c_1$. Type 2 firms concentrate on supplying a similar product, good $B$, to a different market but can also supply product $A$ though at a higher unit cost, $c_2$, than a type 1 firm’s cost. Figure 16.2 depicts the marginal cost curves for the two different types of firms and the market demand curve for good $A$. In the absence of regulation, a competitive equilibrium is achieved in which only type 1 firms supply product $A$ and do so at a price of $c_1$. Suppose the initial situation is that type 1 firms are required by government to set a price no lower than $\bar{P}$, where $\bar{P}$ exceeds not just the unit cost of type 1 firms but also that of type 2 firms. In that case, the equilibrium has (unregulated) type 2 firms supplying $Q^1$ units of product $A$ at a price of $c_2$.

Now suppose that further regulation is imposed in that all firms supplying product $A$ are required to set price no lower than $\bar{P}$. Furthermore, assume consumers slightly prefer the product provided by type 1 firms, so that if both type 1 and type 2 firms sell at the same price, all demand will go to type 1 firms. Under this assumption, the equilibrium under the new regulatory regime has all firms pricing at $\bar{P}$, but only type 1 firms supplying the market. What

---

is the welfare effect of expanding regulation? First, price goes up from $c_2$ to $\bar{P}$, which reduces consumers surplus by $\bar{P}ab_c_2$. Second, more efficient firms are producing, so that industry profits rise by $\bar{P}adc_1$. If rectangle $c_2edc_1$ exceeds triangle $abe$, then the expansion of regulation has actually raised social welfare. The reason for this result is that the initial regulatory regime had less efficient firms producing, because price regulation discriminated against more efficient firms. By bringing all firms under price regulation, the more efficient firms ended up being the ones that supplied the market. These cost savings could be sufficiently large so as to compensate consumers for the rise in price from expanded regulation. This line of reasoning will be relevant when we examine the effects of simultaneous price regulation in the railroad and trucking industries.

**Direct Effects of Price and Entry/Exit Regulation: The Imperfectly Competitive Model**

Let us now consider a market in which minimum efficient scale is not small relative to market demand so that there are only a few active firms in the industry. In this situation, each firm supplies a significant share of the market and presumably recognizes that their output decisions have a noticeable impact on the market price. Our objective is to assess the effects of price regulation in such a market.

To perform this analysis, we will use the Cournot model of oligopoly (which was presented in chapter 5). In this model, each firm chooses its quantity so as to maximize its profit given
the quantity decisions of the other firms in the industry. Of importance is that a firm takes into account how its quantity choice affects the market price. In figure 16.3, we have supposed that the Cournot equilibrium involves three firms, each producing $q'$, so that the resulting market price is $P'$ as $3q' = D(P')$. By the analysis in chapter 5, we know that firms typically earn positive profits at a Cournot equilibrium because firms restrict supply in order to keep price above cost.

Let us consider the welfare effects of price regulation in this setting. Since price is already too high, as $P'$ exceeds the competitive equilibrium price $P^*$, it follows that if regulation causes price to be raised above the Cournot equilibrium price, welfare will fall. However, if price is reduced, welfare will rise as price is pushed closer to the competitive level. As long as the reduction in price is not too great, regulation that reduces price below the equilibrium level may actually raise welfare when the industry is characterized by imperfect competition.

In evaluating the effects of entry regulation on social welfare, the most important point to realize is that free entry does not necessarily result in an efficient market structure when the market is imperfectly competitive. Recall that for a competitive market, firms are small so that their quantity decisions do not affect price. Thus, entry by a single firm does not affect existing firms’ profits, nor does it affect consumer surplus (because it does not affect the market price). The change in welfare from entry is then measured by the profits of the new
In a competitive industry, entry is profitable (and thereby occurs) if and only if it raises social welfare. The interests of society and the interests of individual firms perfectly coincide. Entry then occurs until the point at which social welfare is maximized.

This harmony of interests breaks down when there is imperfect competition. In that situation firms are not small, so that entry and the ensuing change in industry output do affect the market price. The change in welfare from entry is measured by the change in consumer surplus plus the change in industry profits, which comprises the new firm’s profits and the change in existing firms’ profits. The key point is that entry occurs if and only if a new firm’s profits are positive, but entry is welfare-improving if and only if a new firm’s profits plus the change in consumer surplus plus the change in existing firms’ profits are positive. Clearly, the criterion for the private optimality of entry (that is, for entry to be profitable) differs from the criterion for the social optimality of entry. Just because the profits for a new firm are positive, so that it enters, it need not be true that entry is welfare-improving. Although consumer surplus is generally higher with entry, as price typically declines, industry profits are generally lower because of the fall in price. If the latter exceeds the former, then entry can be welfare-reducing even though it is profitable. Similarly, entry could be welfare-increasing but unprofitable. For example, if a new firm expects to greatly intensify competition so that its entry drastically reduces price, entry will probably be unprofitable. At the same time, it would have probably increased welfare because of the rise in consumer surplus emanating from the sharp fall in price. In this case, there would be too few firms in the industry at the free-entry equilibrium.

To be more concrete, consider the Cournot model with firms offering homogeneous products. Assume the market demand function is \( D(P) = 100 - P \), the firm cost function is \( C(q) = 10q \), and the cost of entry is 150. The entry cost might come from the construction of a production facility or advertising to introduce one’s product. Given a fixed number of active firms, one can calculate the Cournot equilibrium. These calculations have been done for a number of firms between 1 and 7. The resulting levels of price, firm profits, consumer surplus, and social welfare are listed in table 16.1. First note that the free-entry equilibrium is defined by six firms producing in the industry. If a seventh entered, the new firm’s profits (as well as those of every other firm) would be negative, so that entry is unprofitable. Furthermore, all six active firms are earning positive profits, so that none has an incentive to exit the industry. Thus, this is a free-entry equilibrium. Note, however, that social welfare is not maximized when there are six firms. Rather, it is maximized when there are only three firms. Although price is higher with three as opposed to six firms (so that consumer surplus is lower), there are considerable cost savings from having only three firms (specifically, entry costs of 450 are saved by having three fewer firms enter). As a result, industry profits are sufficiently higher with three firms so as to compensate for lower consumer surplus. Free entry then results in too much entry. One can show that there are always too many firms relative to the social optimum under Cournot competition with homogeneous products. If instead firms offer
differentiated products and consumers sufficiently value product diversity, it has been shown that free entry can entail too few active firms. The key point is that the private interests of a firm generally do not coincide with the interests of society.

Given that free entry need not result in an efficient industry structure under imperfect competition, it is unclear as to whether entry/exit regulation raises or lowers social welfare. No general conclusions can be drawn; each particular case must be analyzed on its own merits. As an example of such an analysis, let us consider the case depicted in figure 16.3. In the absence of regulation, the equilibrium price is \( P' \), with each firm producing below minimum efficient scale. Thus, price is excessively high, and there are too few firms in the industry. Suppose regulation sets price below \( P' \) at \( \bar{P} \) and prohibits both entry and exit. Each firm now supplies \( \bar{Q}/3 \) instead of \( q' \). Because \( \bar{Q}/3 \) is quite close to minimum efficient scale, it appears that the industry has the optimal number of firms for supplying \( \bar{Q} \). Although there were too many firms when price was \( P' \), there are now the correct number of firms when price is lowered to \( \bar{P} \). In this situation, the prohibition of any entry or exit is in society’s best interests given the regulation of price. As a matter of fact, regulation is clearly welfare-improving in this case.

Although regulation can be welfare-improving in an imperfectly competitive market, it is perhaps unwise to support a regulatory policy that attempts to fine-tune such markets. One major obstacle to implementing such a policy is that it is often extremely difficult to determine which markets are candidates for regulation, as well as at what level price should be set. This latter problem is especially troublesome when cost and demand conditions change substantially over time. That a market has only a few firms does not imply that regulation is welfare-improving, inasmuch as competition may nevertheless be strong because firms are innately competitive or because potential competition forces them to be so. Regulation in such markets is likely to reduce welfare. A policy of trying to fine-tune

<table>
<thead>
<tr>
<th>Number of Firms</th>
<th>Price</th>
<th>Firm Profits</th>
<th>Consumer Surplus</th>
<th>Social Welfare</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>55</td>
<td>1,875</td>
<td>1,013</td>
<td>2,888</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>750</td>
<td>1,800</td>
<td>3,300</td>
</tr>
<tr>
<td>3</td>
<td>33</td>
<td>356</td>
<td>2,278</td>
<td>3,347</td>
</tr>
<tr>
<td>4</td>
<td>28</td>
<td>174</td>
<td>2,592</td>
<td>3,288</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
<td>75</td>
<td>2,813</td>
<td>3,188</td>
</tr>
<tr>
<td>6</td>
<td>23</td>
<td>15</td>
<td>2,976</td>
<td>3,067</td>
</tr>
<tr>
<td>7</td>
<td>21</td>
<td>-23</td>
<td>3,101</td>
<td>2,937</td>
</tr>
</tbody>
</table>

Table 16.1

Profits and Social Welfare for the Cournot Solution

---

imperfectly competitive markets through price regulation is a perilous task that has been shown historically to be self-defeating. A general policy of relying on unfettered competition seems advisable in markets that are not natural monopolies.

Indirect Effects of Price and Entry Regulation

A commonly observed regulatory policy is one in which price is set above cost, allowing firms to earn above-normal profits at least initially, and new firms are prohibited from entering the industry. Entry is usually not expressly forbidden. Rather, the regulatory agency says that a firm may apply to enter but that entry will be permitted if and only if certain (very stringent) standards are satisfied. A common procedure is for an applicant to be given a hearing at which the regulatory commissioners decide whether the standards for entry have been met. Historically, it has been observed that this procedure can effectively be equivalent to a simple prohibition of entry. Finally, let us assume that although price and entry are regulated, product quality is subject to minimal or no regulation. For reasons described in chapter 10, quality regulation is typically difficult.

Excessive Nonprice Competition

By specifying the price at which firms must sell or a relatively small range within which price must be set, regulation effectively eliminates price as an instrument through which firms compete. To increase the demand for their product, one would expect firms to turn to non-price methods. These include improving the quality of the product, changing its characteristics, providing or extending a warranty, and advertising to make the product appear more attractive.

The intensity of nonprice competition varies from industry to industry, as it depends on the available technology for differentiating products as well as the degree of competition. Some products, like automobiles, are naturally easy to differentiate, while others, like natural gas, are inherently similar and thus resistant to differentiation. Even in the latter case, firms can compete by providing better service to go along with the product. A second factor that influences the degree of nonprice competition is the ability of firms to collude. A regulated industry is fertile ground for collusion, as the same firms interact over time and without fear of entry disrupting a collusive arrangement. If firms can cooperate and prevent excessive nonprice competition (recall that regulation has taken care of the need to collude over price), firms may be able to retain above-normal profits. Otherwise, they may end up competing away above-normal profits through nonprice competition.

Let us consider the welfare effects of excessive nonprice competition. To simplify the analysis, suppose that a product can be produced at either high or low quality. All consumers prefer the high-quality product (good $h$) to the low-quality product (good $l$). The products are imperfect substitutes, so that the demand curve of product $h$ depends on the price of product $l$, $P_l$, as well as the price of product $h$, $P_h$. Similarly, the demand curve for product $l$ depends
on both $P_l$ and $P_h$. Because product $h$ is higher quality, consumers will purchase product $l$ only if its price is sufficiently less than the price of product $h$ and, further, more consumers will buy product $l$, as the price of product $l$ is lower than product $h$’s price. It is also assumed that the unit cost of product $h$, denoted $c_h$, exceeds the unit cost of product $l$, denoted $c_l$. This assumption seems reasonable because it says that a better product costs more to make.

Prior to regulation, it is assumed that the industry is at a competitive equilibrium. This is depicted in figure 16.4. The competitive equilibrium has price equal to marginal cost so that $P_h = c_h$ and $P_l = c_l$. The associated quantities are $Q'_h$ and $Q'_l$ for products $h$ and $l$, respectively. $Q'_h$ is the demand for product $h$ when its price is $c_h$ and the demand curve is $D^h(P_h; P_l = c_l)$. Implicit in the demand curve for product $h$ is that the price of the lower quality substitute is $c_l$. An analogous argument applies to the determination of $Q'_l$. Given that product $h$ is priced at $c_h$, note that the demand for product $l$ is zero when product $l$ is also priced at $c_h$. This lack of demand reflects the fact that if the prices are the same, then every consumer prefers the high-quality product, so that the demand for the low-quality product is zero.

Now consider a regulatory policy which specifies that firms must price their products at $c_h$ regardless of quality. Nonprice competition is presumed to take the form of switching to producing the higher-quality product. The demand and supply of product $l$ are then zero. Because now, by law, product $l$ must be priced at $c_h$, the demand curve for product $h$ shifts out to $D^h(P_h; P_l = c_h)$, as the price of a substitute product has increased. The equilibrium under regulation entails $Q''_h$ units of product $h$ being produced and consumed, while product $l$ is no longer produced.

The welfare loss from regulation is measured by the shaded triangle under the demand curve $D^l(P_l; P_h = c_h)$, as this is the consumer surplus forgone from no longer having the option to buy the low-quality, low-priced product. It is important to note that the increased area under the demand curve for good $h$, resulting from its shifting out to $D^h(P_h; P_l = c_h)$, does not represent a welfare gain. Rather, this area measures the increased willingness to pay for product
given that product \( l \) is no longer available. The increased area under the demand curve for product \( h \) tells us that there is a greater welfare loss from eliminating product \( h \) when product \( l \) is not available relative to when product \( l \) is available. A second point to make is that the increase in cost from supplying market demand, which equals \( Q_h c_h - Q_h' c_h - Q_l' c_l \), is not a measure of the loss realized by consumers from regulation. It is true that consumers who previously purchased the low-quality product are now paying a higher price of \( c_h \) as opposed to \( c_l \), but it is also true that they are receiving a higher-quality product than before. Since this is certainly of value to them, it partially offsets the increase in production costs. Finally, in the event that regulation sets price above \( c_h \), say at \( \bar{P} \), the welfare loss is then the sum of the two shaded triangles. In addition to the loss from product \( l \)’s being “regulated out of the market,” there is the welfare loss from product \( h \)’s being priced above cost.

The basic point to be made is that while regulation limits some avenues through which firms compete, it is difficult to restrict all avenues. In their efforts to maximize profits, firms will shift their activities to those avenues that are unimpeded by regulation. In the case of price regulation, firms will compete more intensively through nonprice methods such as product quality and advertising. From a firm’s perspective, such competition tends to reduce the above-normal profits generated by price and entry regulation. From society’s perspective, such behavior is likely to result in excessive nonprice competition. Regulatory-induced nonprice competition played a central role in the airline industry.

**Productive Inefficiency**

If price regulation allows firms to earn above-normal profits and entry regulation prevents these profits from being competed away through the arrival of new firms, then workers, especially if they are unionized, are likely to try and extract part of the surplus. One obvious way is to demand higher wages. Although a straight transfer of rent from shareholders to workers is not necessarily welfare-reducing, higher wage rates result in the firm’s substituting away from labor and toward other inputs like plant and equipment. Although the firm is still choosing the capital-labor ratio that minimizes its cost, the ratio is not optimal from a social perspective because the cost of labor to the firm (that is, the wage rate) exceeds the opportunity cost of labor to society. Alternatively, workers may extract rent by increasing nonpecuniary benefits—for example, better working conditions. Such activities use up valuable resources and reduce productive efficiency, though they are of value to workers.

A second source of productive inefficiency from price and entry regulation is the continued operation of inefficient firms that would have perished under free entry. In an unregulated environment, new firms replace those firms that are relatively inefficient. Entry regulation neutralizes the mechanism by which efficient firms are rewarded and inefficient firms perish. This analysis suggests that if a regulated industry is deregulated, one would expect entry and exit to occur simultaneously: exit taking place by the less efficient firms and entry occurring to replace those firms.
As evidence of how entry restrictions can affect the natural evolution of an industry, a recent study explored the effect of eliminating restrictions on bank branching. Until the 1970s, almost all states had in place regulations that limited the number of branches that a bank could have within the state (as well as prohibiting interstate branching). The study showed that there was a sharp decline in banks’ operating costs and loan losses after a state introduced unrestricted statewide branching. This evidence is consistent with the hypothesis that the elimination of branching restrictions allowed more efficient banks to grow at the expense of their less efficient rivals. By interfering with this selection process, regulation adversely affected industry performance.

Some Indirect Effects of Price and Exit Regulation

Now consider a policy opposite to the one just considered. The regulated price is set below cost and firms are prevented from exiting the market. Given that price is below cost, there is certainly little incentive to enter the industry. Thus, if there is any need for regulation of market structure, it is through the prohibition of firms leaving the market. As before, firms are required to meet all demand at the set price, and there is minimal regulation of product quality.

Cross-Subsidization

A common regulatory policy is to set price below cost in some markets served by the regulated industry and then, in order to cover losses incurred in those markets, to set price above cost in some other markets. This policy is referred to as cross-subsidization and often takes the form of a product’s being identically priced in different geographic markets, even though the cost of supplying the product differs in these markets.

To measure the welfare effects of a policy of cross-subsidization, consider a regulated industry that offers products 1 and 2. Assume that the demand curves for these two products are independent; that is, the price of one product does not affect the demand for the other product. For whatever reason, the regulatory agency desires to raise the supply of the high-cost product, which is presumed to be product 2. The regulated price for product 2 is set at \( P_2 \), where \( P_2 < c_2 \) and \( c_2 \) is the unit cost of producing good 2 (see figure 16.5). The immediate welfare loss from such a policy is, of course, triangle \( abc \). However, note that the industry is incurring losses equal to \( (c_2 - P_2)Q_2 \) or rectangle \( c_2bcP_2 \). If the firm is to earn at least normal profits, which is necessary in order to raise new capital and avoid bankruptcy, the regulatory agency must increase the price of product 1 from the socially efficient level of \( c_1 \) to \( \bar{P}_1 \). Here \( \bar{P}_1 \) is set to allow a regulated firm to earn normal profits: \( (\bar{P}_1 - c_1)Q_1 + (P_2 - c_2)Q_2 = 0 \). The welfare loss of a policy designed to subsidize the supply of product 2 is then

the sum of triangles \(abc\) and \(def\). In pursuing a policy of increasing the supply of one product, cross-subsidization is often used, even though it entails the spread of deadweight welfare losses to other markets. A notable example is setting identical prices for local phone service for both urban and rural consumers, even though it is more costly to provide it in the latter case. Subsidizing rural consumers was argued to be necessary to achieve universal phone service.

### Reduced Capital Formation

As we just observed, if a firm is forced to serve unprofitable markets, it is likely to have a difficult time earning at least normal profits. This can result in long-run problems through its effect on investment. If some investment is financed internally, reduced profits decrease the amount of capital formation. In addition, the firm faces an increased chance of bankruptcy as a result of having to serve unprofitable markets. In order for investors to be willing to take the additional risk associated with a firm that has a relatively high probability of bankruptcy, the firm has to offer a higher return to investors. This increases its cost of capital and therefore reduces the amount of investment. The end result is likely to be a short-run deterioration of the firm’s capital stock, which reduces capacity, productivity, and product quality. Such a deterioration is likely to be myopic from society’s long-run perspective. Such an effect proved to be a substantive force in the deterioration of the capital stock in the railroad industry.

### Regulation and Innovation

Thus far we have considered the static welfare effects of price and entry/exit regulation. Another important impact of regulation is with respect to dynamic efficiency—specifically,
its effect on the incentive to invest in research and development (R&D), as well as the incentive to adopt innovations. The importance of technological innovation in the modern economy cannot be underestimated. In his famous 1957 study, Nobel laureate Robert Solow concluded that 90 percent of the doubling of per-capita nonfarm output in the United States over the period of 1909–1949 was due to technical advance.\(^6\) Given the importance of technological innovation in the economy, it is essential to consider the ramifications of regulation on the pace of technological progress.

Before beginning our analysis, a proviso of sorts is in order. It is one thing to determine how regulation affects the incentive to invest in R&D and the rate of technological innovation, but it is quite another thing to determine whether regulation results in a suboptimal rate of innovation. Dynamic efficiency does not necessarily imply that firms invest at the greatest rate possible, but rather that there is a particular rate of investment that is socially optimal. More innovation is not always better, because resources must be used in order to discover and adopt innovations. Although a competitive equilibrium results in static efficiency, it is not at all clear as to whether it results in dynamic efficiency. Thus, if regulation results in less investment in R&D relative to a competitive equilibrium, this fact need not imply that there are dynamic welfare losses, because there might be too much R&D expenditure at a competitive equilibrium.

Historically, it has been observed that new firms are an important source of innovation. For lack of a better word, new firms are thought to be vital entrepreneurs that play a crucial role in developing and adopting technological advances. Innovation provides a prospective firm with the opportunity to enter an industry profitably. Regulation that prevents entry or keeps price so low that entry is generally unprofitable closes the door to these entrepreneurs.

If regulation keeps price above cost and allows firms to earn above-normal profits, it is then possible for regulation to result in a greater rather than lower rate of innovation. Retained earnings can be an important source of funds for R&D expenditure, and so regulation that increases the level of industry profits could lead to more investment and more innovations. Alternatively, if price is kept below cost and firms incur losses, this approach should reduce the amount of investment in R&D.

A third effect of price regulation on innovation is through nonprice competition. Recall that with price set by the regulatory authorities, firms will attempt to differentiate their product in order to increase demand. Investing in R&D to achieve product innovations so as to offer a “new and improved” product can be an important avenue for nonprice competition. If such improvements are important, then regulation that keeps price excessively high may result in more product innovations taking place.

---

The Effect of Regulatory Lags on Innovation

If a regulatory agency prevents a regulated firm from reaping the return from innovating by always making it price at average cost, then the regulated firm will have little or no incentive to innovate. In this regard, lags in the regulatory process are conducive to regulated firms’ innovating. Any cost savings from adopting an innovation are retained by the firm until the regulatory agency is able to adjust price. Thus, the time between regulatory reviews allows the regulated firm to have price in excess of average (production) cost and thereby receive a return to innovating.

It has also been shown that regulatory lags not only influence whether a regulated firm adopts an innovation but also affect the speed at which adoption takes place. To see why this statement is true, assume that the regulatory agency always sets price equal to average cost and that the regulated firm’s current production technology generates a constant average cost of \(c\). As shown in figure 16.6, the regulated price will be \(c\). Now suppose an innovation becomes available to the regulated firm that would lower average cost to \(c’\). The regulated firm can do one of three things. First, it can choose not to adopt the innovation. Second, it can adopt the innovation, reduce cost to \(c’\), and receive profits measured by the rectangle \(cabc\). Of course, these above-normal profits are received only for one period, where a period is defined as the time between regulatory reviews. Once the regulatory agency meets, it will reduce price to \(c’’\) so that the regulated firm no longer earns a return from adopting the innovation. As long as the cost to adopting the innovation (for example, having to shut down production) is less than \(cabc\), the existence of a regulatory lag provides the regulated firm with the necessary incentive to adopt the innovation. However, a third alternative may be available, which is to adopt the innovation gradually. Suppose the regulated firm can partially adopt the innovation and reduce its cost to \(c’\). After the next regulatory review (which will reduce price to \(c’\)), it can complete the adoption and reduce its cost to \(c’’\). From this strategy, the regulated firm earns profits of \(cadc’\) in the first period and \(c’efe’’\) in the second period. Ignoring discounting and assuming the cost of full and gradual adoption are the same, the regulated firm earns higher profits (measured by \(defb\)) by gradually adopting the innovation. Thus, regulatory lag influences not only the incentives to adopt cost savings innovations but also the speed at which adoption takes place.

Regulation and Productivity Growth

It is well known that the growth rate of productivity in the United States (as measured by, for example, output per person-hour) fell considerably in the 1960s and 1970s. It is also

---

noteworthy that the extent of government intervention in the economy accelerated after the mid-1960s. It is natural to wonder to what extent the slowdown in productivity growth is due to a rise in the amount of regulation.

One study explored this question by estimating the determinants of labor productivity in U.S. manufacturing over 1958–1977. Three different measures of aggregate regulation were used: (1) the cumulative number of major pieces of regulatory legislation in effect, (2) federal expenditures on regulatory activities, and (3) the number of full-time federal personnel engaged in regulatory activities. Depending on which measure of regulatory intensity was used, it was estimated that between 12 percent and 21 percent of the slowdown in the growth of labor productivity in U.S. manufacturing during 1973–1977, as compared to 1958–1965, was due to the growth in federal regulation.

Methods for Estimating the Effects of Regulation

There are two fundamental reasons for which we should be concerned with estimating the quantitative effects of regulation on price, product quality, productive efficiency, and other relevant variables. Having put forth a theory as to the effects of regulation, we now must determine its validity. Are the predictions of the theory consistent with actual regulatory experiences? For example, do we indeed find that price and entry regulation induces excessive nonprice competition? Does price and exit regulation that keeps price above cost result in reduced capital formation? A second motivation for estimating the effects of regulation is to determine, quantitatively, the welfare implications of alternative policies, in particular, of deregulation. Such estimates should be central to public policy debates with respect to regulation. For both of these reasons, this section investigates different methods for quantitatively estimating the effects of economic regulation.9

Overview of Estimation Methods

The situation we are faced with is as follows. An industry is or has been subject to price and entry/exit regulation. Our objective is to estimate the impact that regulation has had on important economic variables. These variables include price, cost, product quality, product characteristics, capital investment, wages, and technological innovation. All these variables are relevant to assessing the welfare implications of regulation, as they affect allocative efficiency, productive efficiency, and dynamic efficiency.

To determine the effects of regulation, we must compare the values these variables would have taken in the absence of regulation with the values they actually did take under regulation. Because the industry is or has been regulated, in principle there is no difficulty in collecting data on these variables under regulation. Of course, in practice, some of these variables can be quite difficult to measure, for example, the quality of a product. This difficulty aside, one of the central tasks in estimating the effects of regulation is to derive a nonregulatory benchmark—that is, to determine the values that these variables would have taken in the absence of regulation. There are three basic methods that have been used to estimate a nonregulatory benchmark, and we now turn to them.

Intertemporal Approach

The *intertemporal* (or *time-series*) approach compares the industry under study during years for which was regulated with years for which it was not regulated. The nonregulatory

benchmark is then the industry under study at a different time. This method requires that the sample period for which one has data include years for which the industry was regulated and years for which it was not. We refer to this as the intertemporal approach because it compares variables across time.

In assessing the effects of regulation it can be quite misleading to compare values for the relevant economic variables in years with regulation with values of those same variables in years without regulation. Given that many factors other than the regulatory environment change over time, the movement in economic variables may be only partly due to changes in regulation. For example, suppose one observes that profits are lower after an industry is deregulated. Lower profitability could be due to regulation’s having kept prices artificially high, but it could also be due to the business cycle. If the economy moved into a recession about the time of deregulation, lower profits might be due to the recession-induced shift of the market demand curve, and not to the fact that deregulation reduced prices. Therefore, when utilizing the intertemporal approach, one must consider other relevant factors that might be changing across time. The demand curve might shift across time because of such factors as the business cycle, changes in preferences, or the degree of foreign competition. In estimating the effects of regulation on productive efficiency, one needs to control for exogenous changes across time in input prices and the production technology.

With the intertemporal approach, a valuable indicator of the perceived effect of regulation on firm profits is the share price of a firm’s common stock. Referred to as an event study, this method entails observing how a firm’s share price changes in response to policy announcements concerning deregulation. Given that the share price is an index of the market value of the firm based on the information available to stock market traders, a fall in it reveals that these traders expect deregulation to reduce the future stream of firm profits. This suggests that regulation was beneficial to firms, perhaps because prices were kept artificially high and entry was prohibited. However, if the share price was to rise in response to an announcement that the industry will be deregulated, then this response reveals a belief among market participants that regulation depressed firm profits. This effect on profits could be due to keeping price too low or stifling innovations. In using the event study approach, it is important to keep in mind that movements in a firm’s share price reflect, at best, all the information that is currently available. If an announcement of deregulation depresses the share price, the decline indicates that, based on current information, regulation was beneficial to firms. Because information is inherently incomplete, whether in fact regulation was beneficial is another matter.

Application: New York Stock Exchange

From its inception in 1792 until major deregulatory legislation in 1975, the New York Stock Exchange (NYSE) set minimum commission rates on transactions conducted by its members.\(^{11}\) Given that its members always chose to set their rates equal to that minimum, in practice the NYSE set commission rates. The NYSE also required that commission rates be independent of the size of the order. Its members were not allowed to offer quantity discounts, even though there are obvious scale economies in performing securities transactions. Finally, we should note that while the NYSE set standards for member behavior, these standards were enforced by the industry’s regulatory arm, the Securities and Exchange Commission (SEC).

Regulation resulted in considerable discrepancies between commission rates and cost. In December 1968, the commission rate set by the NYSE was $.39 per share. Table 16.2 describes the relationship between the established rate and the estimated cost per share. Due to scale economies in transactions, the average cost per share was declining in the number of shares transacted. As the numbers reveal, cross-subsidization took place as consumers with relatively large orders subsidized consumers with relatively small orders.

The deregulation of rates began in the early 1970s. In 1971 the SEC ordered the NYSE to allow its members and their clients to freely negotiate commission rates on large orders, specifically, on the portion of an order in excess of $500,000. This deregulation largely

Table 16.2
Commission Rate, Cost, and Profit on $40 Stock by Order Size (1968)

<table>
<thead>
<tr>
<th>Shares per Order</th>
<th>Commission per Share(^*)</th>
<th>Estimated Cost per Share(^†)</th>
<th>Profit per Share</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>$0.39</td>
<td>$0.55</td>
<td>$(0.16)</td>
</tr>
<tr>
<td>200</td>
<td>0.39</td>
<td>0.45</td>
<td>(0.06)</td>
</tr>
<tr>
<td>300</td>
<td>0.39</td>
<td>0.41</td>
<td>(0.02)</td>
</tr>
<tr>
<td>400</td>
<td>0.39</td>
<td>0.39</td>
<td>0.00</td>
</tr>
<tr>
<td>500</td>
<td>0.39</td>
<td>0.37</td>
<td>0.02</td>
</tr>
<tr>
<td>1,000</td>
<td>0.39</td>
<td>0.32</td>
<td>0.07</td>
</tr>
<tr>
<td>5,000</td>
<td>0.39</td>
<td>0.24</td>
<td>0.15</td>
</tr>
<tr>
<td>10,000</td>
<td>0.39</td>
<td>0.23</td>
<td>0.16</td>
</tr>
<tr>
<td>100,000</td>
<td>0.39</td>
<td>0.21</td>
<td>0.18</td>
</tr>
</tbody>
</table>

\(^*\) Based on commission schedule in effect as of December 5, 1968.


applied to institutional investors such as managers of pension funds. The SEC continued to deregulate throughout the early 1970s by reducing the minimum order size at which negotiation was allowed. The legislative branch of the government entered the deregulatory process by passing the Securities Act Amendments of 1975. This legislation mandated that the SEC prohibit the NYSE from fixing commission rates.

Figure 16.7 provides a time series of average commission rates for individual and institutional investors during the first five years of deregulation. Rates fell drastically. Almost immediately rates dropped about 25 percent in response to deregulation. Because of cross-subsidization, deregulation resulted in commission rates rising for small orders (at least for noninstitutional transactions) and falling for large orders. Rates fell more than 50 percent for orders in excess of 10,000 shares. For another example of the intertemporal approach, see the 44 Liquormart decision below.

Intermarket Approach

If an industry is currently regulated and has been so for some time, the intertemporal approach is not very useful. For one reason, data might not exist that go back to the time when the industry was not regulated. Even when data do exist, they still may not allow one to determine what the effects of regulation would be today. If it has been a considerable length of time since the market was unregulated, the industry is liable to have changed so much that
one could not reasonably determine what the effects of regulation are today or what would be the effect of deregulation.

The deregulation of many industries in the 1980s provided economists with the data that allowed the use of an intertemporal approach. However, before that time, when industries had been regulated for decades, economists looked to other markets for a nonregulatory benchmark. Referred to as the intermarket approach (or the cross-sectional approach), this method compares two markets that offer similar products and have similar demand and cost functions. However, they differ in an essential way: one market is regulated whereas the other is not. By comparing economic variables for these two markets, one can derive estimates of the effect of regulation.

Typically, the markets differ geographically in that firms provide similar products or services with similar technologies but provide the same product or service in different geographic areas, for example, different states. In the case of the intertemporal approach, one has to be careful to control for changes over time in factors other than regulation if one is to isolate the effects of regulation. There is a related concern for the intermarket approach. On the one hand, since one is comparing two markets at the same point in time, one does not have to be concerned with such things as the business cycle or changes in technologies. On the other hand, one is comparing two distinct markets, and they are likely to differ more than just geographically. They could differ in terms of input prices like wages or the elasticity of demand. To use the intermarket approach effectively to estimate the effects of regulation, one needs to control for these differences in the two markets.

The need to control for intermarket differences is made more acute when one realizes that regulation is endogenous. Since a state chooses its regulations, a state with a lot of regulation can be expected to differ from a state with little regulation, with those differences being responsible for their different policies. For example, states like California and New York are often trendsetters in government policies, and it is fair to say that those states are atypical in many ways. Now suppose those same traits that explain different regulations across states also directly influence price and other relevant economic variables that are used to measure the impact of regulation. Simply observing that price is, say, higher for states with more regulation may not reflect a causal relationship between regulation and price. Rather, it may be due to the fact that states with, say, higher income prefer to have more regulation, and in addition have higher prices because demand is higher due to higher income. Higher prices and more regulation are both due to higher income. Controlling for differences across markets is then crucial, given that regulation is typically endogenous.

Application: Advertising of Eyeglasses

In studying regulation, one is often struck by its pervasiveness and its idiosyncrasies. It is not surprising that electric utilities and local telephone companies are regulated. But why should state regulatory agencies control the advertising of eyeglasses and eye examinations?
Yet in the 1960s, approximately three-quarters of states did just that. Some states outlawed the advertising of just price information; others prohibited the advertising of any information concerning eyeglasses and eye examinations.

A ban on advertising may either raise or lower price. By advertising, a firm may be able to differentiate its product and thereby increase the demand for its product. Generally, this tactic would result in a higher price. Based on that effect, an advertising ban would tend to reduce price. However, advertising reduces search costs incurred by consumers, leading to more comparison shopping and thus more intense price competition. Given that advertising makes it less costly for consumers to learn which firm has the lowest price, a firm with a high price will experience lower demand, whereas a firm with a low price will experience higher demand compared to when advertising is prohibited. Because a firm’s demand curve is more sensitive to price, a firm will tend to set a lower price in order to realize a sizable increase in its sales. The end result is more intense price competition among firms and lower prices. This effect suggests an advertising ban would raise prices. Given these two counteracting forces, it is unclear whether regulation that restricts advertising would raise or lower price.

To estimate the effect of advertising regulation on price, a study by Lee Benham compared the price of eyeglasses in states without regulation with that in states with regulation. His data consisted of a 1963 national survey of individuals who had purchased eyeglasses, so that the data set was not a time series but rather a cross-section of individuals in different states at a particular point in time. It was found that the average price paid for eyeglasses in states without advertising restrictions was $26.34, whereas in states with advertising restrictions it was $33.04. This evidence supports the hypothesis that advertising restrictions reduced the intensity of price competition by raising consumer search costs and thereby raised the price of eyeglasses.

In any empirical study, it is critical to play the devil’s advocate by trying to think of other factors that could explain one’s findings. With respect to the case at hand, could the observed price differential be due to factors other than state regulations? Not all eyeglasses are the same. Suppose that consumers with higher income tend to buy higher-quality eyeglasses. If states with advertising regulations also tend to have higher income (for whatever reason), it is then possible that the observed price differential is due not to advertising restrictions but rather to differences in the purchased quality of eyeglasses. To attempt to control for this and other factors, Benham estimated the price paid for eyeglasses as a function of family income, sex, age, family size, and, of course, whether or not the state restricted advertising. He found that state regulations caused price to be higher by $7.48. This price differential is actually higher than the $6.40 estimated previously.

The evidence in this study clearly supports the hypothesis that advertising regulation raises prices.

**Application: 44 Liquormart Decision**

A recent study blended the intertemporal and intermarket approaches in examining the effects of advertising regulation on prices. With the *44 Liquormart* decision in May 1996, the U.S. Supreme Court struck down a Rhode Island state law banning the advertising of liquor prices. The study’s approach to measuring the effect of advertising regulation is to compare liquor prices before and after this judicial decision. However, rather than try to control for all of the factors that could also cause retail liquor prices to change over time—for example, the wholesale price of liquor and wages—the authors compared the change in Rhode Island prices over June 1995 to June 1997 with the change in prices for the same products in the neighboring state of Massachusetts. Throughout the sample period, price advertising was legal in Massachusetts. If liquor stores in Rhode Island and Massachusetts were subject to the same set of factors influencing their cost, then, if liquor prices in Rhode Island decline relative to those in Massachusetts, one can be reasonably assured that it is due to the change in regulation. The study is then using intertemporal variation in Rhode Island prices to draw an intermarket comparison with the intertemporal variation in prices in the unregulated market of Massachusetts. Results showed that prices for advertised liquor products fell by 20 percent in Rhode Island compared with prices of the same products in Massachusetts. Consistent with the study for eyeglasses, regulations on advertising raise prices.

**Counterfactual Approach**

If neither an intertemporal nor an intermarket approach can be used, then one can, so to speak, create a nonregulatory benchmark. This is achieved by using data for the regulated industry to simulate what the industry would look like if it had not been regulated. The *counterfactual approach* has been used, for example, to estimate regulatory-induced allocative inefficiencies from crude oil price controls.

A typical application of the counterfactual approach is as follows. One first estimates the market demand curve and the firm marginal cost curve. The next step is to compare quantity under regulation with the quantity derived by evaluating the estimated market demand curve at a price equal to marginal cost. If a competitive equilibrium would be achieved in the absence of regulation and the estimates of the market demand curve and marginal cost are relatively precise, then one can derive an estimate of industry supply in an unregulated market. With regulated supply and the estimate of unregulated supply, one can then estimate the effect of regulation on consumer surplus.

The counterfactual approach is the least desirable of the three methods of estimation. One reason is that it requires making numerous assumptions about what the industry would have looked like in the absence of regulation. It is typically assumed that the cost curves would be the same with and without regulation and that a competitive equilibrium would be achieved. On the one hand, because regulation often reduces productivity, assuming that cost is the same probably underestimates the benefits from deregulation. On the other hand, the postregulation market may be one of imperfect competition. Assuming a competitive equilibrium would then overestimate the gains from deregulation. A second drawback from the counterfactual approach is that it rarely can shed any light on productive inefficiencies created by regulation. Who knows what innovations would have taken place in the absence of regulation? For example, no one predicted that deregulation of the airline industry would cause the widespread adoption of the hub-and-spoke system.

When the data are available, the counterfactual approach has been used in conjunction with either the intertemporal or intermarket approach. In trying to estimate the effects of regulation, the intertemporal approach compares data in years with and without regulation. Alternatively, one might ask what the regulated years would have looked like if they had not been regulated. To address that question, one can use data from the unregulated years to estimate how exogenous variables like input prices, prices of substitutes, and the business cycle impact the industry equilibrium in terms of price, the number of firms, and other endogenous variables. With this estimated relationship, one then plugs in the values for these exogenous variables for the regulated years to come up with a simulated unregulated industry equilibrium. For the regulated years, one can then compare these simulated values with the actual observed values to derive a measure of the effects of regulation that is distinct from a pure intertemporal approach. If instead data are available for regulated and unregulated markets for only one year, one can perform an analogous experiment by using the intermarket approach in conjunction with the counterfactual approach.

**Application: State Usury Laws**

Aristotle considered money to be sterile, and thus the breeding of money to be unnatural. Attitudes like this one have persisted throughout time and periodically have led to laws that either prevent interest from being paid on loans or, more generally, limit the rate of interest. Regulations that specify a maximum rate of interest that an institution can charge for lending money are known as usury laws.

Most states in the 1970s had some form of usury law. With regard to conventionally financed residential mortgages, only eight states had no usury ceiling, and fifteen states had usury ceilings of 10 percent or lower. A 10 percent usury ceiling means that a bank could not lend money at an interest rate exceeding 10 percent. Many of these laws had been in place for decades, but they had no real economic impact for much of that time because market-clearing interest rates generally fell below the legal maximum. However, the rampant
inflation of the 1970s drove interest rates up, so that suddenly these usury ceilings became a binding constraint faced by lending institutions.

To understand the implications of usury ceilings, one first needs to understand that borrowers and lenders care about the real rate of interest and only indirectly care about the interest rate at which they trade, which economists refer to as the *nominal interest rate*. By subtracting the rate of inflation from the nominal interest rate, one derives the *real interest rate*. For example, if the nominal rate is 9 percent and the inflation rate is 5 percent, then the real interest rate is 4 percent. Consider someone borrowing $10,000 on January 1, 1994, under the agreement that it is to be paid back with interest on December 31, 1994. If the bank lends the money at a 5 percent (nominal) rate, then the borrower must pay back $10,500 at the end of the year. If the inflation rate was also 5 percent in 1994, then the $10,500 the bank receives at the end of the year buys the same amount of goods that $10,000 purchased at the beginning of the year (as prices have risen by 5 percent). For having forgone the use of $10,000 for a year, the bank has nothing to show for it! If instead the inflation rate had been 0 percent, then $10,500 received on December 31, 1994, would have meant $500 more in goods than the bank could have bought at the beginning of the year. In this case the real interest rate is 5 percent, whereas in the former case it was 0 percent. What matters for lending and borrowing decisions is the real interest rate.

To consider the effects of a usury ceiling, plotted in figure 16.8 is the demand curve for real loans (that is, after controlling for the inflation rate), denoted $D(r)$, and the supply curve, $S(r)$, where $r$ is the nominal interest rate (what is observed in the market). Let $i$ denote the inflation rate associated with these demand and supply curves. In an unconstrained market, equating the supply and demand curves yields an equilibrium or market-clearing interest rate of $r'$. Note that the associated real interest rate is then $r' - i$. With a usury ceiling of $r_u$, regulation is not binding, because the market-clearing rate of $r'$ falls below $r_u$. Now suppose the rate of inflation jumps up to $i + d$. Holding the nominal rate fixed, the rise in inflation lowers the real interest rate. As a result, consumers demand more loans at a given nominal rate, so that the market demand curve shifts out to $D'(r)$. Of course, a higher rate of inflation means that lending institutions are less willing to supply loans at a given nominal rate, so that the market supply curve shifts in to $S'(r)$. In the absence of a usury ceiling, the new market-clearing rate would be $r' + d$, which is just the original nominal rate $r'$ plus the change in the rate of inflation. Note that the real amount of loans remains at $L'$. As long as the nominal rate of interest can freely adjust with the rate of inflation, all that happens is that the nominal numbers change. Real economic activity remains the same.

Now let us consider the implications of having a usury ceiling of $r_u$. When inflation rises to $i + d$, the nominal interest rate is prevented by law from fully adjusting to $r' + d$. It is only allowed to rise to $r_u$. Because the real interest rate has fallen from $r' - i$ to $r_u - i - d$ (recall that the real interest rate remains the same if the nominal rate is allowed to rise from $r'$ to
there is excess demand for loans. With an inflation rate of \( i + d \) and a nominal interest rate of \( r_u \), consumers demand \( L^0 \) loans, but only \( L'' \) are supplied. Excess demand is then \( L^0 - L'' \).

This situation arose in the market for residential loans in the mid-1970s. In those states with usury ceilings, the allowed nominal rates often fell below what was required to equate supply and demand. Since more loans were demanded than were supplied by lending institutions, how were these scarce loans allocated among consumers? One argument is that the loans went to those consumers who were willing to accept the least attractive terms. Lending institutions have plenty of demand for loans and are thus in the position of being able to demand terms that make the loan less risky to them. In the case of residential loans, this policy could take the form of requiring a higher down payment (which makes the loan less risky to lenders, since the ratio of loan to property value is higher) or loans of shorter length.

The basic theoretical prediction is as follows. Let \( r^* \) denote the market-clearing nominal interest rate for residential mortgages. If \( r^*_m > r_u \), then the usury ceiling is binding. In that case, the theory predicts that the greater is the distortion, as measured by \( r^*_m - r_u \),
the greater is excess demand and the more attractive are loan terms to lenders. This relationship takes the form of a higher average ratio of loan to property value and a shorter average loan maturity. The difficulty in testing this theory is that the market-clearing rate $r_m^e$ is not observed when the usury ceiling is binding. As we will describe, some researchers solved this problem by jointly using the counterfactual and intertemporal approaches.

To estimate the effects of usury ceilings on loan terms, a study by Steven Crafton examined quarterly data during 1971–1975 for residential mortgages.\textsuperscript{14} What is interesting about that time period is that usury ceilings were binding in some but not all quarters. The research strategy was to use data from those quarters for which the market-clearing interest rate for residential mortgages was observable (that is, the usury ceiling was not binding) to estimate the relationship between exogenous variables and that market-clearing rate. The estimated relationship was as follows:

$$r_m^e = 3.186 + 0.4526r_{AAA} + 1.471r_{adv} + 0.1195(r_m)_{-1}$$

where the variables on the right-hand side of the equality are the exogenous variables. The variable $r_{AAA}$ is the rate of AAA-rated bonds, $r_{adv}$ is the interest rate paid by lending institutions to borrow funds from the Federal Home Loan Bank Board, and $(r_m)_{-1}$ is the mortgage rate from the previous quarter. What all this says is that when the usury ceiling is unbinding, the market-clearing rate for residential mortgages is approximately equal to 3.186 plus 0.4526 multiplied by the value for $r_{AAA}$ plus 1.471 multiplied by the value for $r_{adv}$ plus 0.1195 multiplied by the value for $(r_m)_{-1}$.

This relationship can be used to come up with an estimate of $r_m^e$ when the usury ceiling was binding. For a quarter in which it was binding, one plugs in the values for $r_{AAA}$, $r_{adv}$, and $(r_m)_{-1}$ in the equation. The resulting number is the simulated value for $r_m^e$, which we will denote $\hat{r}_m^e$. The estimated distortion in nominal interest rates due to the usury ceiling is then estimated to be $\hat{r}_m^e - r_u$. The final step in this approach is to determine, when the usury ceiling was binding, whether there was a relationship between the estimated distortion $\hat{r}_m^e - r_u$ and the terms of the average residential mortgage. Recall that the theory predicts that the greater is the discrepancy between the market-clearing rate and the usury ceiling, the more favorable are the terms of the loan to the lenders because of greater excess demand. Consistent with the theory, it was found that the greater was $\hat{r}_m^e - r_u$, the greater was the ratio of loan to property value (as lenders were demanding higher down payments) and the smaller was the maturity of the loan.

Measuring the Return to Price and Entry Restrictions: Taxicab Regulation

Regulatory History

Before the 1920s the taxicab industry was largely unregulated. As the automobile became an integral part of transportation, the demand and supply functions of taxicab services shifted out. The 1920s proved to be a growth phase for both taxicab services and taxicab regulations. Fare regulation became increasingly common, as did other legal restrictions such as the requirement that taxicabs be insured. Though regulation of the taxicab industry grew during the 1920s, local government left entry into the industry largely unregulated.

After the onset of the Great Depression, massive entry took place into the taxicab industry. Some of this entry was due to the sharp rise in the number of unemployed workers. With few other job opportunities available, some individuals took to driving taxicabs. With the increased number of competitors, fare competition heated up and taxicab drivers and owners faced falling profits. What ensued were local governments’ placing restrictions on entry into the taxicab industry:

The regulation movement spread throughout the country. In Massachusetts, Frank Sawyer [owner of Checker Taxi Company] urged the state to regulate taxis, and in 1930 the legislature limited the number of cabs in Boston to 1,525 (the same as in 1980). New York City first limited the number of cabs in 1932 under the sponsorship of Mayor Jimmy Walker, but when Walker was forced to resign when it was discovered that he had been bribed by one of the taxi companies, the attempt at regulation failed. Five years later, however, the Haas Act in New York City froze the number of taxi medallions at 13,500.15

This appears to be a classic example of the economic theory of regulation (see chapter 10). Each taxicab company would gain a lot from regulation, while each consumer would only be harmed a little. Furthermore, there are many fewer taxicab companies than consumers, so that the cost of organizing political support is much lower for the former. As a result, the taxicab companies were more effective in getting regulation put in place than were consumers in preventing regulation.

The regulatory program that took place in the 1930s had a lasting impact on the industry. To an extent that we will describe later, regulation has prevented entry in most taxicab markets. A notable exception is Washington, D.C., which does allow entry. Since 1979 there has been a minor deregulatory movement in a few cities that has reduced restrictions on fares and the number of taxicabs. Nevertheless, the taxicab industry has largely escaped the program of deregulation that began in the 1970s.

Entry Restrictions

Taxicab regulation encompasses control over price, the number of competitors, and certain practices. Cities set either fares or ceilings for fares. Entry restrictions take different forms in different cities. A common method is to mandate that a taxicab, in order to operate, must own a medallion. Medallions are issued by the city and are limited in supply. The number of medallions provides an upper bound on the number of taxicabs (the number of taxicabs can be less than the number of medallions if a taxi company chooses not to use a medallion it owns). In most cities medallions can be sold and their ownership transferred. Cities that have pursued this type of entry regulation include Baltimore, Boston, Chicago, Detroit, New York, and San Francisco. An alternative method of limiting the number of competitors is to limit the number of taxi companies, and possibly the number of taxicabs as well. This practice has been used in Cleveland, Dallas, Houston, Los Angeles, Philadelphia, Pittsburgh, and Seattle.\(^{16}\)

In practice, cities have severely constrained the number of taxicabs. In New York City, 13,566 medallions were issued in 1937. With nearly 2,000 returned to the city around World War II and then 400 sold by the city in 1996–1997, the supply of medallions was 12,187 in 2003. With a growing population, residents have had to suffer with a declining number of taxicabs per capita. Similar experiences have occurred in most major cities. In Boston, the number of taxicabs has been fixed at 1,525 since 1930. In Detroit, the number has been 1,310 for over forty years. The city of Chicago allowed 4,108 taxicabs to operate in 1934, then reduced this number to 3,000 in 1937. Since 1963, 4,600 taxicabs have been allowed to operate.\(^{17}\)

The Value of a Medallion

Perhaps the best method for assessing the value of price and entry restrictions is to determine how much a firm is willing to pay in order to operate in the industry. Although this is often a difficult piece of information to acquire, it is a simple task for the taxicab industry. As in many cities, a taxicab must have a medallion in order to operate, and since the number of medallions is fixed, prospective taxicab operators must purchase a medallion in a secondary market. If the price in the market for medallions is positive, then the number of competitors must be less than the number that would exist under free entry (where, effectively, the price of a medallion is zero).

The market value of a medallion is more than just a rough indicator of the effectiveness of entry restrictions. A medallion’s price tells us exactly what the most informed agents


\(^{17}\) Ibid.
believe to be the discounted stream of above-normal profits from economic regulation. To see this, consider an individual with a taxicab who is faced with two alternatives. He can freely enter and operate in Washington, D.C., or he can buy a New York City medallion and operate there. The equilibrium price of a medallion is set by the market so that a prospective firm is indifferent between the two alternatives. If it were not, the market for medallions would not be in equilibrium; if firms could expect to earn more by buying a medallion and operating in New York City, then this expectation would increase the demand for New York City medallions and drive the price up until the price made firms indifferent between the two alternatives. Because the Washington, D.C., market is subject to free entry, it is logical to presume that normal profits are being earned by firms there. Thus, the price of a medallion in New York City must equal the additional profits that can be earned by operating in a regulated market as opposed to an unregulated market. Specifically, it is equal to the discounted sum of future excess profits that are earned by a taxicab operating in New York City. For example, suppose operating a taxicab in a regulated market yields above-normal profits of $10,000 per year for the infinite future. If the interest rate is 4 percent, then the market value of a medallion is $10,000/0.04, or $250,000. Competition for medallions should then drive the price up to $250,000.

Figure 16.9 presents the market price of a New York City medallion over 1973–2003. The 2003 price was around $250,000, which meant that the total value of medallions was in excess of $3 billion. This amount represents above-normal profits achieved through fare and entry regulation. It is then not surprising that entry restrictions persist: as a current holder, a New York City cab owner would stand to lose $250,000 if free entry were allowed. In comparison, the value of deregulation to each consumer of taxicab services would be much lower. Furthermore, the holding of medallions is typically concentrated in a few large taxicab companies, making this interest group more effective in providing political support in exchange for continued entry restrictions. Nevertheless, the New York City Council passed a law in July 2003 that authorized the sale of up to 900 additional medallions in April 2004. The primary motivation seemed to be to raise revenue for the city. But at the same time, the plan was to raise taxi fares by 25 percent, with the express purpose of preserving the value of a medallion. Entry regulation is weakened at the same time that price regulation is strengthened. Isn’t politics beautiful?

Summary

This chapter had two objectives. First, we sought to provide an introductory analysis of the effects of price and entry/exit regulation on allocative and productive efficiency. We found that the static welfare effects of price regulation depend on whether entry is also regulated, on whether there exist unregulated substitutes for the regulated product, and on whether the
industry is imperfectly competitive. Some indirect effects of price and entry/exit regulation were also identified. Setting price above cost can result in excessive nonprice competition and productive inefficiencies, while setting price below cost can spread welfare losses to other markets (through cross-subsidization) and result in reduced capital formation. Although dynamic welfare effects are more difficult to classify, we argued that they could be substantial. Price regulation reduces the incentive to innovate because it limits the returns to innovating. Fortunately, regulatory lags can offset this and thus provide greater incentive to adopt cost-saving innovations. Entry regulation cuts off a source of innovation in the form of entrepreneurs.

The second objective of this chapter was to briefly describe alternative methods for estimating the quantitative effects of regulation. We found that one can measure the impact of regulation by comparing regulated and unregulated markets at a point in time, by comparing a market before and after regulation, and by comparing a regulated market with projections of what it would look like if it were deregulated. The ensuing chapters focus on applying this theory and the methods for measuring the effects of regulation. In particular, the regulation
of the transportation industry, reviewed in the following chapter, provides a most relevant application of the concepts provided in this chapter.

Questions and Problems

1. Assume the market demand function is \( D(P) = 100 - P \) and the firm cost function is \( C(q) = 20q \). The industry is populated by many small firms that offer identical products. In the absence of regulation, a competitive equilibrium would be achieved. However, regulation is in place and requires that a firm’s price be at least as great as 30. Derive the effect of regulation on quantity, firm profits, and social welfare.

2. In following up question 1, now suppose that a firm in the industry discovers and patents a technological innovation that reduces its unit cost to 10.
   a. Derive the welfare gain from deregulation.
   b. How does your answer to part a change if the firm with the innovation can license other firms to use its technology?

3. For the case in table 16.1, find the per-firm subsidy or tax that results in the socially optimal number of firms.

4. Assume the market demand function is \( D(P) = 1,000 - 20P \), the firm cost function is \( C(q) = 10q \), and there are twenty firms in the industry. Regulation requires that each firm set a price of 20. Suppose that an innovation becomes available to all firms that will reduce unit cost from 10 to 5. The cost of adopting this innovation is 50.
   a. Derive the value of this innovation under regulation.
   b. Derive the value of this innovation under deregulation.

5. Suppose the government considers regulating the price of automobiles. What difference does it make if it also regulates the quality of the automobile sold?

6. The number of seats on the New York Stock Exchange is a set number. Ownership of a seat is transferable and sold for about $850,000 in 1994. What does the market value of a seat on the NYSE measure?

7. In 1970, seats on the NYSE traded between a low price of $130,000 and a high price of $320,000. In 1974, they traded between $65,000 and $105,000. What do you think caused this reduction in the market value of a NYSE seat? Was it deregulation or something else?

8. In 1979, New Jersey had a usury law that restricted the interest rate on conventionally financed residential mortgages to not exceed 9.5 percent. This usury ceiling was more restrictive than the one in the state of New York. Recognizing that people who work in New York City can choose to live in suburbs located either in New Jersey or in New York, what effect does the more restrictive usury ceiling in New Jersey have on housing prices in the New Jersey suburbs relative to housing prices in the New York suburbs?

9. Until the 1977 Supreme Court decision in *Bates v. State Bar of Arizona*, all states prohibited advertising by attorneys. That decision gave constitutional protection to an attorney’s right to
advertise the availability of his or her services and the fees to perform routine legal services. What effect do you think the 1977 decision has had on legal fees?

10. In 1998, a New York City taxicab medallion had a market value of $230,000. If the interest rate is 6 percent and entry restrictions are not expected to change in the future, what is the amount of above-normal profits earned annually?

11. If you bought a New York City taxicab medallion and operated a taxicab, would you be earning an above-normal rate of return on your investment?

12. Suppose a taxicab company is expected to earn annual above-normal profits of $10,000. Assume the interest rate is 5 percent.

   a. Derive the price of a medallion if entry regulation is expected to continue forever.

   b. Suppose it is announced by the city government that they will allow free entry into the taxicab industry in exactly two years. Derive the change in the price of a medallion in response to this announcement. Would this price differ from the price it would be selling for next year?
17 Economic Regulation of Transportation: Surface Freight and Airlines

From the mid-1970s to the early 1980s, the United States witnessed an unprecedented program of deregulation. Industries that had been long under government control found themselves “free at last.” From the perspective of society, this period of deregulation is important because it generated considerable welfare gains. From the perspective of economic science, it is also a most significant period. Episodes of deregulation provide economists with natural experiments for testing theories concerning the effects of regulation. By investigating how prices, product quality, product variety, productive efficiency, and other important economic variables respond to deregulation, we can learn about the effects of regulatory policies. In other words, we can gain information as to what would have taken place in the absence of government regulation.

The objective of the current chapter is to put the theory of chapter 16 into use by exploring the impact of regulatory policies in several important markets of the U.S. transportation industry. We should note that our interest in the transportation industry is solely as a case study of the effects of economic regulation. It is not intended to be a comprehensive review of the transportation industry.¹

Transportation Industry

Simply stated, the service provided by a transport firm is the physical movement of a good from one point in geographic space to a second point in geographic space. In attempting to define the transportation industry, it becomes immediately apparent that a wide array of markets is encompassed. Taxicabs transport travelers and small packages within metropolitan area. Airlines move travelers and small packages (as well as some larger ones) across metropolitan areas. Railroads transport large loads of coal and grain across long distances. Although they are confined to transporting a small number of raw materials like natural gas and oil, pipelines also provide transportation services. Even intercity telecommunication companies provide transportation services of a sort in that they move information between local exchanges. All these firms provide some form of transportation service, but it should be obvious that they do not all serve the same market. When it comes to the transportation of manufactured goods from the plant to the wholesaler, trucks and railroads can effectively compete, though taxicabs cannot. Airlines, railroads, and buses compete to transport travelers long distance, but pipelines are hardly capable of providing such a service.² The transportation industry, broadly defined, comprises many varied markets.

¹. For the reader interested more generally in the economics of transportation as opposed to just the effects of regulation on the transportation industry, see Clifford Winston, “Conceptual Developments in the Economics of Transportation: An Interpretive Survey,” *Journal of Economic Literature* 23 (March 1985): 57–94, and the references cited within.

². A fictional exception took place in the 1987 James Bond movie, *The Living Daylights*. A Russian defector is transported across the Iron Curtain in a capsule propelled through a pipeline.
We can define a market in the transportation industry as comprising those consumers who demand a particular type of product to be transported from one geographic location to a second geographic location and those firms that can effectively compete to provide that service. It is important to note that firms can offer services that are quite good substitutes for one another yet use very dissimilar technologies in providing those services. In transporting travelers between geographic locations, airlines and railroads can, in some markets, be quite effective substitutes, for example, in the Washington–Philadelphia market. However, they use very different technologies. In light of this definition, a market would then be, for example, the transportation of travelers from Boston to Dallas or the transportation of steel from Pittsburgh to Chicago. How it is done is not important as long as consumers perceive firms as providing reasonable substitutes for one another’s services.

For the purpose of exploring the effects of economic regulation, our interest is not so much in any particular market but rather in classes of markets. By a class of markets we are referring to markets that have some essential properties in common, specifically, those properties that are essential in assessing the impact of regulation. One such property is the distance over which a good is transported. It is important to differentiate between local transportation (for example, within a metropolitan area) and long-distance transportation (across metropolitan areas). A second property is the type of good being transported. At a minimum we need to differentiate between the transportation of passengers and of freight. Freight can be partitioned into bulk goods and nonbulk goods, where bulk goods include many raw materials like coal, grain, and oil, whereas nonbulk goods include many manufactured goods.

The transportation industries we will examine in this chapter include long-distance freight and long-distance passenger. With regard to long-distance freight transportation, the main suppliers include railroads, trucks, water barges, pipelines, and airlines (see table 17.1). Our concern will be with goods that are generally too large for airlines to be an effective competitor. We will also not be concerned with goods that can be moved by pipelines. As a result, the relevant competitors are railroads, trucks, and water barges. Since the impact of regulation has been most strongly felt in the railroad and trucking industries, we will concentrate on surface freight transportation and ignore water barges.

With respect to long-distance passenger travel, the most important mode of transportation is airlines. Railroads and passenger buses can be adequate substitutes in some markets, but there are many markets in which they are not effective competitors to airlines because of the distance being traveled and consumers’ valuation of time. In those markets, there is no adequate substitute for airline travel. Hence, it is not too severe a restriction to focus solely on airlines in the provision of long-distance passenger travel.

These two transportation subindustries, surface freight and airlines, are common in that they have a long history of economic regulation. In addition, both were deregulated in the last few decades and, as a result, offer informative case studies of the impact of economic regulation.
In this section we will explore the regulation of the railroad and trucking industries in the United States. In assessing the effects of the regulation of surface freight transportation, it is important to consider the cross-effects between the regulation of one mode of transportation and other modes. One would expect the regulation of rail rates to affect the supply and demand not only for rail services but also for trucking services, since the two are competitors in some markets. In light of this fact, we will consider both modes simultaneously.

### Regulatory History

To understand the historical roots of regulation in the transportation industry, one has to start with the railroad industry in the second half of the nineteenth century. At that time railroads were the predominant form of long-distance transportation, whether passenger or freight. The most interesting feature of the railroad industry in the 1870s and 1880s was the volatile movements in rail rates, with many episodes of aggressive price wars. In response, the railroads attempted to coordinate their pricing decisions in order to stabilize prices at profitable levels. This effort led to the formation of the Joint Executive Committee (JEC) in 1879. As it turned

---

### Table 17.1

<table>
<thead>
<tr>
<th>Year</th>
<th>Rail</th>
<th>Truck</th>
<th>Water</th>
<th>Pipeline</th>
<th>Air</th>
</tr>
</thead>
<tbody>
<tr>
<td>1929</td>
<td>74.9</td>
<td>3.3</td>
<td>17.4</td>
<td>4.4</td>
<td>0.0</td>
</tr>
<tr>
<td>1939</td>
<td>62.4</td>
<td>9.7</td>
<td>17.7</td>
<td>10.2</td>
<td>0.0</td>
</tr>
<tr>
<td>1944</td>
<td>68.6</td>
<td>5.4</td>
<td>13.8</td>
<td>12.2</td>
<td>0.0</td>
</tr>
<tr>
<td>1950</td>
<td>56.2</td>
<td>16.3</td>
<td>15.4</td>
<td>12.1</td>
<td>0.0</td>
</tr>
<tr>
<td>1960</td>
<td>44.1</td>
<td>21.7</td>
<td>16.9</td>
<td>17.4</td>
<td>0.0</td>
</tr>
<tr>
<td>1970</td>
<td>39.8</td>
<td>21.3</td>
<td>16.4</td>
<td>22.3</td>
<td>0.2</td>
</tr>
<tr>
<td>1980</td>
<td>37.5</td>
<td>22.3</td>
<td>16.4</td>
<td>23.6</td>
<td>0.2</td>
</tr>
<tr>
<td>1988</td>
<td>37.0</td>
<td>25.2</td>
<td>15.5</td>
<td>21.9</td>
<td>0.3</td>
</tr>
<tr>
<td>1996</td>
<td>40.0</td>
<td>27.7</td>
<td>14.2</td>
<td>17.7</td>
<td>0.4</td>
</tr>
</tbody>
</table>

* Includes both for-hire and private carriers.


---

out, the JEC was only mildly effective in keeping price above cost, as there were episodes of price wars in 1881, 1884, and 1885.\(^4\)

**Interstate Commerce Act of 1887**

By the mid- to late 1880s the railroads came to the realization that for rates to be stabilized at profitable levels, a more effective authority than the JEC would be required. The JEC was replaced by the federal government with the Interstate Commerce Act of 1887. This act established the Interstate Commerce Commission (ICC) for the purpose of regulating the railroads. It was the job of the ICC to see that rail rates were “reasonable and just,” that higher rates were not charged for short hauls than for long hauls, and that the railroads did not discriminate among persons or shippers.

Only with additional legislation did the ICC actually acquire the requisite powers for controlling the railroad industry. The Hepburn Act of 1906 gave the ICC the power to set maximum rates, whereas the Transportation Act of 1920 allowed the ICC to set minimum rates and to control the entry and exit of firms from rail routes. The power to control the range over which firms could set rates, as well as the power to make entry and exit decisions, was extensively used by the ICC throughout the period of railroad regulation.

Until the 1920s the ICC had no difficulty in ensuring that the railroads earned at least a fair rate of return. However, in that decade, both trucks in the surface freight market and buses in the passenger market arose as vigorous competitors. Partly because of lobbying pressure from the railroads, the Motor Carrier Act of 1935 was passed, which brought motor carriers under ICC regulatory control. In addition, the Transportation Act of 1940 placed certain water barge transportation within the domain of the ICC. As in the case of railroads, the ICC controlled both rate setting and entry into and exit out of markets served by regulated motor carriers and water barges.

**Path to Deregulation**

Rather interestingly, it was not long after this spurt of additional regulation that the railroads began lobbying for less ICC control, at least of railroads. When the Motor Carrier Act was passed, poor road conditions and the limited size of trucks prevented the diversion of much rail traffic. This situation changed drastically in the 1950s with the development of the interstate highway system and the presence of an unregulated trucking sector comprising owner-operators, who carried exempt commodities, and manufacturers and wholesalers providing their own freight transportation.\(^5\) Railroads found that ICC regulations made it increasingly

---


difficult to respond to this increased competition from alternative modes of transportation. All this led the industry to lobby for increased flexibility in setting rail rates. What ultimately came to pass was the Transportation Act of 1958. As a result of this legislation, the ICC approved some of the lower rates requested by the railroads and allowed railroads to discontinue passenger service in some markets that were considered unprofitable. However, the ICC turned down many other rate-change requests, and the railroads still demanded increased rate flexibility. By the 1970s the railroads were demanding increases because of rising fuel costs as result of the oil price shocks. The bottom line is that by the mid-1970s, the railroad industry was lobbying hard for reduced regulatory control.

Spurred on by lobbying pressure from the railroads and the bankruptcy of Penn Central, the Railroad Revitalization and Regulatory Reform Act of 1976 (4R Act) was passed. The 4R Act set up a “zone of reasonableness” within which railroads could adjust rates, with the exception that if railroads had “market dominance” over a certain route, the ICC could maintain its strict control. Using an encompassing definition of “market dominance,” the ICC maintained considerable control over rail rates. A second important aspect of the 4R Act was to give railroads increased freedom in abandoning unprofitable routes. Although the 4R Act was the first step in deregulation, it did very little to relieve the pressures for reduced government control.

Quite in contrast to the case of railroads, the ICC took it on itself to deregulate the trucking industry. As early as 1975, the ICC made entry into trucking routes easier. Deregulation was moving at quite a strong pace by the late 1970s, as the ICC had reduced restrictions over entry into routes and reduced the power of rate bureaus to establish rates, thus allowing more rate freedom for trucking firms.

**Deregulation**

The major pieces of legislation mandating deregulation came in 1980 with the Staggers Rail Act and the Motor Carrier Act of 1980. The Staggers Act overturned much of the Interstate Commerce Act of 1887 by giving railroads considerable freedom in setting rates except where there is “market dominance” (which was not to be so often appealed to by the ICC), and in allowing freedom of entry and exit. The Motor Carrier Act of 1980 put into law much of the deregulation that the ICC had pursued in the late 1970s, though it did limit or rescind some ICC regulatory reforms. Nevertheless, after 1980 the surface freight transportation market was largely deregulated. Firms were free to compete and to move in and out of markets.

**Why Was There Regulation?**

It would appear that the formation of the ICC was a response to the inability of the railroad industry to maintain stable prices at profitable levels. One explanation for such price instability is that the railroads were attempting to keep rail rates artificially high so as to reap above-normal profits. If that is true, then price wars were not indicative of “destructive
competition” but rather of collusive pricing, which created the incentive for a firm to undercut the agreed-on price for the purpose of raising short-run profits (though at the cost of future profits, since it was likely to induce a price war). Under this view, the ICC’s role is as a cartel rate-setter, which is clearly not in society’s best interests. Another explanation is a natural monopoly argument that can be made for railroad regulation. An examination of the production technology suggests that average cost might have been declining in output. There are several components of cost that do not rise proportionately with traffic volume, including right-of-way, the cost of track, and certain equipment like locomotive power and train stations. If marginal cost lies significantly below average cost and competition leads to marginal cost pricing, then firms will earn below-normal profits when they are unable to coordinate their pricing decisions. In that case, there is an economic rationale for regulation.

Regardless of the rationale, empirical evidence reveals that financial markets expected the profitability of the railroads to improve with regulation. Robin Prager examined movements in the stock prices of railroad firms in response to events surrounding the passage of the Interstate Commerce Act. Using monthly stock-price data from January 1883 to December 1887, the study revealed that the members of the JEC earned excess returns of 13.4 percent in the month in which the Senate passed the Cullom bill, which was the Senate’s version of the Interstate Commerce Act. Non-JEC members earned even higher returns.

In contrast, there would appear to be no natural monopoly argument that one can make for the trucking industry. Economies of scale would appear to be exhausted at relatively low rates of production. The most plausible hypothesis for why trucking was regulated is that the presence of an unregulated trucking sector made it difficult for the ICC to effectively regulate the railroads. Given that the railroads and truckers compete in many markets, it would be difficult to achieve a particular outcome for the railroad industry as long as the ICC could only control what the railroad companies did.

The regulation of surface freight transportation appears to be consistent with the economic theory of regulation (see chapter 10). The railroads formed a small and well-organized group that recognized the potential gains from price stability and how price and entry regulation would achieve that stability. Although the benefits of regulation were concentrated among a few firms, the cost of regulation was spread out across the many users of rail services. It is then not surprising that regulation should have passed that favored the railroads. It also appears that the ICC was “captured” by the railroads in that the regulation of the trucking industry was apparently a response to the competitive pressures felt by railroads. Of course, eventually all this was to change. Ultimately, regulation came to reduce the profitability of the railroad companies and increase the profitability of truckers, as reflected in their respective lobbying practices.

Description of Regulatory Practices

Although regulation encompasses a wide array of restrictions, the most important restrictions are generally those placed on price and on entry into and exit from markets. A reading of the relevant legislation suggests that railroads and trucking were subject to similar ICC control. In both industries, rate changes had to be requested and approved by the ICC and, in order to operate in a particular market, a certificate of convenience from the ICC was required. Exit from a market also required ICC approval. Despite the similarities in their respective pieces of legislation, in practice, ICC control of price and entry and exit was quite different between railroads and trucking.

Price Regulation

In railroads, the ICC was very active in exercising its control over the setting of maximum and minimum rates. Its pricing practice was dictated by two basic principles: value-of-service pricing and equalizing discrimination. Value-of-service pricing entails charging higher rates for higher-valued commodities, regardless of whether there is a cost difference in transporting goods of different value. Because of value-of-service pricing, for many decades the railroads were forced to set higher rates for manufactured goods than for raw materials and agricultural products. The principle of equalizing discrimination is that rates should not discriminate between shippers or between the size of the shipment, despite the fact that cost may vary across shippers (for example, some ocean ports are more costly to transport to) as well as the size of the shipment (typically, smaller shipments have a higher per-ton cost than larger shipments). The principle of equalizing discrimination meant that cost played a subordinate role in determining rail rates and typically resulted in rates for manufactured goods subsidizing rates for nonmanufactured goods. This practice is referred to as cross-subsidization (see chapter 16).

The setting of rates in the trucking industry was quite different from the practice observed in the railroad industry. To begin, the ICC allowed the trucking industry to establish rate-making bureaus that were exempt from antitrust prosecution by the Reed-Bulwinkle Act of 1948. Typically, the ICC automatically approved rate changes made by these bureaus unless they were protested by a shipper or a trucking firm. However, like the railroads, trucking rates were not allowed to be selectively cut. In particular, rates could not vary across routes of similar distances, even though the density of the routes might vary and higher density results in lower unit cost, ceteris paribus. The density of the route refers to the volume of goods transported per mile. As a result, rates on high-density routes subsidized rates on low-density routes. In contrast to railroads, rates were allowed to vary across other dimensions that affect cost, including the size of the shipment and the characteristics of the shipper.

The hypothesis that rail rates were set to equalize discrimination whereas truck rates were set in accordance with cartel pricing has been tested. The empirical method used was to
determine how freight rates varied with demand elasticities and costs of service. To understand how freight rates vary with cost, depending on whether rates are set to equalize discrimination or maximize industry profit, let us suppose marginal cost is related to shipment size according to the following formula: 

\[ MC = \frac{12}{1 + S} \]

where \( S \) is the shipment size. Thus, higher shipment size reduces marginal cost. For example, if the shipment size is 1, then marginal cost is 6, while if shipment size is doubled to 2, then marginal cost is reduced by one-third, to 4. Equalizing discrimination would imply the same price being charged for all shipment sizes, say a per-unit price of 8, so that a shipment of size \( S \) would cost \( 8 \cdot S \) to ship.

To consider joint profit maximization, let \( P^*(S) \) denote the profit-maximizing price (per unit being shipped) for a shipment of size \( S \). Note that the total required payment for a shipment of size \( S \) is then \( P^*(S) \cdot S \). Let \( \eta \) denote the absolute value of the elasticity of demand with respect to price. Recall that it measures how responsive demand is to a change in price. One can show that if price is set to maximize industry profit, then \( P^*(S) \) must equal \( MC \cdot \left[ \frac{\eta}{\eta - 1} \right] \). Substituting our formula for marginal cost, one derives \( P^*(S) = \left[ \frac{12}{1 + S} \right] \left[ \frac{\eta}{\eta - 1} \right] \), which implies that price is decreasing in \( S \). Hence, if price was set so as to maximize industry profit, then the per-unit price should be lower for larger shipments. Intuitively, because the profit-maximizing price must equate marginal revenue and marginal cost, if marginal cost is lower for larger shipments, then price must be set so that marginal revenue is lower, and doing so requires a lower price (assuming marginal revenue is declining in price). For example, if demand elasticity is constant at a value of 2, then the profit-maximizing price equals \( \frac{24}{1 + S} \). In that case, if the shipment size is 1, then a profit-maximizing cartel would set a per-unit price of 12, whereas if the shipment is of size 2, then a per-unit price of 8 would be charged.

While a profit-maximizing cartel sets a lower per-unit price for larger shipments, this is not the case under equalizing discrimination, where, for our example, a per-unit price of 8 is charged regardless of shipment size. Also note that industry profit maximization implies that the per-unit price is higher for submarkets for which demand is less responsive to a change in price (that is, demand is more inelastic, which is represented by a smaller value for \( \eta \)). No such relationship is observed under equalizing discrimination.

Using 1972 rail and truck rates (the rate being cents per hundredweight shipping charges), the dependence of these rates on cost and demand factors was estimated.7 The proxy for cost of service was the size of the shipment, because marginal cost is typically less the larger the shipment (as we assumed in our example). The empirical analysis revealed that rail rates were unaffected by shipment size, whereas truck rates were lower for larger shipments. It was also

found that truck rates were higher the more inelastic the demand. These empirical findings are supportive of the hypothesis that rail rates were set so as to equalize discrimination, whereas truck rates were set in line with maximizing industry profits. It is interesting to note that the railroad industry was in support of deregulation, whereas the trucking industry was not. This difference would make sense if truck rates were set in a profit-maximizing manner and rail rates were not.

**Entry and Exit Regulation**

Although both the railroad and trucking industries required the approval of the ICC to enter and exit markets, the constraint varied between the two industries because of the ratemaking practices described in the preceding paragraphs. From the railroads’ perspective, the effective constraint was with regard to exiting rather than entering markets. The policy of equalizing discrimination resulted in cross-subsidization, so that railroads were operating in some markets at a loss. The ICC was adamant that the railroads not abandon these markets. In contrast, the pricing policy for the trucking industry permitted considerable profits to be earned. This profit would normally be dissipated through entry (as market demand gets divided among more firms), but the ICC prevented any such entry. Specifically, a petition for entry into a particular route required the petitioner to establish that demand could not be effectively met by existing suppliers. This requirement clearly placed the burden of proof on the firm seeking entry, and in fact most petitions for entry were turned down by the ICC. Generally, the only way to enter was to purchase the operating license of another trucking firm. Of course, this did not expand the number of competitors, but only changed the identities of the firms. It is important to note that the ICC limited entry into routes as well as into the industry, so that even existing motor carriers found it difficult to enter routes that were currently being served.

**Effects of Regulation**

As described in chapter 16, economic regulation can induce welfare losses through both allocative and productive inefficiencies. Allocative inefficiency is the misallocation of resources among different goods. It typically results from prices deviating from marginal cost. If the price of a good exceeds its marginal cost, we know that a suboptimal amount of that good is produced, ceteris paribus. It is also true that if two goods, say $X$ and $Y$, are substitutes and the price of $X$ is above its marginal cost, it may be optimal to have the price of $Y$ above its marginal cost as well, by the theory of second best. Therefore, in considering welfare losses from price regulation, one is concerned not only with the relation between rail rates and the cost of rail service but also with the relation between rail rates and truck rates, inasmuch as they are substitutes in some markets. If rail rates are priced high relative to truck rates, then there are apt to be greater truck services supplied than is socially optimal. This
statement is true even if truck rates are above cost, for the service is more efficiently supplied by railroads. Productive inefficiencies occur when inputs are not effectively used in production. Likely culprits are distorted input prices that result in a suboptimal input mix and wasted inputs resulting from the lack of competitive pressure. There are also dynamic sources of productive inefficiencies, including the stifling of the discovery and adoption of innovations and distorted investment decisions. The ensuing analysis will consider both allocative and productive inefficiencies created by regulation.

**Price and Quality of Service**

To begin, let us assess the regulatory impact on surface freight rates. Our initial approach is to compare rates during the period of regulation with rates after deregulation. This approach is intertemporal in that it involves analyzing prices over time. Presented in figure 17.1 is the time path of average real revenue per ton-mile of freight for rail services, while figure 17.2 shows the same variable for motor carrier services. (Except where noted, all dollar figures in this chapter are in 1985 dollars.) These rates are calculated by taking a weighted average of the actual rates charged. It is clear from figure 17.1 that rail rates declined after the Staggers Act. Real rail rates fell by over 12 percent between 1981 and 1985. Recall that the deregulation of motor carrier rates began prior to the Motor Carrier Act of 1980, as the ICC deregulated rates beginning in the late 1970s. During the period of the late 1970s, figure 17.2 shows that motor carrier rates did indeed decline. Interestingly, however, rates rose after the Motor Carrier Act of 1980, having gone up over 5 percent between 1981 and 1985.

Prior to drawing any conclusions about the effect of regulation on rail and truck rates from figures 17.1 and 17.2, we should carefully consider the data at hand. In using an

![Figure 17.1](image_url)

**Figure 17.1**

Average Real Revenue per Ton-Mile for Railroads

intertemporal approach, it is important to recognize that there are likely to be many important variables that can influence rates and that they may be changing over the sample period. Deregulation is not the only event that would be expected to affect average revenue per ton-mile. There are at least two other important changes over this period worth taking account of. One is the recession that took place in 1981–1983. By shifting in the demand curve for transportation services, the recession would be expected to decrease average revenue per ton-mile in that firms would be chasing fewer consumers, and this would intensify competition and lower rates. But the recession is probably not wholly responsible for the movement in surface freight rates, because rail rates continued to fall through the recovery of 1984 and 1985. Trucking rates did indeed rise in 1983, but they declined in 1984, only to rise again in 1985.

A second factor confounding the effect of deregulation on surface freight rates is one that is endogenous to deregulation: the composition of traffic. Even if actual rates remained the same, average revenue per ton-mile could change if the types of commodities being transported were to change over the sample period. For example, if railroads switched to transporting more bulk goods, which have a lower rate than other goods, then average revenue would decline even if rates remained constant over time. An empirical analysis partially controlled for changes in the composition of traffic in order to assess the effect of deregulation on rail rates.8 The proxy for the change in traffic composition used is the average weight of freight trains, as such a variable would be expected to rise if railroads transported more bulk

commodities. Using the rail rates in figure 17.1, it was found that 90 percent of the change in rates over the period 1971–1985 was due to the change in the average weight of freight trains. This finding suggests that deregulation is apt to have had a large impact on the mix of commodities being transported by railroads. After controlling for changes in the average weight of freight trains, it was found that rail rates were actually higher by 2 percent in the deregulated period 1980–1985.9

Examining the overall effect of deregulation on railroad services can cover up many important facts because there are good reasons to expect deregulation to have affected rail rates in different ways for different commodities. To begin, the price in any market depends on two factors, cost and markup (of price over cost). Deregulation can affect cost by, for example, reducing the constraints placed on railroads and allowing them to adopt more efficacious practices. However, the exact change in cost depends on the traits of the commodity—for example, is it manufactured or bulk?—and the route—for example, is it long distance? Turning to the issue of markups, regulation almost certainly resulted in markups varying across markets as a result of the practice of equalizing discrimination. Markets with relatively low markups under regulation would be expected to experience more of an increase in price. Furthermore, the postderegulation markup of a railroad should depend on the intensity of competition in that market from other railroads and competing modes such as trucking and water barges. Because the degree of competition varies across markets, one would expect the postderegulation markup to vary as well. Hence, even if two markets had the same cost before and after deregulation and the same markup under regulation, they may have different prices under deregulation because of differences in the degree of competition.

Recognizing that the effect of deregulation may vary across commodity markets, a study estimated the impact of deregulation on rail rates for thirty-four different commodity classifications.10 These included farm products, coal, apparel, and machinery, among others. It estimated both the immediate impact of deregulation and its impact over time. Based on the estimates of the immediate impact, regulation kept rail rates artificially low in twenty-two of the thirty-four commodity markets (and this effect was statistically significant for nine of those twenty-two markets). For example, deregulation resulted in a 13 percent rise in rail rates for forest products and a 5 percent rise for coal. For the other twelve commodity markets, regulation had kept rates artificially high (and this effect was statistically significant in four of the twelve). For example, rates for transporting farm products by railroad fell 8.5 percent

9. There are, however, studies which find that deregulation caused average rail rates to fall. For example, a 16.5–18.5 percent reduction is found by Christopher C. Barnekov and Andrew N. Kleit, “The Efficiency Effects of Railroad Deregulation in the United States,” *International Journal of Transport Economics* 17 (1990): 21–36. One reason for such variation in findings is that studies differ in how they implicitly weight rail rates for different commodities.

upon deregulation. Though the initial impact of deregulation on rail rates varied across commodities, the cumulative effect by 1988 was much more uniform in that not a single market had experienced a statistically significant increase in rates as a result of deregulation. More specifically, rail rates were statistically significantly lower in twenty of the thirty-four markets, and in the other fourteen markets deregulation did not have a statistically significant effect on rates. Weighting these price changes by the size of the market (as measured by ton-miles), the cumulative effect of deregulation on average rail rates was a decline of 30 percent by 1988, which is to be contrasted with an initial rise of 10 percent. It is conjectured that the fall in rail rates can be attributed to cost savings emanating from deregulation. Later evidence shows that regulation did reduce the rate of productivity growth, a finding that is supportive of this hypothesis.

With respect to motor carrier rates, figure 17.2 fails to provide a clear picture of the impact of regulation. More disaggregated data show that regulation kept motor carrier rates too high. One piece of evidence dates from the mid-1950s, when the ICC exempted poultry and fruits and vegetables from the regulation of motor carrier rates. In response to this exemption, motor carrier rates fell 19 percent for fruits and vegetables and 33 percent for poultry. More recent evidence comes from surveys of shippers (who are demanders of motor carrier service) taken after passage of the Motor Carrier Act of 1980. A survey of some thirty-five shippers revealed that truckload rates fell by 25 percent from 1975 to 1982, and less-than-truckload rates fell by 11 percent over the same period. A larger survey of 2,200 shippers of manufactured goods taken shortly after deregulation found that 65 percent of those polled said that truck rates were lower. In contrast, only 23 percent found rail rates to be lower. Finally, one study examined rates charged by sixty-one motor carriers of general freight from 1975 to 1985. It found that deregulation had lowered rates by 15 to 20 percent by 1983 and by 25 to 35 percent by 1985. The growing effect of deregulation on motor carrier rates may be due to increased productivity growth. Such growth will be reviewed later.

Additional evidence comes from the state level. Around the same time as the Motor Carrier Act of 1980, a number of states also deregulated intrastate trucking. In fact, the first full deregulation in the U.S. transportation industry occurred when Florida deregulated trucking on July 1, 1980. State regulation was quite comparable to that on the federal level, in terms of both pricing and entry policy. One study examined the effect of deregulation in Florida and in Arizona, which was deregulated in July 1982. Controlling for several factors that

12. Both survey data are from Moore, “Rail and Trucking Deregulation.”
Influence rates, including the commodity class, the shipment size, and the type of motor carrier, changes in motor carrier rates were examined for Arizona from January 1980 to October 1984 and for Florida from January 1979 to October 1984. It was found that deregulation caused average intrastate motor carrier rates to fall for half of the routes in Arizona and all of the routes in Florida. A second study focused on motor carrier rates for Florida for the more limited time period of June 1980 to September 1982. It found that average rates fell by almost 15 percent.

Interpreting the evidence on the effect of regulation on surface freight rates is difficult because there are so many different commodities and so many different routes. Nevertheless, the available evidence is supportive of the hypothesis that regulation kept rail rates too low, on average, though too high for some commodities, and kept motor carrier rates too high. To learn more about the effects of regulation, let us analyze what happened to traffic composition. The evidence here shows that the transportation of manufactured goods shifted away from railroads to trucks. In contrast, for bulk goods, in particular fruits and vegetables, railroads increased their share while trucking reduced its share. A sample of shares of manufactured commodities in table 17.2 reveals that deregulation reinforced the trend in which railroads were providing a smaller market share of these services. On the other hand, as shown in table 17.3, the transportation of fruits and vegetables by railroads more than doubled from 1978 to 1982.

Having recognized the change in traffic composition, the data in figures 17.1 and 17.2 are now more understandable. On the one hand, average rail revenue per ton-mile declined even though some rates increased because railroads have moved to transporting bulk commodities, which have lower rates. On the other hand, motor-carrier rates fell in spite of average revenue per ton-mile rising over 1981–1985, because they moved away from transporting goods with lower rates.


<table>
<thead>
<tr>
<th>Year</th>
<th>Rail</th>
<th>For-Hire Motor Carrier</th>
<th>Private Motor Carrier</th>
</tr>
</thead>
<tbody>
<tr>
<td>1963</td>
<td>32.8</td>
<td>46.6</td>
<td>16.5</td>
</tr>
<tr>
<td>1967</td>
<td>34.0</td>
<td>45.3</td>
<td>18.0</td>
</tr>
<tr>
<td>1972</td>
<td>30.8</td>
<td>45.1</td>
<td>20.6</td>
</tr>
<tr>
<td>1977</td>
<td>22.9</td>
<td>43.0</td>
<td>28.6</td>
</tr>
<tr>
<td>1983</td>
<td>16.0</td>
<td>47.4</td>
<td>26.0</td>
</tr>
</tbody>
</table>

It is quite apparent that regulation distorted rail and motor carrier rates. Though calculating the amount of welfare loss from allocative inefficiency is an extremely difficult task, many economists have tackled this important problem. Although the estimates are so far-ranging as to question their usefulness, most studies from the 1980s fall in the range of $1 to $1.5 billion per year. One of the more recent studies uses counterfactual analysis and estimates a much higher number. Using shipper, carrier, and labor behavior during the deregulated year of 1985, it estimated what 1977 would have been like if the industry had been deregulated in that year. The simulated data for an “unregulated” 1977 were then compared with the actual (regulated) data for 1977. Looking at the effect on the welfare of shippers, who are the consumers of transportation services, it was estimated (in 1977 dollars) that the deregulation of motor carrier rates increased their welfare by almost $4 billion per year. Although the deregulation of rail rates was found to reduce grain rates, thereby raising shippers’ welfare by about $280 million per year, it raised all other rates (on average), and these increases reduced welfare by $1.35 billion. Thus, shippers’ welfare went up by $2.89 billion annually as a result of the change in surface freight rates from deregulation. Of course, this measure does not net out the transfer from shippers to carriers.


17. Winston et al., “Economic Effects.”

18. Winston et al. in “Economic Effects” (1990) also consider some indirect effects of rate deregulation on service, in particular, rail transit time, and find very large welfare gains even after netting out the loss in carriers’ profits and labor’s wages. Large estimates of the welfare gain from deregulation are also found by Barnekov and Kleit, “Efficiency Effects.”
Static Productive Inefficiency

One important source of productive inefficiency is restrictions on entry and exit. Entry restrictions prevent more efficient firms from replacing less efficient firms, whereas exit restrictions can keep firms producing in markets that are not socially efficient to serve. Let us initially consider the effect of entry and exit restrictions in the surface freight transportation industry.

As described in the description of regulatory practices, the ICC was influential in affecting the exit decisions of firms in the railroad industry. The evidence on rail rates shows that the ICC kept rates too low in some markets and, by prohibiting exit, forced railroads to serve unprofitable markets. In response to the Staggers Act giving firms the freedom to exit markets, railroads have abandoned many routes. A notable example is Conrail, which immediately abandoned 2,600 route miles after the Staggers Act. This amount represented 15 percent of total track miles for Conrail, but it generated only 1 percent of revenue. It has been estimated that to provide rail services at 1969 levels at minimum cost would have required only 20 to 25 percent of existing capacity. The annual cost saving from reduced capacity was estimated to be on the order of $750 million to $1.5 billion.

In contrast, the binding restriction on the motor carrier industry was the limitation of entry. Because motor carrier rates were above cost, trucking firms were earning above-normal profits. The prohibition on entry prevented these profits from being competed away. The response to the open-door policy for prospective firms brought about by the Motor Carrier Act of 1980 was remarkable. From 1978 to 1985, the number of ICC-certified motor carriers doubled, from 16,874 to 33,823, and it exceeded 40,000 by 1990. Nearly 17,000 companies entered between 1978 and 1985; over 6,000 failed. As shown in figure 17.3, bankruptcies rose dramatically in the early 1980s because of deregulation and the recession.

Besides maintaining inefficient firms, ICC regulation of the motor carrier industry resulted in productive inefficiency by raising wages. Because trucking firms earned above-normal profits, the Teamsters Union was able to extract some of these rents through higher wages. Using data from 1973–1978, it was estimated that a wage premium of 50 percent was paid to union workers in the trucking industry; that is, union workers earned a wage 50 percent higher than nonunion workers performing comparable work with comparable skills. In

contrast, from 1979 to 1985 (in the absence of regulation) the union premium was only 27 percent, which is very close to the national average of 28 percent. By opening these markets to competition, laborers in trucking realized that they would have to settle for lower wages or watch workers being laid off as firms went bankrupt. Labor as an input was priced too high under regulation, which led to a reduced supply of transportation services and a socially inefficient input mix biased toward other inputs.

There are also many idiosyncratic sources of productive inefficiencies from the regulation of surface freight regulation. Let us briefly mention two examples. With respect to the motor carrier industry, the ICC required truckers to charge the same rates on backhauls, even though this rule resulted in many of them being empty. It would have been both welfare-improving and profit-improving to allow them to charge lower rates because the opportunity cost of providing the service is lower (as the trucks are traveling to their next

Figure 17.3
Bankruptcies among Trucking Firms
*Source:* Data from Dun and Bradstreet’s *Business Failure Record.*

haul anyway). It was estimated that the annual welfare loss due to empty backhauls was over $300 million.\textsuperscript{23} Since deregulation, empty backhauls have fallen from 28 percent to 20 percent.\textsuperscript{24}

Finally, did you know that the regulation of rail rates reduced the competitiveness of the flour market? The higher cost of transporting flour under regulation reduced the geographic size of flour markets. Because each market had fewer firms competing, railroad regulation reduced the intensity of price competition in flour markets.\textsuperscript{25}

### Dynamic Productive Inefficiency

By requiring railroads to serve unprofitable markets, regulation restricted the ability of railroads to finance investment. By the late 1970s, $15 billion of investment on track maintenance had been deferred or postponed. With deregulation and the ensuing ability to profitably adjust rates to market conditions, there was a tremendous increase in investment. During 1981–1985, $27 billion was spent on railroad structures, roadways, and maintenance of way, while $30 billion was invested in rail cars, locomotives, and other equipment.\textsuperscript{26}

A second source of dynamic inefficiency is the stifling of innovations. If regulation restricts railroads in their utilization of innovations (for example, through proper pricing), fewer innovations will be adopted. It is very difficult to directly measure how regulation affects the discovery and adoption of innovations. However, we can estimate what productivity growth would have been in the absence of regulation. Of course, this estimate will encompass all sources of reduced productivity growth, including reduced innovations and reduced investment.

For this purpose, a cross-country analysis was performed that compared the growth in total productivity for U.S. railroads over 1956–1974 with the growth achieved for Canadian railroads during the same time period. Growth in total productivity of inputs is equivalent to a decline in cost per unit of output. Although both industries had access to the same innovations, the Canadian railroads were subject to much less regulation than U.S. railroads. As shown in table 17.4, for 1956–1974 total productivity growth in the railroad industry was 3.3 percent in Canada, but only 0.5 percent in the United States. Since there is a large number of small and relatively weak railroads in the United States, the study sought to control for differences in U.S. and Canadian railroads by selecting a comparable subsample. Those chosen were Canadian National and Canadian Pacific for Canada and Atchison, Topeka and


\textsuperscript{24} Moore, “Unfinished Business.”


Santa Fe and Southern Pacific for the United States. At least for the period 1963–1974, this subsample supports the hypothesis that regulation reduced productivity growth. If U.S. railroads had experienced the growth in productivity that Canadian railroads had, it was estimated that the cost of providing rail services in 1974 would have been $13.83 billion lower (in 1985 dollars). This is striking evidence of the productive inefficiencies brought about by the regulation of the railroad industry.

The most recent studies find that real revenue per ton-mile had decreased by 10 to 25 percent by 1997 (depending on the commodity being transported), and at least one-third of this decline is attributable to deregulation. It is argued that dynamic efficiency gains were crucial:

Railroad deregulation in the United States is more a story of cost reduction than of demonopolization. Surface freight deregulation as a whole surely increased competition, but the ensuing rate reductions were too large to be explained merely by the elimination of monopoly profits. Since most railroads earned far less than their cost of capital during the 1970s, there was little monopoly profit to redistribute to shippers. To simultaneously cut rates, improve service, and improve their financial performance, railroads had to cut costs. The experience of U.S. railroads thus illustrates the importance of dynamic or “productive” efficiency in creating improvements in economic welfare.

**Lessons from Regulation**

Summarizing many of the studies that estimated the welfare loss from the regulation of surface freight transportation, Robert Willig and William Baumol state:

Various studies estimated, for example, that between 1950 and 1980 more than a billion dollars a year was wasted in transporting freight by truck rather than by rail. Another billion dollars a year was wasted

---

in transporting freight on rail routes that were too long or were utilized with too little traffic density. Another $1.5 billion a year or more (in 1977 dollars) was wasted on unnecessary mileage traversed by empty cars, unnecessary demurrage-time between car unloadings and loadings, and circuitous loaded routings.29

There are several lessons to be learned from the regulation of surface freight transportation. First, once put in motion, regulation can be imperialistic. Beginning with the railroad industry, ICC control soon spread to motor carriers and water barges. There was no economic basis for regulating truckers except that it made for more effective regulation of the railroad industry. A second lesson is the potentially large welfare losses from product substitution. Much of the allocative inefficiencies were due not so much to rail and truck rates being different from marginal cost, but rather to rail and truck rates being set nonoptimally with respect to one another. As a result, traffic that would have been more efficiently carried by railroad—agricultural commodities, for example—was instead carried by trucks because rail rates were set too high. Similarly, there are examples of intermodal substitution going the other way. A third lesson is that reduced price flexibility brought on by regulation can have disastrous consequences. It seriously threatened the long-term profitability of the railroad industry.

With its regulatory jurisdiction evaporating, the ICC was abolished in 1995. The small number of remaining regulatory tasks were transferred to the newly created Surface Transportation Board in the Department of Transportation. Thomas Gale Moore marked this occasion by writing the ICC’s obituary, entitled “In Memoriam”:

Federal regulation of trucking which, during its illustrious career, provided a competition-free system, guaranteeing huge profits for owners and high wages for unionized workers, died January 1. Last summer, its cousin, state oversight of intrastate carriers, sustained a mortal blow as all controls on interstate firms hauling intrastate loads were abolished. Interstate Commerce Commission regulation was in its 60th year. Ravaged by the Motor Carrier Act of 1980, Washington’s supervision of motor carriers suffered a long enfeeblement, leading ultimately to its demise, despite the heroic efforts of union workers to preserve their benefactor. Unfortunately for the health of federal curbs, a combination of free market economists, shippers groups, and a surprising group of liberal politicians, including Senator Ted Kennedy and President Jimmy Carter, had undermined its support system and it died, not with a bang but a whimper.

Born at the height of the New Deal in 1935 from the unlikely marriage of federal and state regulators, large trucking interests, and railroads, ICC management of motor carriers evinced a long and successful career of prescribing prices, enjoining entry, and curtailing competition. By the early 1970s, Washington bureaucrats were forcing trucks to travel empty on return trips; to carry goods on circuitous routes, adding hundreds of miles to their transport; and to distinguish between carrying ordinary horses and those destined for the slaughterhouse.

During the nearly six decades of ICC rulemaking, the economy suffered hundreds of billions of dollars in waste, loss and abuse.

In addition to continued controls over household good carriers, federal regulation is survived by its adopted offspring, the International Brotherhood of Teamsters. No services are planned.\(^{30}\)

**Airlines**

Regulation has been a feature of the airline industry from almost its inception. Then, quite suddenly, things began to change in the late 1970s. An intensive program of deregulation was begun that ultimately led to the complete absence of price and entry controls. There is perhaps no industry for which deregulation has caused such radical changes. The issue is what exactly has changed and whether it has been for the better.\(^{31}\)

**Regulatory History**

The commercial airline industry began in the late 1920s with the hauling of mail for the U.S. Postal Service. Passenger service followed shortly thereafter in the early 1930s. Initially, the U.S. Postal Service had authority over mail rates, though this came under the realm of the Interstate Commerce Commission (ICC) with the Airmail Act of 1934. To allocate mail routes, the ICC set up a competitive bidding system whereby the airline that offered the lowest price per mile received a route franchise. In response to this system, the existing airlines pursued a strategy of submitting very low bids to ensure the retention of their routes, with the anticipation that they could raise rates afterward. As it turned out, the ICC did not permit such rate increases, so that many airlines found themselves on the verge of bankruptcy from having to provide mail service at a price considerably below cost. This episode is significant with regard to later airline regulation, as it was used as evidence by supporters of government regulation that an unregulated airline industry would be plagued by “destructive competition.” Regulation was seen by some as a necessary condition for the airline industry to develop into a stable and healthy segment of the transportation sector of the U.S. economy.

**Civil Aeronautics Act of 1938**

With the Civil Aeronautics Act of 1938, the airline industry came under federal regulation. This act created the Civil Aeronautics Authority, which two years later became the Civil


Aeronautics Board (CAB). The CAB was given control over the setting of maximum and minimum rates where the procedure for ratemaking was taken from the Interstate Commerce Act of 1887. In addition, the number of competitors was placed under the domain of the CAB. This action meant not only that entry into and exit out of the industry was subject to CAB approval but also that the CAB had the power to govern the route structures of firms. It could prevent an existing airline from entering a route or, alternatively, abandoning a route it was currently serving. In addition to control over price and market structure, the CAB was initially given responsibility for airline safety. This task was transferred to the Federal Aviation Administration (FAA) in 1958.

From its inception until the wheels of deregulation began to turn in the mid-1970s, the CAB was not shy about controlling air fares and the number of competitors. In a rather strong display of this power, the CAB implemented a route moratorium in the early 1970s. Due to excess capacity among airlines and declining industry profits, the CAB prohibited any firm from entering route markets. This policy kept new firms from entering the industry and existing firms from expanding the number of routes they served. The CAB feared that route expansion would intensify competition and have a negative impact on industry profit.

Path to Deregulation

By the mid-1970s, pressure was building to reform airline regulation. Academics had long argued that regulation stifled competition and generated considerable welfare losses.32 In 1975, Senate hearings held by Senator Edward Kennedy seriously explored the idea of regulatory reform. The Department of Transportation was also engaged in designing regulatory reform, and even the CAB strongly supported not just regulatory reform but full deregulation.

As CAB chairman, John Robson provided the initial step in deregulation by relaxing entry restrictions. He lifted the route moratorium and allowed entry into currently served markets for the first time since the 1960s. With Alfred Kahn’s arrival as the new CAB chairman in June 1977, the pace of deregulation accelerated as he further reduced entry restrictions and controls over fares. Major fare cutting followed.

Airline Deregulation Act of 1978

In response to these CAB reforms, fares were lower and industry profits were actually up in 1978. With such positive results from increased competition, Congress passed the Airline Deregulation Act (ADA) in 1978, which called for the phased deregulation of the airline industry. As prescribed by the ADA, the CAB’s authority over routes was to terminate on December 31, 1981, its authority over fares was to terminate on January 1, 1983, and, finally,

32. The first academic study to propose deregulation was Lucile S. Keyes, Federal Control of Entry into Transportation (Cambridge, Mass.: Harvard University Press, 1951).
its very existence was to terminate on January 1, 1985. The actual pace of deregulation turned out to be considerably faster than outlined in the ADA. Within a year of its enactment, airlines were free to serve any route. By May 1980 the CAB allowed unlimited downward flexibility in fares and considerable upward flexibility. Independent pricing was being strongly encouraged by the CAB. Even prior to January 1, 1983, the date for which CAB control over entry and fares was to end, airlines were competing in an unregulated environment.

Description of Regulatory Practices

The main objectives underlying CAB policies during the regulatory period were to keep the airline industry financially sound and to promote air service. Toward this end, the CAB extensively exercised control over price, the number of competitors, and route structures.

Price Regulation

The setting of fares by the CAB was characterized by four properties. First, fares were set with the intention of allowing airlines to earn a reasonable rate of return. Though this objective was an implicit feature of regulatory policy for a long time, it was not made explicit until 1960. In fact, before World War II, airfares were typically set at the first-class rail fare, railroads being the main competitor to airlines. In 1960 the General Passenger Fare Investigation set fares so as to allow the industry to earn a rate of return of 10.5 percent, whereas in 1970 the Domestic Passenger Fare Investigation (DPFI) was designed to generate an industry rate of return of 12 percent.

A second important feature of CAB pricing policy was that prices were generally set independent of cost. At least until 1970, the key criterion was profitability. Then the DPFI did introduce some role for cost in that it related fares to the length of the route, which is indeed related to cost of service. However, the CAB set fares above cost for routes exceeding 400 miles and below cost for routes less than 400 miles. This cross-subsidization from long-haul to short-haul markets was done primarily to promote air service on less dense routes.

A third property of CAB rate-setting was that fare changes were typically across the board rather than selective. This practice reduced fare flexibility and contributed to distortions in the fare structure. The fourth and final property was that the CAB strongly discouraged price competition. Requests for fare changes that were deemed as being a competitive response to other airlines’ fares were not treated well by the CAB. Presumably, the CAB feared that fare competition would make it more difficult for the industry to maintain profitability. Related to an earlier point, fare competition might have been interpreted by the CAB as leading to destructive competition.
Entry and Exit Regulation

With the introduction of regulation in 1938, sixteen airlines, referred to as trunk carriers, were “grandfathered” and became certificated carriers. From 1938 to 1978, the CAB did not allow the entry of a single new trunk carrier. In preventing such entry, it denied seventy-nine applications during 1950–1974. At the time of deregulation, only ten trunk carriers remained, six having disappeared through merger.

Whether a prospective firm wanted to enter the airline industry or an existing airline wanted to enter a currently served route, the CAB placed the burden of proof on the prospective entrant. The process by which a firm applied for entry into a route was long and expensive. Although the CAB did not allow entry that would directly compete in a major way with the trunk carriers, it did allow entry by local-service carriers, which picked up short-haul business abandoned (with approval of the CAB) by the trunk carriers.

Entry into routes by existing airlines was permitted to a limited degree, though we know by the route moratorium of 1969–1974 that the CAB could prevent that as well. Only 10 percent of applications by existing airlines to enter an existing route were approved from 1965 to 1974. In high-density markets, the CAB was more apt to allow entry, preferring an expansion in the number of competitors to a reduction in fares. As a general principle, the CAB chose not to have more than two or three carriers serving the same route.

Comparison to Motor-Carrier Regulation

In practice, airline regulation is quite comparable to the regulation of motor carriers. Prices were generally set to allow reasonable profits and entailed cross-subsidization from high-density to low-density markets. Similarly, entry was controlled into routes as well as into the industry. However, it is generally believed that the CAB was more lenient than the ICC in terms of allowing existing firms to enter currently served routes. Regardless of the similarities in regulatory practices, the effects of regulation differed considerably between trucking and airlines. As we will observe, this difference was largely due to technological differences in the service being provided, as well as to airlines’ responding to regulation in a quite different manner from that of motor carriers.

Effects of Regulation

In comparing the effects of price and entry/exit regulation across industries, certain regularities arise. One common effect of regulation is that it reduces productivity growth. This was observed for the railroad industry, and we will find it also to be true for the airline industry. Another common effect of regulation is that the prevention of entry maintains inefficient firms

33. Breyer, Regulation.
34. Ibid.
in the industry. This was found to be true for the motor carrier industry, inasmuch as deregulation induced a large number of bankruptcies while simultaneously inducing considerable entry. Entry regulation will be seen to have had the same effect in the airline industry. But there are typically also effects of regulation that are unique to an industry, or at least the extent to which these effects are significant is more extreme than in other industries. These effects are due to interindustry differences with respect to the product or service being offered and the type of production technology, as well as the idiosyncratic features of any particular regulatory structure.

The regulation of the airline industry offers a case study of two classic effects of price and entry/exit regulation. First, if the government takes away price as a competitive instrument, then firms will turn to competing in other ways. In the case of airlines, firms aggressively competed through quality of service. Second, it is very difficult to predict the effects of regulation, because it is difficult to foresee the new and innovative means of providing a better product at a lower cost than competition would have brought about. The development of the hub-and-spoke system since deregulation is such an innovation. The stifling of the adoption of the hub-and-spoke system was an unanticipated yet substantial effect of regulation.

### Price and Quality of Service

Our first task is to establish the effect of CAB regulation on airfares. One method for doing this is to compare airfares in regulated and unregulated markets over the same time period. In chapter 16, this method was referred to as the intermarket approach. Although all interstate markets were subject to CAB regulation, intrastate markets were not. We can then assess the effect of regulation on airfares by comparing fares on intrastate and interstate routes that are of similar length and density. To be able to make such a comparison, one must consider intrastate routes in large states like California and Texas, where routes are of sufficient length and density to allow a reasonable comparison to be made with some interstate routes.

Table 17.5 compares fares for intrastate routes in California and comparable interstate routes subject to CAB regulation. Fares in unregulated intrastate markets were considerably

<table>
<thead>
<tr>
<th>Table 17.5</th>
<th>Differential between Intrastate Fares in California and Interstate Fares, 1972</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intrastate Fare/Mile (in cents)</td>
</tr>
<tr>
<td>Very short haul (65 miles)</td>
<td>16.923</td>
</tr>
<tr>
<td>Short haul (109 miles)</td>
<td>9.363</td>
</tr>
<tr>
<td>Short-medium haul (338–373 miles)</td>
<td>5.021</td>
</tr>
</tbody>
</table>

below fares for CAB-regulated markets. Also note that this differential was greater, the longer the length of the route. Fares for some specific routes in Texas for 1975 are provided in table 17.6. Southwest Airlines was the primary intrastate carrier in Texas and consistently provided lower fares than were set by CAB-regulated carriers. Cross-sectional data on fares during the time of CAB regulation establish that airfares were kept excessively high in some markets.

An alternative method for determining the effects of regulation on airfares is to observe how fares changed over time as the industry moved from a regulated to an unregulated status. Figure 17.4 shows the change in average real airfares between 1978 and 1993. Fares fell substantially for routes exceeding 1,000 miles but rose substantially for routes less than 500 miles. From a different perspective, airfares between major large metropolitan areas declined by 8.7 percent for long-haul markets and by 14.5 percent for short-haul markets between 1976 and 1983, whereas fares rose on routes between small cities by 13.2 percent for short-haul markets and by more than 50 percent for medium-haul markets.35 Consistent with the intermarket evidence in tables 17.5 and 17.6, the intertemporal evidence confirms the hypothesis that CAB price regulation was characterized by cross-subsidization, with fares in high-density markets being set above cost and fares in low-density markets being set below cost.

Another important change in the fare structure since deregulation is the sizable increase in discount fares and the ensuing increase in the number of passengers traveling on discount fares. In major markets only about a quarter of passengers traveled on discount fares in 1976, whereas almost three-quarters did so in 1983. In low-density markets, the rise in the use of

Table 17.6
Comparison of Interstate and Intrastate Fare Levels in Selected Texas Markets, December 1, 1975

<table>
<thead>
<tr>
<th>Fare Type</th>
<th>Dallas–Houston</th>
<th>Dallas–Harlingen</th>
<th>Dallas–San Antonio</th>
<th>Houston–Harlingen</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAB Interstate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First class</td>
<td>$48.00</td>
<td>$51.00</td>
<td>$51.00</td>
<td>$42.00</td>
</tr>
<tr>
<td>Coach</td>
<td>35.00</td>
<td>37.00</td>
<td>33.00</td>
<td>38.00</td>
</tr>
<tr>
<td>Economy</td>
<td>32.00</td>
<td>51.00</td>
<td>33.00</td>
<td>38.00</td>
</tr>
<tr>
<td>Southwest Intrastate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“Executive class”</td>
<td>25.00</td>
<td>40.00</td>
<td>25.00</td>
<td>25.00</td>
</tr>
<tr>
<td>“Pleasure class”</td>
<td>15.00</td>
<td>25.00</td>
<td>15.00</td>
<td>15.00</td>
</tr>
</tbody>
</table>


discount fares was even greater.\footnote{Ibid.} This pattern reveals a greater use of price discrimination as airlines differentiate between the relatively price-elastic demand of pleasure travelers—who can, in many instances, easily adjust their day and time of departure in response to low fares—and the relatively price-inelastic demand of business travelers, who often lack scheduling flexibility.

If fares were set above cost in some markets, an airline would be very interested in increasing its sales in those markets. Because fare reduction was discouraged by the CAB, airlines were forced to compete by making the service they provided more attractive to consumers. To understand the form of this competition, one must examine a consumer’s decision as to which airline to fly. To begin, the effective price of air travel to a consumer includes more than just the price of the airline ticket. Since consumers value their time, one measure of the cost of travel is the airfare plus the value of a consumer’s time used during travel. The farther is the actual departure time of a flight from a consumer’s ideal departure time, the greater the effective price of travel. The longer the travel time, the greater the effective price of travel.

Figure 17.4
Percentage Change in Airfares by Distance, Adjusted for Inflation
Thus, even if all airlines charge the same airfare, the effective price can differ across airlines as well as across consumers. The effective price depends on travel time, departure times, and a consumer’s value of time.

In addition to affecting the time cost of travel through the frequency of flights and the availability of nonstop flights, airlines can influence the pleasure derived from air travel. An airline that is less crowded and has better on-board services and a better safety record is more desired by consumers. Ceteris paribus, such an airline will have greater demand. Table 17.7 provides estimates of the average consumer’s willingness to pay for improvements in some of these nonprice factors. It shows that, on average, a consumer is willing to pay $5.67 for a ten-minute reduction in travel time. In other words, if two airlines offered identical services but one airline’s travel time was ten minutes less, the average consumer (flying the average route) would prefer the faster airline, as long as its airfare was no more than $5.67 more expensive than the slower airline. If an airline raised the reliability of a flight being on time by ten percentage points, consumers would be willing to pay a higher airfare by the amount of $12.13.

As the decision of which airline to fly is influenced by factors other than price, airlines are able to compete through nonprice methods in order to raise the demand for their service. One important nonprice variable is flight frequency and, associated with that, load factor. Load factor is the number of passengers on a flight divided by the number of seats. Under fare regulation, airlines competed by offering a wide array of flights that resulted in low load factors. By doing so, an airline was more likely to have the departure time desired by a consumer, and in addition, low load factors resulted in less crowded and thus more pleasant flights.

By examining time-series evidence we can consider the effect of regulation on load factors. From a cost perspective, it is optimal for airlines to use larger aircraft for longer distances. Given that the cost of a seat rises with distance, the optimal load factor also rises with distance. One would then predict load factors to rise with distance in an unregulated environment. However, if regulation induced excessive nonprice competition, then one would expect load factors to fall with distance. Fares were more above cost on long-haul markets, so
nonprice competition would be stronger in those markets, which means that load factors should have been lower.

The estimated relationship between load factors and distance for the regulation year 1969 and for the deregulation years of 1976 and 1981 is shown in figure 17.5. As expected, load factors did indeed fall with distance under regulation and rose with distance after deregulation. By keeping fares above cost, regulation induced heightened nonprice competition in terms of increased flight frequency and low load factors. Although consumers valued the lower load factors, regulation induced a nonoptimal mix of fares and load factors. Consumers would have been willing to accept lower fares and higher load factors than those achieved under CAB regulation.37

Excessive nonprice competition was also observed in terms of on-board services. From 1976 to 1982, the Consumer Price Index for food rose 62 percent, whereas the cost of food for airlines rose only 40 percent.38 Because deregulation reduced the difference between fares and cost, it reduced the level of competition in terms of the quality of food. Over the same time period, the number of flight attendants per passenger fell 16 percent.39

---

39. Ibid.
these changes show that firms achieved a higher level of nonprice competition under regulation.

Fares in some markets have fallen since deregulation, but the preceding evidence suggests that quality has also fallen (though further evidence to be presented later puts that conclusion into question). A rough indicator that, on net, consumers are better off is to look at what has happened to the volume of air travel. It has exploded since the industry was deregulated. From 1977 to 1990, domestic air travel rose by 120 percent. A bit more refined measure shows that in 1988 air travel was 41 percent above the level projected by the trend established during the regulated period 1955–1978.

**Development of the Hub-and-Spoke System**

From the evidence thus far, it would seem that regulation resulted in airlines offering a high-priced, high-quality product. The welfare loss to consumers was then in not having the option of choosing a low-priced, low-quality product. In assessing the welfare implications of regulation, it would then seem that one must consider both the welfare loss from higher fares and the welfare gain from better quality of service. It is indeed true that regulation raised quality by having lower load factors and greater on-board services, but in terms of one important quality dimension, regulation actually resulted in lower quality of air service. Flight frequency has risen since the airline industry was deregulated. Consumers now have a wider array of departure times from which to choose. That flight frequency could increase in spite of higher load factors is due to the adoption of the hub-and-spoke system.

To understand the economics of the hub-and-spoke system, consider an airline that serves two large destinations that we will refer to as cities B and C. It also serves two smaller destinations, called towns A and D, with A reasonably near B and D reasonably near C. (Imagine these destinations along a line running from A to B to C to D.) The array of route markets consists of A-B (that is, traveling from A to B or B to A), A-C, A-D, B-C, B-D, and C-D. Given these six markets, one route system is to have six pairs of direct flights, so that, for example, someone in town A who wanted to travel to city C would take a direct flight from A to C. An alternative route system is to make hubs out of cities B and C, so that all travelers go through these two points. For example, a traveler desiring to go from A to C would first fly from A to B and then take a second flight from B to C. The virtue of this system is that it concentrates traffic. So, for example, a flight from B to C would include not only those whose initial point of departure and final point arrival are B and C, respectively, but also those who started in A and want to go to C or D. Cost savings are created by allowing the use of larger aircraft for the portion from B to C. In contrast, with direct flights, smaller

---


aircraft with higher per-unit costs would have to be used to transport someone directly from A to C. This concentration of traffic allows for a larger array of flights. If the density of traffic from A to C is sufficiently light, it may only be sufficient to support one daily direct flight from A to C. However, combining travelers desiring to go from A to C with those wanting to go from A to B and A to D, there may be several daily flights from A to B, and, of course, there are many flights between B and C, given the high density of travelers between those two hubs.

One of the unanticipated developments since deregulation has been the widespread adoption of the hub-and-spoke system. To see how drastically route structures have changed, figure 17.6 shows the route structure of Western Airlines before and after deregulation. As should be apparent, the two hubs are Salt Lake City and Los Angeles. By restricting the entry of existing carriers into currently served markets, CAB regulation prevented the massive route restructuring required to move to a hub-and-spoke system.

In spite of reduced nonprice competition, the development of the hub-and-spoke system puts into doubt the claim that the overall quality of service is lower under deregulation, because the hub-and-spoke system trades off longer travel time (as there are fewer direct flights, so that more flights require a transfer) for a wider array of departure times. To address this question, a study analyzed 812 city pairs with and without regulation. It used actual data on fares and flight frequency for 1977 as the regulatory benchmark. To ascertain the effect of regulation, these data are compared with projected data for 1977 conditional on the route structure and fares being what they were under deregulation (specifically, 1983). Table 17.8 shows that real fares declined in medium hub–large hub and large hub–large hub routes but rose in all other markets. This finding is consistent with airfares being set above cost in dense markets and below cost in less dense markets. With the movement to

| Table 17.8 |
| Weighted Average Percentage Change in Fares and Frequency from Deregulation (1977) |

<table>
<thead>
<tr>
<th>Category of Route</th>
<th>Number of City Pairs</th>
<th>Coach Fare</th>
<th>Discount Fare</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonhub–nonhub</td>
<td>51</td>
<td>21.2</td>
<td>22.1</td>
<td>33.9</td>
</tr>
<tr>
<td>Nonhub–small hub</td>
<td>52</td>
<td>22.5</td>
<td>12.3</td>
<td>1.4</td>
</tr>
<tr>
<td>Nonhub–medium hub</td>
<td>45</td>
<td>5.4</td>
<td>-0.4</td>
<td>24.3</td>
</tr>
<tr>
<td>Nonhub–large hub</td>
<td>53</td>
<td>16.3</td>
<td>9.1</td>
<td>28.7</td>
</tr>
<tr>
<td>Small hub–hub</td>
<td>60</td>
<td>15.3</td>
<td>11.3</td>
<td>33.9</td>
</tr>
<tr>
<td>Small hub–medium hub</td>
<td>69</td>
<td>18.7</td>
<td>10.4</td>
<td>20.8</td>
</tr>
<tr>
<td>Small hub–large hub</td>
<td>57</td>
<td>25.0</td>
<td>8.1</td>
<td>19.2</td>
</tr>
<tr>
<td>Medium hub–medium hub</td>
<td>69</td>
<td>15.6</td>
<td>2.0</td>
<td>-4.3</td>
</tr>
<tr>
<td>Medium hub–large hub</td>
<td>161</td>
<td>17.4</td>
<td>-6.8</td>
<td>14.4</td>
</tr>
<tr>
<td>Large hub–large hub</td>
<td>201</td>
<td>8.6</td>
<td>-17.6</td>
<td>-3.5</td>
</tr>
</tbody>
</table>

Figure 17.6
The Adoption of the Hub-and-Spoke System by Western Airlines

hub-and-spoke systems, flight frequency increased for most markets. For example, there were a third more flights for nonhub–nonhub and small hub–small hub. Overall, the route-weighted average frequency increased by 9.2 percent as a result of the restructuring of the route system.

The development of the hub-and-spoke system points out an important lesson from economic regulation which was perhaps best said by Alfred E. Kahn, former chairman of the CAB:

The essence of the case for competition is the impossibility of predicting most of its consequences. The superiority of the competitive market is the positive stimuli it provides for constantly improving efficiency, innovating, and offering consumers diversity of choices.42

Before deregulation, economists generally believed that regulation caused a nonoptimal product mix; fares were too high and quality of service was too high. What was not recognized was that regulation, through restrictions on entry, was holding back a restructuring of the route system that would result in lower costs and greater flight frequency. Deregulation reduced quality of service by raising load factors and travel time and reducing on-board services, but it raised quality by increasing the number of departures. The development of the hub-and-spoke system points out that it is only in retrospect that we can adequately understand the effects of regulation.

Welfare Estimates from Changes in Price and Quality

In considering the welfare effects of regulation, let us begin by focusing on the initial period of deregulatory activity by looking at movements in fares and travel times and delays from the end of 1976 to the end of 1978.43 This period captures the initial effect of price deregulation but not the effect of the restructuring of the route system. The average real standard coach fare fell by almost 5 percent, while first-class fares went from being 150 percent of the coach fare to 120 percent. Because lower fares brought forth greater volume of traffic, load factors rose from 55.6 percent to 61.0 percent. Once the effect of deregulation on travel time and delays is taken into account, the estimated gain to consumers was about 10 percent of an average prederegulation round-trip fare or about $25 to $35 per round trip (recall that all figures are in 1985 dollars). Any undesirable changes in travel time and delay were more than compensated for by the reduction in fares. It is important to note that this welfare measure does not take into account lower welfare due to reduced on-board services and higher load factors.


Table 17.9 provides the most up-to-date estimates of the welfare gains from deregulation. Consumers are gaining $12.4 billion annually from lower fares under deregulation and $10.3 billion from greater flight frequency. On the downside, increases in travel restrictions, travel time, load factors, and the number of connections have all reduced consumer welfare. On net, the gains are quite impressive, in excess of $18 billion annually.

Regulation not only reduced consumers’ welfare but also negatively affected industry profits. If the industry had not been regulated in 1977, industry profits would have been higher by over $4 billion. This is a surprising finding inasmuch as the airline industry has performed rather poorly since deregulation. Although it earned an annual rate of return of 1.30 percent over 1970–1977, it earned an even more feeble annual return of 0.10 percent over 1979–1986. However, the airline industry was hit by a number of negative profit shocks after deregulation, in particular the recession of 1981–1983, a sharp rise in fuel prices, and several strikes by union employees (which can be perceived as part of the transitional costs from deregulation).

Dynamic Productive Inefficiency

To assess the effect of regulation on productivity growth, a cross-country study was performed that examined productivity growth for twenty-one U.S. airlines (including all trunk airlines) with that achieved by twenty-seven non-U.S. airlines for the period 1970–1983. This is a reasonable comparison, inasmuch as countries’ airline industries are similar in many respects except for the regulatory environment and labor costs. Aircraft and fuel are sold in

Table 17.9
Annual Gains to Travelers from Airline Deregulation (billions of 1993 dollars)

<table>
<thead>
<tr>
<th>Category</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fares</td>
<td>12.4</td>
</tr>
<tr>
<td>Travel restrictions</td>
<td>-1.1</td>
</tr>
<tr>
<td>Frequency</td>
<td>10.3</td>
</tr>
<tr>
<td>Load factor</td>
<td>-0.6</td>
</tr>
<tr>
<td>Number of connections</td>
<td>-0.7</td>
</tr>
<tr>
<td>Mix of connections (on-line/interline)</td>
<td>0.9</td>
</tr>
<tr>
<td>Travel time</td>
<td>-2.8</td>
</tr>
<tr>
<td>Total</td>
<td>18.4</td>
</tr>
</tbody>
</table>


world markets, while operation and maintenance of aircraft are governed by strict international standards.

Dating the beginning of deregulation at 1976, table 17.10 compares the annual percentage decline in unit costs over periods of regulation and deregulation and across U.S. and non-U.S. airlines. Operating characteristics include changes due to traffic, route density, firm size, load factor, and capacity utilization. Technical efficiency is composed of all other sources of change in factor productivity. Under regulation, the annual decline in unit cost for U.S. airlines was only 3.0 percent, compared to 4.5 percent for non-U.S. airlines. After deregulation, U.S. airlines outperformed non-U.S. airlines by experiencing a 3.3 percent annual decrease in unit cost. Consistent with the experience in the railroad industry, airline regulation reduced productivity growth.

### Airline Safety

A constant fear expressed in public forums is that deregulation would cause airline safety to deteriorate. On the surface, this would seem to be an irrational fear, because airline safety is controlled by the Federal Aviation Administration and safety regulations were unaffected by the Airline Deregulation Act. Further reflection, however, raises the potential for concern because nonprice competition may generate levels of safety exceeding that mandated by law. If deregulation reduced nonprice competition and intensified efforts to keep costs down, then perhaps it could have reduced safety. How might this outcome occur? Perhaps a greater need to keep costs low resulted in airlines using less experienced pilots and skimping on aircraft maintenance. Furthermore, safety is also influenced by the amount of traffic and congestion, both of which have risen since deregulation. On the other hand, there are some factors that could have caused safety to rise since deregulation. Increased air travel requires more planes; therefore, the average age of the fleet is lower. Increased competition could result in more efficient practices so that costs are lower and safety is higher. Finally, there is continual technological progress that increases airline safety.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>U.S.</td>
<td>Non-U.S.</td>
</tr>
<tr>
<td>Operating characteristics</td>
<td>1.8</td>
<td>3.1</td>
</tr>
<tr>
<td>Technical efficiency</td>
<td>1.2</td>
<td>1.4</td>
</tr>
<tr>
<td>Total productive efficiency</td>
<td>3.0</td>
<td>4.5</td>
</tr>
</tbody>
</table>

There are numerous ways in which to measure airline safety, including fatalities, accidents, and incidents (for example, the number of near midair collisions). These variables could be measured per departure or per passenger mile. Regardless of which measure is used, the conclusion is the same: the long-term rise in airline safety has continued during the years of deregulation.46 Figure 17.7 shows the steady decline in the total accidents per million departures. If one looks at the number of fatalities per passenger mile (not shown in figure 17.7), there is no clear trend, as the measure is quite volatile. The crash of a single large aircraft typically dominates the annual number. Still, the fatality risk for a passenger on a U.S. domestic jet flight was 1 in every 2.5 million flights in 1971–1978 and 1 in every 7.4 million flights in 1979–1986.47 Further evidence that safety has not deteriorated during the years of deregulation is that aviation insurance rates have not risen.48 As perceived by insurance specialists, deregulation has not compromised airline safety.


That airline safety did not suffer, with a 52 percent increase in passenger-miles over 1976 to 1986, is striking. As mentioned, this record may simply reflect a long-term trend in improvements in safety due to better equipment and practices. There is also some evidence that deregulation has resulted in more efficient practices. One study examined the maintenance practices of seven airlines and found that the length of time between maintenance shop visits for jet engines increased since deregulation but that there was no effect on the probability of an engine shutdown.\footnote{D. Mark Kennet, “Did Deregulation Affect Aircraft Engine Maintenance? An Empirical Policy Analysis,” \textit{RAND Journal of Economics} 24 (Winter 1993): 542–58.}

\section*{Competition and Antitrust Policy after Deregulation}

If an industry is subject to price and entry regulation, then there is a greatly reduced need for antitrust policy. With prices being set by the regulatory agency, collusion in price is irrelevant, though one might still need to be concerned about firms colluding along unregulated dimensions like product quality. With entry restrictions, there is little need for concern about entry-deterring or predatory tactics designed to increase concentration. However, once an industry is deregulated, an active role for antitrust policy should emerge. Because the benefits from deregulation are predicated upon firms not conspiring to raise price, not acting to deter entry, and not predatorily driving new competitors out of the industry, antitrust policy can be critical in achieving those benefits. The Airline Deregulation Act of 1978 took account of the important postderegulation role for antitrust policy in its mandating that the CAB put additional emphasis on the standards for mergers outlined by Section 7 of the Clayton Act. All mergers or acquisitions were to require prior approval.

\section*{Concentration of the Airline Industry}

What has happened to industry concentration since deregulation? To measure industry concentration, we will use the number of “effective competitors,” which is defined as the inverse of the sum of each firm’s market share squared (or, the inverse of the HHI which is defined in chapter 6). If \( s_i \) is the market share of firm \( i \) and there are \( n \) firms, then the number of effective competitors equals \( 1/(s_1^2 + s_2^2 + \ldots + s_n^2) \). The effective number of competitors for an industry is the number of equal-sized firms that would give the same level of HHI. For example, if there is but a single firm, so that its market share is one, then the number of effective competitors is one (which makes sense). If there are \( n \) firms with equal market share, then the number of effective competitors is \( 1/[(1/n)^2 + \ldots + (1/n)^2] = 1/[n(1/n^2)] = n \). However, if there are \( n \) firms and market share is skewed so that a few firms have a large part of the market, then the number of effective competitors is less than \( n \).

Using market share data at the national level, figure 17.8 shows how the number of effective competitors has changed between the passage of the ADA in 1978 and 1994. During the
In the early years after deregulation, entry by new airlines and the expansion of regional airlines resulted in a rise in the number of effective competitors (concentration went down). Then a consolidation process started to emerge that was particularly intense around 1986. Some airlines went bankrupt and some were acquired. Although deregulation initially increased the effective number of competitors at the national level, unconstrained merger activity caused it to retreat to a level below the number under regulation.

Because competition occurs at the level of a route—for example, Boston to Chicago—measures of concentration at the national level can be misleading and are typically inappropriate for understanding the competitiveness of airline markets. To get a clearer picture of what has happened to competition since deregulation, figure 17.9 plots the number of effective competitors for each route, averaged over all routes, routes in excess of 2,000 miles, and routes under 500 miles. Similar to what was occurring at the national level, competition steadily rose in the immediate aftermath of deregulation as the average number of effective competitors (for all routes) climbed from 1.7 to 2.5 by 1986. Consistent with figure 17.8, the industry merger wave reduced the number of effective competitors in the mid- to late 1980s. However, contrary to what occurred at the national level, overall route markets are more competitive today than they were under regulation. This statement is especially true for routes in excess of 2,000 miles, though competition has not really changed much for routes under 500 miles.

**Figure 17.8**

Industry Concentration: National Level

*Source:* Market share data used in calculating the number of effective competitors is a carrier’s share of domestic passenger miles. Figure from Steven A. Morrison and Clifford Winston, *The Evolution of the Airline Industry* (Washington, D.C.: Brookings Institution, 1995).
In the deregulated mid-1990s with the regulated late 1970s, concentration is higher today at the national level, but those airlines are serving more routes, so that on average, concentration is lower in the route markets. An important caveat is that the hub-and-spoke system has resulted in some airports being dominated by a single carrier. This topic will be discussed later.

Antitrust policy was conspicuously absent in the decade after regulation. The CAB was quite lenient in its standards for airline mergers, and the Department of Transportation (DOT) (which has handled airline mergers since 1984) has approved every airline merger, some against the recommendation of the Department of Justice (see table 17.11). Under this laissez-faire policy, twenty airlines merged in 1985–1987. In response to these events, former CAB chairman Alfred E. Kahn stated,

\[\text{[The] reconcentration of the industry reflects in part the deplorable failure of the Department of Transportation to disallow even one merger, or, in all but one case, even to set conditions to mitigate possible anti-competitive consequences. The DOT seems to have no appreciation whatever of the dangers our antitrust laws were set up almost a century ago to forestall.}\]

\[50. \text{Alfred E. Kahn, “I Would Do It Again,”} \textit{Regulation} \textit{vol.} 12, no. 2 (1988): 22–8.\]
The airline industry has been a hotbed of contentious antitrust issues. Numerous practices have come under government and academic scrutiny for possibly making entry more difficult, serving to drive out competitors, and, more generally, creating an anticompetitive advantage for some or all of the established firms.

### Airport Gates and Slots

The most obvious obstacles to entry are the restrictions on the right to take off or land at a given time of day. These takeoff and landing slots are in short supply at four major airports—O’Hare in Chicago, LaGuardia and Kennedy in New York, and Ronald Reagan National Airport in Washington, D.C. In these cases, therefore, potential competition is less effective in disciplining carriers with high fares. Furthermore, these slots are largely controlled by a small number of airlines. In 1998, at LaGuardia, American, Delta, and US Airways controlled

### Table 17.11

<table>
<thead>
<tr>
<th>Year</th>
<th>Carriers</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1979</td>
<td>Pan American–National</td>
<td>Approved by CAB</td>
</tr>
<tr>
<td></td>
<td>Texas International–National</td>
<td>Approved by CAB</td>
</tr>
<tr>
<td></td>
<td>Eastern–National</td>
<td>Not consummated</td>
</tr>
<tr>
<td></td>
<td>Continental–Western</td>
<td>Anti-competitive finding by CAB</td>
</tr>
<tr>
<td></td>
<td>North Central–Southern</td>
<td>Not consummated</td>
</tr>
<tr>
<td>1980</td>
<td>Republic–Hughes Air West</td>
<td>Approved by CAB</td>
</tr>
<tr>
<td>1981</td>
<td>Continental–Western</td>
<td>Approved by CAB</td>
</tr>
<tr>
<td></td>
<td>Texas International–Continental</td>
<td>Approved by CAB</td>
</tr>
<tr>
<td>1982</td>
<td>Air Florida–Western</td>
<td>Approved by CAB</td>
</tr>
<tr>
<td>1985</td>
<td>United–Pan American (Pacific Division)</td>
<td>Opposed by DOJ</td>
</tr>
<tr>
<td>1986</td>
<td>Delta–Western</td>
<td>Approved by DOT</td>
</tr>
<tr>
<td></td>
<td>Texas Air–People Express</td>
<td>Not-anti-competitive finding by DOJ</td>
</tr>
<tr>
<td></td>
<td>Texas Air–Eastern</td>
<td>Approved by DOT after sale of slots to Pan Am Shuttle</td>
</tr>
<tr>
<td></td>
<td>Trans World Airline–Ozark</td>
<td>Opposed by DOJ</td>
</tr>
<tr>
<td></td>
<td>Northwest–Republic</td>
<td>Approved by DOT</td>
</tr>
<tr>
<td>1987</td>
<td>US Air–Piedmont</td>
<td>Approved by DOT</td>
</tr>
<tr>
<td></td>
<td>American–Air California</td>
<td>Approved by DOT</td>
</tr>
<tr>
<td></td>
<td>US Air–Pacific Southwest</td>
<td>Approved by DOT</td>
</tr>
</tbody>
</table>


### Issues of Anticompetitive Nonpricing Practices

The airline industry has been a hotbed of contentious antitrust issues. Numerous practices have come under government and academic scrutiny for possibly making entry more difficult, serving to drive out competitors, and, more generally, creating an anticompetitive advantage for some or all of the established firms.
62 of 98 slots, and at O'Hare, American and United controlled 82 of 101 slots—well above their holding of 66 of 100 slots in 1986. A second source of difficulty in airport access is the concentration of control over gates at hubs as carriers may control a large fraction of gates. After its acquisition of Ozark Airlines in 1986, TWA controlled over 75 percent of the gates at the St. Louis airport. Shortly thereafter, TWA raised its fares by 13 to 18 percent on its flights out of St. Louis. Exacerbating matters is the nature of the contracts between an airline and the airport in that they are often long-term and give an airline exclusive use of a gate. These factors can make entry into some airports difficult.

**Computer Reservation Systems**

Airline flights are booked using a computer reservation system (CRS), and their ownership by individual airlines has come under scrutiny for being anticompetitive. The original CRSs were introduced in 1976 with American Airlines’ Sabre system and United Airlines’ Apollo system. Ownership of the CRS used by a travel agent could give an airline an anticompetitive advantage in two ways. First, the airline could program the CRS to provide more information on its flights and require fewer keystrokes to access that information. Second, at least prior to the prohibition of the practice by the CAB in 1984, an airline could make its own flights more prominent on the screen. For example, in response to a request for a 3:00 P.M. flight, the Apollo system might show a 1:00 P.M. United flight prior to showing a 3:00 P.M. flight on another airline. Although CRSs clearly resulted in better servicing of customers by travel agents, they could provide anticompetitive advantages to established carriers. It is also argued that this advantage is accentuated by merger.

**Frequent Flier Programs**

In 1981, American Airlines introduced an innovation that has since become an industry fixture: frequent flier programs. These programs boost an airline’s demand by offering awards of free flights to passengers who have flown some minimum number of miles on the airline. Such programs are especially effective with business travelers, given they are not typically paying for the flight though often are the recipients of the free flights. It is argued that these programs give an anticompetitive advantage to airlines with more routes, since a traveler interested in accumulating frequent flier miles will prefer to concentrate her mileage on one airline rather than split it between two smaller airlines. For example, if it takes 40,000 miles to receive a free domestic flight and a passenger flies 50,000 miles, she will prefer to fly


50,000 miles on one airline rather than to fly 25,000 miles on each of two airlines. In addition, because larger airlines fly to more places, a free flight on one of them will be a more attractive award.

**Marketing Alliances**

A new antitrust issue emerged with the announcement in 1998 of marketing alliances between U.S. airlines. In that year, three pairs of airlines—American Airlines and US Airways, United Airlines and Delta Air Lines, and Continental Airlines and Northwest Airlines—proposed to consolidate some of their activities. While maintaining separate companies, these alliances would have airlines link their route systems, so that passengers would not have to pay a higher fare for traveling on, say, American Airlines and US Airways on the same trip. A second feature is the consolidation of the airlines’ frequent flier programs. The attractiveness of these alliances is that airlines can reap some of the benefits of a merger without incurring many of the costs. They have already taken place between U.S. and foreign airlines with, for example, United and Lufthansa in 1994 and, soon thereafter, British Airways and American Airlines. The latter was subject to scrutiny because the two airlines control a large percentage of the gates at London-Heathrow. However, initial studies suggest that these alliances have mixed effects. They can cause fares to rise because of coordination by two competitive carriers but may also result in lower fares by allowing for more efficient pricing of flights involving multiple airlines. One such study finds alliances to be welfare-enhancing.54

**Issues of Anticompetitive Pricing Practices**

**Predatory Pricing**

A recurrent issue since deregulation is whether large carriers have periodically responded to new entrants with a pricing strategy intended to drive them out of the carriers’ markets.55 In commenting on the exit of many of the early postderegulation entrants, former CAB Chairman Alfred Kahn said:

I take perverse satisfaction in predicting the demise of price-cutting competitors like World and Capital Airways if we did nothing to limit the predictable geographically discriminatory response of the incumbent carriers to their entry.56

While this issue died down after many of those new airlines departed, it has now reappeared as a contentious antitrust issue. Smaller airlines have complained to the antitrust

55. For a more general treatment of predatory pricing, the reader is referred to chapter 9.
In chapter 9, we defined predatory pricing

authorities that they have suffered at the hand of predation. As an example, Frontier Airlines accused United Airlines of using predatory pricing to drive it out of the Billings–Denver route.57 Frontier entered in 1994 with a fare of around $100, about half of United’s preentry fare (see figure 17.10). United responded with a comparable fare. In response to these low fares, demand rose by 60 percent. After about a year, Frontier withdrew from the route and logged a complaint with the DOJ. United’s fare rose precipitously upon Frontier’s departure.58

What is one to make of this episode? Is United just responding as any good competitor would, or are they acting in a predatory manner? In chapter 9, we defined predatory pricing


58. This encounter is not necessarily typical. When Southwest Airlines entered United’s Los Angeles–Sacramento route, it set a fare of $56, well below United’s fare of around $100. Three years later, both carriers were still serving that route and fares remained low. In the case of Southwest’s entry of the Baltimore–Cleveland route, intense fare competition did induce exit, but by the incumbent firm, US Airways! (Ibid.)
as “pricing at a level calculated to exclude from the market an equally or more efficient competitor.” In practice, a necessary condition to establish predatory pricing in an antitrust case is to show that the predator priced below some appropriately defined measure of cost per unit. For our purposes, we will suppose that it is average cost. So, did United price below its average cost? Could its pricing behavior only be rationalized if it induced the exit of Frontier? To try to address these questions, a little notation will be useful. Let \( D_F(P_F, P_U) \) and \( D_U(P_U, P_F) \) represent the demand functions for Frontier and United, respectively, with \( P_F \) and \( P_U \) denoting the prices for Frontier and United. The variables \( AC_F(q_F) \) and \( AC_U(q_U) \) are their average cost functions with \( q_F \) and \( q_U \) denoting the quantities of Frontier and United.

We will assume that average cost is declining because of the fixed costs associated with air travel. In light of the low marginal cost of transporting one more passenger (given that the fixed cost of flying a plane has already been incurred), it is unlikely that either United or Frontier was pricing below marginal cost at a price of $100. The issue of average cost is more difficult. If Frontier accurately forecasted demand and presuming it did not expect United to respond to entry by raising price (in that it would hardly make any sense for United), then we would expect \( 100 > AC_F(D_F(100; 200)) \). The average cost of Frontier when it produces to meet demand of \( D_F(100; 200) \) is \( AC_F(D_F(100; 200)) \). If this were not true, then it would be difficult to rationalize Frontier’s entry as it would have been pricing below its average cost given United’s current price and thus entry would be unprofitable at this or any lower price for United. We also know that Frontier eventually exited after United matched its price, so that it is reasonable to infer that Frontier was incurring losses, which means \( 100 < AC_F(D_F(100; 100)) \).

That Frontier’s price of 100 exceeds its average cost when United prices at 200 but falls below its average cost when United prices at 100 is depicted in figure 17.11(a), where \( q_F^* = D_F(100; 200) \) and \( q_F'' = D_F(100; 100) \). In other words, \( AC_F(q_F^*) > 100 > AC_F(q_F'') \). Market demand, \( D(P) \), is the total number of units demanded when all firms price at \( P \). Because the firm with a lower price should be expected to have a higher market share, Frontier’s demand curve is closer to the market demand curve, the more its price undercuts United’s price. Also note that Frontier’s demand is half of market demand when it prices the same as United. These properties presume that Frontier and United offer comparable service and can supply as much as is demanded. The former assumption will be discussed later. Finally, note that Frontier’s demand curve shifts in when United lowers its price from 200 to 100. As these airlines are offering substitutes, a lower price for United causes some passengers to switch from Frontier to United, so that Frontier’s demand is lower.

59. The issue of whether it should be average cost, average variable cost, or marginal cost is a difficult one and is discussed in chapter 9.

60. At a lower price for United, Frontier’s demand would be less, implying that its average cost is higher, since we have assumed average cost is declining.
Figure 17.11
Evaluation of Postentry Pricing for Predation
If United were no more efficient than Frontier—that is, $AC_U(q) \geq AC_F(q)$ for all $q$—and it had a comparable demand function, it follows from the preceding analysis that United’s price would also be below its average cost: $100 < AC_U(q^*_U)$, where $q^*_U = D_U(100; 100)$.\(^{61}\) This situation is depicted in figure 17.11(b). In responding to Frontier’s price by matching it, United would be creating a situation that is not stable in the long run, as both firms are incurring losses. To rationalize such a pricing response, United would have to expect higher profits in the future. This could happen if Frontier exited and United responded by raising its price. This result is of course what did occur. A second scenario is that the punishment of Frontier contributes to United’s reputation for responding aggressively to entry. Even if Frontier does not exit, this enhanced reputation may benefit United by deterring entry into its other markets. A third scenario is that United’s price matching induces Frontier to raise its price with the anticipation that United would follow and the two firms would achieve a long-run equilibrium at a price above 100 that (perhaps) would allow both airlines to earn positive profit.

All of this discussion presumes that United and Frontier had comparable demand and cost functions. Alternatively, United could have lower cost that might allow it to earn positive profit at a price of 100. This situation is depicted in figure 17.11(b), where now United’s average cost function is $\bar{AC}_U(q_U)$. Even if it were still the intent of United to drive out Frontier, it would be driving out a less efficient firm. Another variant in our discussion is the possibility that United’s flights may be perceived to be of higher quality by passengers. At a price comparable to Frontier’s, United’s demand would then have been higher, allowing United’s average cost to be less than Frontier’s. In that case, a price of 100 might have exceeded United’s average cost (though falling below Frontier’s average cost). This line of argument rationalizes price matching without the intention of inducing Frontier to exit. A caveat is warranted, however. If United did offer a superior service so that passengers were willing to pay more to fly United, why did it match Frontier’s price rather than take advantage of its premium service by setting a premium price? Perhaps there was then predatory intent? This discussion only begins to encompass the many treacherous issues involved in determining the presence of predatory pricing.

**Price Signaling**

As discussed in chapter 5, one of the necessary tasks for an effective price-fixing arrangement is that firms coordinate on what prices to set. Failure to do so can result in some firms being discontent with the current situation and thus lead to a breakdown of collusion and an

---

\(^{61}\) An incumbent firm like United and one of these new competitors like Frontier are most likely to have different cost functions, as an incumbent’s system is typically hub-and-spoke while a new competitor’s system is typically point-to-point. However, which system should yield lower cost is unclear. While there appear to be economies from the hub-and-spoke network, Southwest is point-to-point, and its cost was lower than that of all the major airlines (Borenstein, “Evolution”).
intensification of competition. Because there are many collusive prices, coordination is a non-
trivial task. Direct communication through such means as a clandestine meeting is the most
effective means of communication but also the most risky. Discovery of direct contact
between firms by the antitrust authorities can be extremely damaging evidence in a price-
fixing case. Indirect communication through their prices and announcements surrounding
price changes avoids the “smoking gun” but is less effective.

Tacit communication through price signaling was the center of an investigation of the major
U.S. airlines and the Airline Tariff Publishing Company (ATPCO) launched by the DOJ in
1991.62 Each day airlines submit fare changes and fares for new routes to ATPCO, as well as
ancillary information such as first and last ticket dates (that is, when a travel agent can begin
and end selling seats at a particular fare). If a first ticket date is some time in the future, then
an airline is preannouncing a fare change. ATPCO disseminates this information to all major
airlines as well as the computer reservation systems.

Severin Borenstein consulted for the DOJ on its antitrust charge and wrote that the DOJ
claimed that

the airlines had carried on detailed conversations and negotiations over prices through ATPCO. It
pointed to numerous instances in which one carrier on a route had announced a fare increase to take
effect a number of weeks in the future. Other carriers had then announced increases on the same route,
though possibly to a different fare level. In many cases cited, the airlines had iterated back and forth
until they reached a point where they were announcing the same fare increase to take effect on the same
date.63

Attached to these fare changes was further information in the form of fare basis codes and
footnote designators that, it was argued, allowed airlines to signal a connection between fares
on different routes. For example, suppose airlines A and B are competing on routes
Atlanta–Albuquerque and Boston–Buffalo and that Atlanta–Albuquerque is an important
route to A and Boston–Buffalo is an important route to B. Next suppose that A lowers its
fare on Boston–Buffalo. It is then argued that airline B could lower its fare on Atlanta–
Albuquerque and attach information that would link it to fares on Boston–Buffalo. The
intended signal to A is that B’s reduced fare on Atlanta–Albuquerque is in response to A’s
reduced fare on Boston–Buffalo, and if A wants B to raise its fare on Atlanta–Albuquerque,
then A had better raise its fare on Boston–Buffalo.

In its defense, the airlines and ATPCO argued that the preannouncement of fare changes
was done for the benefit of consumers. And, for that reason, they contended that the case
should be subject to a rule of reason rather than the more standard per se rule used for

62. This section is based on Severin Borenstein, “Rapid Price Communication and Coordination: The Airline Tariff
63. Ibid, p. 314.
price-fixing cases because, even if these practices did promote collusion, they were also beneficial to consumers.\textsuperscript{64}

This case was settled with a consent decree which was to last for ten years. The airlines are prohibited from preannouncing price increases except when they are heavily publicized (so that they are clearly of value to consumers). They are allowed to convey only basic information associated with fare changes and cannot link fares with special codes. Since there was no judicial decision, the legal status of price signaling remains ambiguous. As of 2004, some of the airlines’ decrees had expired, with those for the other airlines soon to follow.\textsuperscript{65}

\textbf{Concentration and Air Fares}

What has a smaller number of airlines and deterrents to entry meant in terms of airfares? This question must be examined at the level of the route for that is the relevant geographic market. Industry-wide concentration is still relevant, however, as the number of U.S. airlines provides a short-run upper bound on the number of competitors for a particular route. Fewer airlines may mean fewer active competitors and certainly means fewer potential competitors.

Since deregulation, we know from figures 17.8 and 17.9 that there are fewer effective competitors at the national level but that these competitors are serving more routes, so that concentration is generally lower at the route level. Of course, this statement is only true on average. Unfortunately, the hub-and-spoke system has resulted in a number of routes being dominated by a single carrier. For example, in 1996 Northwest had 77 percent of enplane-
ments in Detroit, US Airways had 87 percent in Pittsburgh, and Delta had 78 percent in Atlanta.\textsuperscript{66}

The significance of airport concentration is as follows. Suppose a passenger travels from Baltimore to Los Angeles through Pittsburgh. Despite US Airways’ large market share in Pittsburgh, it has no real market power over the Baltimore–Los Angeles route, for a traveler could easily go from Baltimore to Los Angeles by using, for example, American through Dallas/Fort Worth or United through Chicago. However, a passenger whose final destination is Pittsburgh has little choice but to travel on US Airways. Passengers whose origin or destination is a hub with a single dominant firm may be forced to pay higher fares.

A study by Severin Borenstein found that the average fare per mile is higher for travelers whose origin or destination is the carrier’s major hub. A dominant airline with 70 percent of the traffic might be able to charge a price 2 to 12 percent higher than a rival with only a

\textsuperscript{64} The use of the per se rule and the rule of reason in evaluating claims of violation of antitrust laws is discussed in chapter 5.

\textsuperscript{65} I thank Severin Borenstein for assistance here.

10 percent market share.\footnote{Severin Borenstein, “Hubs and High Fares: Dominance and Market Power in the U.S. Airline Industry,” \textit{RAND Journal of Economics} 20 (Autumn 1989): 344–65.} For the thirty largest airports, table 17.12 provides the airport premium, which is the average ratio of fares on local routes from these airports compared to the national average fares on routes of the same distance. The airport premium measures the additional fare that must be paid for traveling on routes for which a hub carrier is likely to have a dominant position. For example, travelers using Chicago’s O’Hare for local travel paid a premium of almost 15 percent. In Minneapolis–St. Paul, where Northwest controls 82 percent of enplanements, the premium is 31.5 percent. The evidence is that market dominance does result in abnormally high fares for some consumers.

Table 17.12
Hubbing and Fares at the Thirty Largest U.S. Airports

<table>
<thead>
<tr>
<th>Airport</th>
<th>Percent Changing Planes</th>
<th>Airport Fare Premium (percent)</th>
<th>Rank by Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Charlotte</td>
<td>75.7</td>
<td>18.8</td>
<td>20</td>
</tr>
<tr>
<td>Atlanta</td>
<td>69.0</td>
<td>17.2</td>
<td>3</td>
</tr>
<tr>
<td>Memphis</td>
<td>67.7</td>
<td>27.4</td>
<td>29</td>
</tr>
<tr>
<td>Dallas/Ft. Worth</td>
<td>65.8</td>
<td>20.5</td>
<td>2</td>
</tr>
<tr>
<td>Pittsburgh</td>
<td>62.1</td>
<td>15.9</td>
<td>16</td>
</tr>
<tr>
<td>Salt Lake City</td>
<td>61.3</td>
<td>19.1</td>
<td>28</td>
</tr>
<tr>
<td>St. Louis</td>
<td>56.2</td>
<td>-4.0</td>
<td>13</td>
</tr>
<tr>
<td>Chicago, O’Hare</td>
<td>55.7</td>
<td>14.8</td>
<td>1</td>
</tr>
<tr>
<td>Denver</td>
<td>54.1</td>
<td>15.3</td>
<td>7</td>
</tr>
<tr>
<td>Minneapolis</td>
<td>51.0</td>
<td>31.5</td>
<td>15</td>
</tr>
<tr>
<td>Houston, Intercontinental</td>
<td>49.5</td>
<td>15.6</td>
<td>19</td>
</tr>
<tr>
<td>New York, Kennedy</td>
<td>47.3</td>
<td>2.9</td>
<td>6</td>
</tr>
<tr>
<td>Detroit</td>
<td>43.6</td>
<td>-0.7</td>
<td>11</td>
</tr>
<tr>
<td>Baltimore</td>
<td>40.5</td>
<td>9.1</td>
<td>26</td>
</tr>
<tr>
<td>Phoenix</td>
<td>33.1</td>
<td>-28.4</td>
<td>9</td>
</tr>
<tr>
<td>Miami</td>
<td>31.0</td>
<td>-14.3</td>
<td>14</td>
</tr>
<tr>
<td>Seattle</td>
<td>27.3</td>
<td>8.7</td>
<td>24</td>
</tr>
<tr>
<td>San Francisco</td>
<td>25.3</td>
<td>-1.5</td>
<td>5</td>
</tr>
<tr>
<td>Los Angeles</td>
<td>25.2</td>
<td>-5.3</td>
<td>4</td>
</tr>
<tr>
<td>Philadelphia</td>
<td>24.9</td>
<td>11.2</td>
<td>22</td>
</tr>
<tr>
<td>Honolulu</td>
<td>22.4</td>
<td>-20.8</td>
<td>17</td>
</tr>
<tr>
<td>Newark</td>
<td>19.6</td>
<td>11.5</td>
<td>12</td>
</tr>
<tr>
<td>Las Vegas</td>
<td>18.9</td>
<td>-27.8</td>
<td>23</td>
</tr>
<tr>
<td>Houston, Hobby</td>
<td>17.5</td>
<td>-23.4</td>
<td>30</td>
</tr>
<tr>
<td>Orlando</td>
<td>16.8</td>
<td>-15.6</td>
<td>21</td>
</tr>
<tr>
<td>Boston</td>
<td>13.8</td>
<td>9.0</td>
<td>10</td>
</tr>
<tr>
<td>Washington, D.C., National</td>
<td>11.1</td>
<td>10.7</td>
<td>18</td>
</tr>
<tr>
<td>Tampa</td>
<td>11.0</td>
<td>-12.4</td>
<td>27</td>
</tr>
<tr>
<td>San Diego</td>
<td>6.6</td>
<td>-18.1</td>
<td>25</td>
</tr>
<tr>
<td>New York, La Guardia</td>
<td>6.2</td>
<td>9.5</td>
<td>8</td>
</tr>
</tbody>
</table>

The preceding analysis used an intermarket approach in that it compared fares across routes that varied in the degree of concentration. One can also use an intertemporal approach by examining how fares along a particular route change as concentration changes. This approach was implemented by examining how airline mergers which increased concentration along some routes affected air fares. An analysis of twenty-seven mergers over 1985–1988 revealed that the merging airlines, on average, raised airfares by 9.4 percent more on those routes for which they both previously provided service. Furthermore, rival airlines on those routes responded by increasing their air fares by 12.2 percent. The majority of the evidence supports the hypothesis that increasing concentration translates into higher fares.

**Lessons from Regulation and Deregulation**

An important lesson to be learned from the deregulation of the airline industry is that if we are to realize the potential gains from deregulation, then regulatory policy and antitrust policy must work together. Dismantling regulation creates a void that needs to be filled by antitrust policy. It is important to understand that deregulation does not eliminate a role for government. Rather, it means a change in the role for government from one of controlling the industry to one of maintaining a competitive environment.

Although the trucking and airline industries were subject to similar regulatory practices, it is clear that the effects of regulation were quite distinct. In the case of trucking, firms did not vigorously engage in nonprice competition, and therefore they could earn substantial above-normal profits. In contrast, airlines competed away most of the potential profits that regulation created through abnormally high fares. They did so through the provision of frequent flights, low load factors, and high-quality on-board services. Why did regulation induce excessive nonprice competition in the airline industry but not in the trucking industry? One possible answer is that the demand for passenger air service is more responsive to nonprice factors than is the demand for surface freight service. If demand is very elastic with respect to these factors, a firm has a strong incentive to invest in them, since a relatively small investment can result in a sharp rise in firm demand. Of course, if all firms do so, then there may be a lot of canceling out of these effects, so that everyone’s demand is only a little higher and considerable expenditure has been incurred.

There are at least two important lessons to be learned from CAB regulation of passenger air service. First, competition will follow the path of least regulatory resistance. Because no authority can fully control every dimension of competition, firms will compete via the least regulated dimensions. In the case of airline regulation, the CAB prevented fare competition but left uncontrolled competition over quality of service. As a result, regulation raised fares.

---

and raised the quality of service. A second lesson to be learned is that it is very difficult to predict the effects of regulation. Because competition entails coming up with new ways to provide a better product at a lower cost, it is typically beyond the creative abilities of a few economists to anticipate innovations that might have arisen in the absence of regulation. Although economists long predicted the welfare gains from deregulation, they were unable to anticipate the development of the hub-and-spoke system, a development stifled by the entry restrictions put in place by the CAB. As a result of the hub-and-spoke system, deregulation has reduced fares in many markets, but it has not reduced quality of service as much as was anticipated.

Summary

The transportation industry is an informative case study for understanding the effects of regulation. There are several lessons to be stressed. The regulation of surface freight transportation highlights the imperialistic tendencies of regulation. It is difficult enough to regulate an industry; it is even more difficult to regulate only part of an industry. This is what the ICC discovered when it regulated the railroads but not the motor carriers. With rail rates being set above cost, truckers were free to set rates so as to undercut the railroads and take demand that would be more efficiently served by the railroads. Due to the difficulties inherent in such a situation, the Motor Carrier Act of 1935 extended ICC control from railroads to motor carriers.

A second lesson highlighted by the airline industry is that when regulation controls price, firms will find other ways to compete. The implication is that firms engage in nonprice competition. In the case of the airline industry, firms provided high-quality service at a high price. Consumers valued the higher quality resulting from regulation, but they were hurt by the absence of a low-priced, low-quality alternative.

A third lesson is that regulation has a tendency to maintain inefficient practices. This may entail railroads being prevented from abandoning unprofitable lines or allowing inefficient truckers to continue to operate. Restricting freedom of entry and exit destroys one of the most important features of competition: the efficient survive and the inefficient perish.

The final and perhaps most important lesson is that regulation can reduce welfare in ways that are difficult to anticipate. Regulation can stifle innovations that would have occurred in an unregulated market. It was found for both the railroad and airline industries that regulation reduced growth in factor productivity. A case in point is the development of the hub-and-spoke system since airline deregulation. As a result of this unanticipated restructuring of the industry, deregulation allowed for lower fares and improved flight frequency.
Questions and Problems

1. How would one measure the effect of regulation on the profitability of the trucking industry? Did regulation cause profits to go up or down?

2. Why were the railroads in favor of regulation in the 1880s but in favor of deregulation in the 1950s?

3. The Staggers Act gave railroads considerable freedom in setting rates except on routes where there was market dominance. From the perspective of a social planner, how would you choose to define market dominance?

4. During the years of CAB regulation, the unit cost for U.S. airlines declined 3 percent annually, whereas the unit cost for non-U.S. airlines declined 4.5 percent. What are the ways in which regulation might have caused smaller productivity gains?

5. Was the deregulation of the railroad industry a Pareto improvement? How about the deregulation of the trucking industry? If these were not Pareto improvements, who gained and who lost from deregulation?

6. What was the role of antitrust policy after airline deregulation? What do you think should be the role of antitrust policy after an industry is deregulated?

7. In both the airline and trucking industries, rates were generally set too high on many routes. Why did nonprice competition tend to compete away most of the rents in the airline industry but not in the trucking industry? Was this nonprice competition welfare-improving?

8. Have all consumers been made better off by airline deregulation?

9. Why did the deregulation of prices cause wages in the airline and trucking industries to fall?

10. Using the economic theories of regulation from chapter 10, can one explain the deregulation of the transportation industry?

11. What are the different ways in which to estimate the effect of regulation on air fares?

12. In response to truckers taking away an increasing amount of business from the railroads, the trucking industry was regulated in 1935. Would it have been better to have left the trucking industry deregulated? Would it have been better to have deregulated the railroad industry?
At least since the industrial revolution, the energy industry has been critical in the world economy. Energy is an essential input in almost all production processes, whether it is the use of gasoline by truckers to transport commodities, the use of electricity to run computers in an office, the use of natural gas to heat residential homes, or the use of coal to generate electricity by a power plant. Due to its importance to economic activity, the price of energy and the availability of energy sources can have substantial ramifications on an economy.

Not only is there considerable diversity in the sources of energy, their relative importance has changed significantly over time. Since 1850, the major sources of energy in the United States have been wood, coal, crude oil, natural gas, hydroelectric power, and nuclear power. Before 1900, wood and coal were the dominant sources of energy. During the period 1900–1950, wood became quite insignificant, while oil and gas joined coal as the central energy sources. Since 1950, oil and gas have increasingly replaced coal, while nuclear power has entered the market as a new energy source. Because these products are very different in a physical sense, the standard unit of measurement for energy is the British thermal unit (BTU), where one BTU equals the amount of heat required to raise the temperature of one pound of water by one degree Fahrenheit.

In considering the energy market, it is important to recognize that it is inherently international in scope. Because coal and oil are relatively inexpensive to transport, there are significant international flows of these energy sources. Because gas is more expensive to transport, there are considerably smaller flows of natural gas across countries. The international nature of the energy market creates a role for government policies, and therefore countries as well as producers and consumers are players in the world market for energy.

The objective of this chapter is to review and evaluate the economic regulation of the U.S. energy market. Our focus will be on government intervention in the crude oil and natural gas industries. State and federal regulation of the production of crude oil dates back to the early part of the twentieth century, whereas federal regulation of crude oil prices took place during the 1970s. Although the natural gas market has been subject to federal regulation since the 1930s, price regulation did not take place until the mid-1950s.

As we will observe, the regulatory regimes for the oil and gas industries were quite distinct, but they are common in that both entailed the establishment of price ceilings that acted to constrain the price that oil and gas producers could charge for their products. Recall that a similar pricing constraint was placed on the railroad industry during its time of regulation. However, a critical distinction exists between railroad price regulation and energy price regulation. The railroad was considered to be a common carrier, so that it was required by the Interstate Commerce Commission to meet all demand at the government-established price.

2. For a discussion of the regulation of electric power generation, see chapter 12.
In contrast, oil and gas producers were under no such obligation. As a result, price ceilings generated shortages in the oil and gas markets as consumers demanded more units than firms were willing to supply.

In the next section, we briefly review the welfare implications of economic regulation that is characterized by price being set below the competitive equilibrium price and in which firms’ supply decisions are left unregulated. With this foundation, the regulation of the domestic crude-oil-producing industry is examined. In addition to reviewing the implications of federal price regulation, we will investigate the state regulation of production through what is called prorationing. Our analysis will then turn to assessing the implications of price regulation in the natural gas industry.

The Theory of Price Ceilings

The objective of this section is to provide an analysis of the welfare implications of price ceilings. Once we review the regulatory practices used in the oil and gas industries, it will become apparent that price regulation was considerably more complex than the regulatory structure analyzed in this section. Nevertheless, our analysis should be relevant to more complex systems of price regulation as long as they involve some form of a price ceiling.

Consider the market for a product, and let us suppose that in the absence of any government regulation it is competitive. In figure 18.1 the market supply function, \( S(P) \), is shown along with the market demand function, \( D(P) \). At a competitive equilibrium, the market clears at a price of \( P^* \), with \( Q^* \) units being sold. Now impose a price ceiling on this market; that is, the government restricts firms to setting price no higher than the ceiling. If the price ceiling is at least as great as \( P^* \), then government regulation is nonbinding. Given current supply and demand conditions, firms and consumers have no desire to change their decisions so as to move price toward the price ceiling and away from \( P^* \). Of course, if market demand shifted out or market supply shifted in, then the market-clearing price would rise, in which case the price ceiling might become binding.

To consider the case of a binding price ceiling, suppose it is set at \( \bar{P} \) and, in addition, \( \bar{P} \) is less than the competitive equilibrium price \( P^* \) (see figure 18.1). At a price of \( \bar{P} \), market demand of \( D(\bar{P}) \) exceeds \( S(\bar{P})(=\bar{Q}) \) which is the amount that firms are willing to supply at a price of \( \bar{P} \). The implication of this price ceiling is that output is reduced from \( Q^* \) to \( \bar{Q} \). In contrast to the case of unfettered competition, consumers gain the rectangle \( P^*df\bar{P} \) from paying \( P^* - \bar{P} \) less per unit on \( \bar{Q} \) units. However, consumers lose surplus measured by triangle \( bcd \) as \( Q^* - \bar{Q} \) less units are supplied. Thus, the net gain to consumers is the difference between rectangle \( P^*df\bar{P} \) and triangle \( bcd \). Firms clearly lose by the imposition of a price ceiling. In its absence, producer surplus is the triangle \( P^*cg \). With the price ceiling, producers lose surplus of triangle \( dcf \) on the reduced supply of \( Q^* - \bar{Q} \) and lose rectangle \( P^*df\bar{P} \).
Figure 18.1
The Effects of a Price Ceiling
to consumers through the lower price. Summing the change in consumer and producer surplus, the net effect of the price ceiling is a welfare loss measured by the shaded triangle $bcf$.

In confining the welfare loss to the triangle $bcf$, an implicit and important assumption is made. We have assumed that the $\bar{Q}$ units supplied by the market go to the consumers that value them the most, that is, the consumers whose demand is represented by the demand curve from 0 to $\bar{Q}$. Because there is excess demand of $D(\bar{P}) - \bar{P}$ how the $\bar{Q}$ units are allocated among consumers is important from a welfare perspective.

To analyze this issue in a simple manner, assume that each consumer wants to buy at most one unit. Consumers differ by their reservation price, which is defined to be the highest price that a consumer is willing to pay for the good. That is, at a price for the good equal to his reservation price, a consumer is indifferent between purchasing the good at that price and not purchasing the good. Obviously, if the market price is less than his reservation price, a consumer strictly prefers to buy the good. According to this formulation, we can then interpret the market demand function as stating that there are a total of $D(P)$ consumers with a reservation price at least as great as $P$.

Now suppose that the $\bar{Q}$ units are randomly allocated to consumers, perhaps by government mandate. At a price $\bar{P}$, a total of $D(\bar{P})$ consumers want to buy the available $\bar{Q}$ units. Therefore, only a fraction $\bar{Q}/D(\bar{P})$ of consumers who want to buy the good at a price of $\bar{P}$ are able to do so. With random allocation of the $\bar{Q}$ units, this means that, for example, only $[\bar{Q}/D(\bar{P})]D(P')$ of the $D(P')$ consumers with a reservation price at least as high as $P'$ will be able to purchase the good. The allocation rule is then depicted as the curve $[\bar{Q}/D(\bar{P})]D(P)$ in figure 18.2.

Let us now reexamine the welfare loss due to the imposition of a price ceiling. In addition to the welfare loss of triangle $bcf$ resulting from reduced supply, there is a welfare loss of triangle $abf$ resulting from the limited supply not being allocated to those consumers who value the good the most. For the good to be properly allocated, the $\bar{Q}$ units should go to those consumers with reservation prices at least as high as $P^\circ$. Their surplus equals trapezoid $abf$.

Alternatively, with random allocation of the $\bar{Q}$ units, only a fraction $\bar{Q}/D(\bar{P})$ of those consumers get to buy the good. The remaining units are allocated to consumers who value them less—specifically, those consumers with reservation prices ranging from $\bar{P}$ to $P^\circ$. Because now the consumers’ surplus is only triangle $af\bar{P}$, random allocation results in an additional welfare loss of triangle $abf$ and a total welfare loss of triangle $acf$.

An important implicit assumption in the preceding analysis is that consumers cannot resell the good (or, alternatively, cannot resell the right to buy the good). If they are allowed to do so, then the welfare loss $abf$ may be avoided. Consumers fortunate enough to be allocated the good can resell it to the highest bidder through the secondary market. If the costs of engaging in this transaction are small, then the units should end up in the hands of those consumers who value them the most, regardless of to whom they were
Figure 18.2
The Effects of a Price Ceiling with Random Allocation
initially given. In that case, the only effect of random allocation is to distribute surplus
to those consumers who are lucky enough to be given property rights over this valuable
commodity.

Suppose instead that the allocation of the $\bar{Q}$ units is determined not by random allocation
but rather by the suppliers. One would expect consumers to compete for the $\bar{Q}$ units and for
those consumers with the highest reservation prices to spend the most in trying to get them.
This might entail providing some form of bribe to a firm, or perhaps simply waiting in line.
Competitive activity for the purpose of gaining access to products in short supply was regu-
larly observed in the former Soviet Union for standard household items and in the United
States for such items as concert tickets. With competition being the mechanism for allocat-
ing the good, there is no longer the welfare loss that occurred from random allocation.
However, there is instead a different type of welfare loss. If consumers use real resources to
compete for these goods, then the value of those resources to society represents an additional
welfare loss from price regulation. If instead consumers compete by using pecuniary
resources, such as paying a financial bribe to a firm, then there is only a transfer and thus no
associated welfare loss.

Two basic points are to be derived from this analysis. First, the imposition of a binding
price ceiling reduces social welfare by decreasing the amount exchanged in the market.
Second, in light of the excess demand, how the good is allocated to consumers can create
additional welfare losses. The good may not end up with those consumers who value it the
most, and/or consumers may use resources in competing with one another for the right to
purchase the good at the government-restricted price.

Price and Quantity Regulation of the Crude Oil Industry

Although the first oil well in the United States dates from 1859, oil did not become an impor-
tant energy source until the early twentieth century. Even as late as 1920, oil made up only
11 percent of energy consumption. The term *petroleum* is often equated with crude oil, but
it actually refers to all natural hydrocarbons except for those in the coal family. In particu-
lar, petroleum includes natural gas as well as crude oil. In this section, we will focus on the
regulation of the production and pricing of domestic crude oil.

Technological Structure

The oil industry comprises three divisions: production, refining, and distribution. Since oil is
normally found in underground reservoirs, the first step in production is exploration to dis-

3. See chapter 12.
cover an oil reservoir. The second step is development of the reservoir, which entails drilling a hole that creates a low-pressure point in the reservoir, which then forces oil up to the surface. The third and final step in production is extraction of oil from the reservoir. On its extraction, the next step in the process is the refining of crude oil into usable products. Finally, refined oil is distributed to retailers and consumers, where distribution entails both the transportation and marketing of the product. Many firms in the oil industry are vertically integrated in that they perform production, refining, and distribution.

**Economic Background**

There are several notable features to the market for crude oil. Because of the relatively low cost of transportation and the geographic concentration of oil reserves, the market for oil is international in scope. As a result, major players include not only firms and consumers but also governments. Historically, two particularly important players are, on the demand side, the United States and, on the supply side, the Organization of Petroleum Exporting Countries (OPEC).

A second important characteristic of the oil market is that it has a rather volatile history. There are several sources of volatility. First, energy demand moves with the business cycle. When an economy moves into a recession, demand for manufactured goods falls, so that inventories build up. In response to increased inventories, manufacturing production slows down. Because energy is an input in manufacturing, demand for energy then falls as an economy moves into a recession. By the same logic, an economic boom raises the demand for energy inputs. Thus, one source of volatility in the energy market is the sensitivity of energy demand to the business cycle. Of course, this is by no means sufficient to explain all volatility, because many products are sensitive to the business cycle yet are not as volatile in terms of price and supply as is the market for energy.

On the supply side, an important source of volatility is that new discoveries of oil reserves can cause quite drastic increases in the size of known reserves. Typically, the result is an upward shock in supply and a downward shock in price. The geographic concentration of oil reserves creates the opportunity for short-run supply shocks due to disruptions in production and/or transportation.

**OPEC**

Even the briefest of descriptions of the oil industry must cover the greatest cartel of our time—OPEC. Formed in September 1960, OPEC initially comprised the major oil-exporting nations of Iran, Iraq, Kuwait, Saudi Arabia, and Venezuela. In addition to these five

---

countries, OPEC consists today of Algeria, Indonesia, Libya, Nigeria, Qatar, and the United Arab Emirates.

Generally, the OPEC cartel has functioned by establishing a price for crude oil and then allocating output quotas among its members so as to maintain this price. The key member of the cartel is Saudi Arabia. As the member with the largest oil reserves, Saudi Arabia has been the “swing producer” whose responsibility was to adjust its production in order that the cartel price be achieved. Saudi Arabia has been regularly called upon to produce below its quota in order to compensate for overproduction by its fellow OPEC members. However, after persistent cheating by cartel members throughout the 1980s, Saudi Arabia eventually chose to forsake its role as the swing producer.

As a cartel, OPEC was not very successful until the early 1970s. At the beginning of October 1973 the posted price of Saudi Arabian light crude oil was $2.80 per barrel. However, by October 16 the price had jumped to $4.76 per barrel. This dramatic price increase was caused by an Arab embargo on oil shipments to the United States as a result of the fourth Arab–Israeli war. But the worse was yet to come for the oil-importing nations. An even greater price shock occurred in January 1974, when the price of crude oil rose to $10.84 per barrel. In the span of less than four months, the price of crude oil had increased almost fourfold.

Throughout much of the 1980s, OPEC was subject to vast overproduction by its members, with an ensuing decline in the price of crude oil. Also weakening its position has been the discovery of oil reserves outside of OPEC, including in the North Sea. As of 2002, only 38 percent of crude oil production was by OPEC producers, which compares to a high of 52 percent in 1973. Still, OPEC retains some market power, though it is clearly not the force it was in the 1970s.

Regulatory History

Beginning with the Standard Oil case in 1911, there is a long history of government involvement in the U.S. oil industry. Of course, that case concerned the oil-refining industry, and the focus of this chapter is on the oil-producing industry. The consumers in our study are made up of domestic refining companies, whereas the producers are domestic and foreign firms that extract oil from wells. To provide an overview, state and federal regulation from the early part of this century to about 1970 was designed to limit the supply of crude oil. This was achieved through restrictions on the production of domestic firms and on exports to the United States by foreign oil producers. From 1970 to 1981, federal regulation switched from this pro-producer stance to one less favorable to the oil industry as regulation constrained the price that domestic oil companies could charge for their product.

Prorationing

The goal of regulating domestic oil production dates to 1909, when Oklahoma limited the production of wells. Similar powers were given to the Railroad Commission in Texas in 1919. However, not until 1928 were the first individual field proration orders issued. These orders limited production by allocating the total production allowed pro rata among wells. Such a mechanism is referred to as prorationing.

Events in 1926–1931 spurred a wave of state intervention in the production of oil. One critical event was the discovery of new reserves and, in particular, the East Texas oil field in 1930. This was a 5.5-billion-barrel reservoir that by 1933 had 1,000 firms with a total of 10,000 wells pumping oil from it.6 Because of this massive increase in supply, along with the reduction in oil demand resulting from the Great Depression, oil prices fell sharply. Pursuing the lead of Oklahoma, Texas instituted prorationing in 1930 and Kansas followed in 1931. The federal government aided the oil-producing states by passing the Connally “Hot Oil” Act of 1935, which prohibited the interstate shipment of oil that was extracted in violation of state regulations.

In addition to prorationing, oil production was restricted through the requirement of minimum spacing between oil wells. States typically required at least twenty acres per well, while the federal government, owing to the need for conservation during World War II, required a minimum of forty acres. After World War II, an additional twenty-two states added conservation laws that gave state government the right to restrict oil production.

Mandatory Oil Import Program

Before 1957 the United States had no controls on the import of crude oil. In that year, motivated by rising imports, President Dwight Eisenhower called for a voluntary reduction by domestic oil refiners. Previously, the oil-producing states had responded to an increase in oil imports by reducing prorationing orders so as to maintain the domestic price. Needless to say, this policy of voluntary restraint did not work. The program then became involuntary in 1959 with the Mandatory Oil Import Program (MOIP).

Oil Price Controls

With rising inflation, President Richard Nixon instituted an economy-wide price freeze in August 1971. Two years after this freeze, Phase IV decontrolled all prices with the exception of crude oil. These controls were set to expire in April 1974, but oil price regulation continued unabated until President Ronald Reagan, in his first act, decontrolled oil prices. A synopsis of crude oil price controls from 1971 to 1988 is provided in table 18.1.

By the early 1970s the other two forms of regulation had become extinct. States had
stopped restricting domestic oil production, and the MOIP was ended in 1973. In November
1973, Phase IV was replaced with the Emergency Petroleum Allocation Act (EPAA). It insti-
tuted price ceilings on oil and an entitlements program for allocating this price-controlled oil
to refiners. To be able to process one barrel of price-controlled domestic oil, a refiner needed
one entitlement. Control over oil prices moved from the Cost of Living Council to the Federal
Energy Administration (FEA) in May 1974.

Although President Gerald Ford favored decontrol of oil prices, the Energy Policy Con-
servation Act (EPCA) went into effect in December 1975. This program rolled back some oil
prices, but it called only for gradual decontrol starting in early 1976. Interested in further
decontrolling oil prices, President Jimmy Carter put forth a plan of gradual decontrol from
June 1979 to September 1981, at which time EPCA was to expire. Concerned with the trans-
fer of wealth from consumers to domestic oil producers under decontrol (as prices would
presumably rise), President Carter instituted the Crude Oil Windfall Profits Tax of 1980. In
actuality, this was not a profits tax but rather an excise tax on oil. Ahead of the schedule out-
lined by President Carter, President Reagan lifted all remaining oil price controls in January
1981.

Table 18.1
Crude Oil Price Controls, 1971–1988

<table>
<thead>
<tr>
<th>Program</th>
<th>Period</th>
<th>Price Regulations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Economic Stabilization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phase I</td>
<td>8/71 to 11/71</td>
<td>Economy-wide price freeze.</td>
</tr>
<tr>
<td>Phase II</td>
<td>11/71 to 1/73</td>
<td>Controlled price increases to reflect cost increase with profit limitations.</td>
</tr>
<tr>
<td>Phase III</td>
<td>1/73 to 8/73</td>
<td>Voluntary increases up to 1.5% annually for cost increases.</td>
</tr>
<tr>
<td>Special Rule No. 1</td>
<td>3/73 to 8/73</td>
<td>Mandatory controls for 23 largest oil companies.</td>
</tr>
<tr>
<td>Phase IV</td>
<td>8/73 to 11/73</td>
<td>Two-tier pricing: old oil at level of 5/15/73 plus $0.35, new oil, stripper oil,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>and “released” oil uncontrolled.</td>
</tr>
<tr>
<td>Reaction to Shortage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emergency Petroleum Allocation Act (EPAA)</td>
<td>11/73 to 12/75</td>
<td>Same as Phase IV plus entitlements program.</td>
</tr>
<tr>
<td>Energy Policy and Conservation Act (EPCA)</td>
<td>12/75 to 9/81</td>
<td>Lower-tier (old) oil at $5.25, upper-tier (new) oil at $11.28, stripper oil decontrolled (9/76), composite price at $7.66, provision for inflation increases.</td>
</tr>
<tr>
<td>Compromising Decontrol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Windfall Profits Tax</td>
<td>3/80 to 10/88</td>
<td>Tax on difference between controlled prices and market price.</td>
</tr>
</tbody>
</table>

Oil Prorationing

Regulatory Practices

In that regulatory procedures varied across states, the following is representative only of the process by which oil was prorated. To begin, each oil well in a state is assigned a maximum allowable rate of production. Let $Q^e$ denote the sum of these maximum allowable rates across all wells. For each month, the prospective market demand is estimated at the current price for oil. Adjusting this demand projection for anticipated changes in inventories, the resulting number is the target rate of production. Let us denote it $Q^t$. If $Q^t$ exceeds $Q^e$, then each well is restricted to producing no more than its maximum allowable rate. If instead $Q^e$ exceeds $Q^t$, then total oil production in the state is restricted to $Q^t$. In that case, $Q^t$ is allocated in the following manner. First, production is allocated to those special wells that are exempt from prorationing. Let the total production of these wells be denoted $Q^e$. The remainder, $Q^t - Q^e$, is allocated to nonexempt wells in proportion to their respective maximum allowable rates. For example, a well with an allowable rate of $Q^o$ would be allowed to produce $(Q^o/Q^e)(Q^t - Q^e)$. Table 18.2 shows the average percentage of the maximum allowable rates in several oil-producing states during 1948–1966. Prorationing clearly imposed severe restrictions on oil production.

Rationale for Prorationing

State Government as a Cartel Manager

One interpretation of prorationing is that it is a scheme whereby the state acts as a cartel manager. By restricting production, the price of oil is kept above the competitive level, which in principle allows above-normal profits to be earned by the oil companies. Because the loss in consumer surplus from output restrictions is distributed over the entire United States but the gain in firm profits is largely concentrated in the oil-producing states, one could imagine a policy of prorationing to be in the welfare interests of an individual state, though certainly not in the interests of the United States as a whole.

Common Pool Problem

An alternative rationale, and one with quite different welfare implications, is that prorationing solves a common pool problem. A common pool problem arises when two or more

individuals share property rights over some resource. As we will explain, these common property rights can result in the inefficient use of a resource.

To begin, let us consider the case of a resource that is owned by a single individual. Consider a newly discovered oil reservoir, and suppose it is entirely contained within the landholdings of a single individual. The landowner (or the company to which he has assigned mineral rights) has to decide how fast to extract the oil. For simplicity, suppose there are just two periods, period 1 (today) and period 2 (tomorrow). We will need to introduce a little notation. Let \( P_t \) and \( Q_t \) denote the price and extraction rate, respectively, in period \( t \). The variable \( MC_t(Q) \) denotes the marginal cost of extracting at a rate \( Q \) in period \( t \).

Suppose that the landowner is initially thinking about pumping \( Q_1 \) barrels of oil today and \( Q_2 \) tomorrow. If he considers pumping a little more oil today, the change in today’s profit is \( P_1 - MC_1(Q_1) \). The landowner sells that additional barrel for \( P_1 \), and it costs him \( MC_1(Q_1) \) to extract it. In addition, extracting more oil today has an impact on tomorrow’s profit. The technology of oil reservoirs is such that pumping at a faster rate reduces the amount of oil that

Table 18.2
Annual Average of Monthly Market Demand Factors (Percentage)

<table>
<thead>
<tr>
<th>Year</th>
<th>Texas</th>
<th>Louisiana</th>
<th>New Mexico*</th>
<th>Oklahoma</th>
</tr>
</thead>
<tbody>
<tr>
<td>1948</td>
<td>100</td>
<td>†</td>
<td>63</td>
<td>‡</td>
</tr>
<tr>
<td>1949</td>
<td>65</td>
<td>†</td>
<td>61</td>
<td>‡</td>
</tr>
<tr>
<td>1950</td>
<td>63</td>
<td>†</td>
<td>69</td>
<td>‡</td>
</tr>
<tr>
<td>1951</td>
<td>76</td>
<td>†</td>
<td>74</td>
<td>‡</td>
</tr>
<tr>
<td>1952</td>
<td>71</td>
<td>†</td>
<td>68</td>
<td>‡</td>
</tr>
<tr>
<td>1953</td>
<td>65</td>
<td>90</td>
<td>63</td>
<td>‡</td>
</tr>
<tr>
<td>1954</td>
<td>53</td>
<td>61</td>
<td>57</td>
<td>‡</td>
</tr>
<tr>
<td>1955</td>
<td>53</td>
<td>48</td>
<td>57</td>
<td>60</td>
</tr>
<tr>
<td>1956</td>
<td>52</td>
<td>42</td>
<td>56</td>
<td>53</td>
</tr>
<tr>
<td>1957</td>
<td>47</td>
<td>43</td>
<td>56</td>
<td>52</td>
</tr>
<tr>
<td>1958</td>
<td>33</td>
<td>33</td>
<td>49</td>
<td>45</td>
</tr>
<tr>
<td>1959</td>
<td>34</td>
<td>34</td>
<td>50</td>
<td>41</td>
</tr>
<tr>
<td>1960</td>
<td>28</td>
<td>34</td>
<td>49</td>
<td>35</td>
</tr>
<tr>
<td>1961</td>
<td>28</td>
<td>32</td>
<td>49</td>
<td>31</td>
</tr>
<tr>
<td>1962</td>
<td>27</td>
<td>32</td>
<td>50</td>
<td>35</td>
</tr>
<tr>
<td>1963</td>
<td>28</td>
<td>32</td>
<td>54</td>
<td>31</td>
</tr>
<tr>
<td>1964</td>
<td>28</td>
<td>32</td>
<td>54</td>
<td>28</td>
</tr>
<tr>
<td>1965</td>
<td>29</td>
<td>33</td>
<td>56</td>
<td>27</td>
</tr>
<tr>
<td>1966</td>
<td>34</td>
<td>35</td>
<td>65</td>
<td>38</td>
</tr>
</tbody>
</table>

* Southeast area only.
† No fixed allowable schedule.
‡ Comparable data not available.


8. The ensuing analysis is from McDonald, *Petroleum Conservation*, chap. 5.
can ultimately be recovered. If extraction is too fast, then there is a loss in subsurface pressure, so that pockets of oil become trapped. Less oil can then be retrieved from the well. Let \( b \) denote the number of units of oil that cannot be extracted tomorrow because an additional barrel is extracted today. The discounted loss in tomorrow’s profit from extracting a little more today is then \( b(1/(1+r))(P_2 - MC_2(Q_2)) \), where \( r \) is the interest rate. In other words, for each additional barrel pumped today, \( b \) fewer barrels can be pumped tomorrow, and each of those barrels represents forgone profit of \( P_2 - MC_2(Q_2) \). This loss is discounted, since it is not incurred until tomorrow. We conclude that the discounted marginal return to pumping another barrel today is

\[
P_1 - MC_1(Q_1) - b[1/(1+r)][P_2 - MC_2(Q_2)].
\]

The landowner can go through the same thought exercise concerning tomorrow’s rate of extraction. By pumping a little more oil tomorrow, he will receive additional revenue of \( P_2 \) at a cost of \( MC_2(Q_2) \), so that his discounted marginal return is \( [1/(1+r)][P_2 - MC_2(Q_2)] \).

Let \( Q_1^* \) and \( Q_2^* \) denote the rates of extraction for today and tomorrow, respectively, that maximize the present value of the landowner’s profit stream. We will argue that these rates must be set so as to equate the marginal return from pumping another barrel today and pumping another barrel tomorrow:

\[
P_1 - MC_1(Q_1^*) - b[1/(1+r)][P_2 - MC_2(Q_2^*)] = [1/(1+r)][P_2 - MC_2(Q_2^*)].
\]

(18.1)

To see that profit maximization requires this equality to hold, suppose that the marginal return from pumping more today exceeded the marginal return from pumping more tomorrow:

\[
P_1 - MC_1(Q_1) - b[1/(1+r)][P_2 - MC_2(Q_2)] > [1/(1+r)][P_2 - MC_2(Q_2)].
\]

By shifting the pumping of one barrel from tomorrow to today, the landowner loses discounted profit of \( [1/(1+r)][P_2 - MC_2(Q_2)] \) but gains discounted profit of

\[
P_1 - MC_1(Q_1) - b[1/(1+r)][P_2 - MC_2(Q_2)].
\]

Because the latter expression is bigger, the net change in the present value of the profit stream is positive. But if this conclusion is true, then the original extraction rates must not have been optimal. Therefore, if the marginal returns from pumping more today and more tomorrow are not equated, the landowner can shift extraction between the two periods so as to increase the present value of his profit stream. We conclude that profit maximization requires that those marginal returns be equated, as expressed in equation 18.1. Furthermore, because \( P_t \) is presumed to measure the marginal social benefit from a unit of oil in period \( t \), profit maximization also achieves the rates of extraction that maximize social welfare.

Where the common pool problem arises is if the oil reservoir spans the property of two or more individuals. According to U.S. law, property rights over the oil reservoir are determined
by the rule of capture. This rule states that any extracted oil belongs to the landowner who captures it through a well on his land. Thus, if an oil reservoir spans several properties, there are several individuals who have the right to extract oil from it. To capture the effect of other landowners extracting oil from a common reservoir, let $x$ be the fraction of a barrel drained by neighbors by postponing pumping another unit today. That is, if a landowner considers pumping one less barrel today, he will not find that entire barrel in the reservoir tomorrow. Only a fraction $1 - x$ of that barrel would remain, for his neighbors would have extracted the other fraction $x$ of it. The condition for profit maximization is no longer (18.1) but instead the following expression:

$$P_1 - MC_1(Q_1) - (1 - x)b[1/(1 + r)][P_2 - MC_2(Q_2)] = (1 + x)[1/(1 + r)][P_2 - MC_2(Q_2)]$$

where $Q_1$ and $Q_2$ denote the new profit-maximizing rates of extraction. Note that if $x = 0$ (as is true if there is no common pool problem), (18.2) is the same as (18.1). The key implication of $x > 0$ is that a landowner has an incentive to extract at a faster rate today; that is, $Q_1 > Q_1^*$. For every barrel of oil not pumped today, a landowner loses $x$ of that barrel to his neighbors. Because the pool is shrinking over time because of one’s neighbors also drawing from it, an individual landowner has an incentive to speed up extraction so as to acquire oil before it is acquired by the other landowners. Each landowner engages in this practice of fast extraction, so that the overall rate of extraction is higher than when it is owned by one individual.

With a common pool problem such as this one, profit maximization by each landowner results in a rate of extraction that exceeds the socially optimal rate. Each owner is induced to pump more today, relative to the social optimum, because postponing extraction is costly, as the other landowners will drain the field in the meantime. To see this result graphically, we can rewrite equation 18.1 as follows:

$$P_1 = MC_1(Q_1) + SUC$$

where

$$SUC = (1 + b)[1/(1 + r)][P_2 - MC_2(Q_2)].$$

and $SUC$ stands for “social user cost.” The variable $P_1$ is the marginal revenue from pumping one more barrel today, while $MC_1(Q_1) + SUC$ is the marginal cost to society from pumping that additional barrel. It comprises the marginal cost of pumping it, $MC_1(Q_1)$, and $SUC$, which is the future cost to society of pumping another unit today. When there is just one landowner, private user cost is the same as social user cost. It follows that the profit-maximizing rate of extraction then satisfies

$$P_1 = MC_1(Q_1^*) + SUC.$$
This relationship is depicted in figure 18.3. When instead there is a common pool problem, the cost to an individual landowner from extracting one more barrel is less than the cost to society. The reason is that the landowner is concerned about leaving a barrel in the ground because it may be extracted by one’s neighbors. However, society does not care who extracts the oil (assuming that all landowners are equally efficient in extraction). When there is more than one landowner extracting oil from the same reservoir, an individual landowner’s private user cost equals \((1 - x)SUC\), which falls short of social user cost. The profit-maximizing rate of extraction, \(\bar{Q}_t\), then satisfies

\[
P_t = MC_t(\bar{Q}_t) + (1 - x)SUC.
\]

As shown in figure 18.3, the common pool problem results in the rate of extraction exceeding the socially optimal rate.

Figure 18.3
The Effect of Oil Prorationing on the Extraction Rate
Competition did indeed appear to cause excessive early extraction in the Texas oil fields during the 1920s and 1930s. It is estimated that the actual recovery of oil from a particular reservoir was around 20 to 25 percent of total oil reserves. In contrast, with controlled withdrawal, one could have extracted 80 to 95 percent. Of course, this difference is not necessarily evidence that extraction was too fast because the social optimum might have called for quick extraction to the point of low ultimate recovery if the marginal social benefit from a barrel of oil in the early years was high relative to what it was expected to be in later years. However, the lack of variation in price over time and the substantial amount of oil lost would suggest that it is unlikely that such a fast rate of extraction was socially optimal.

Thus far we have shown competition results in excessive extraction for a given number of wells. In addition, owners might find it optimal to build more wells than is socially optimal in order to increase the rate of extraction. For example, the social optimum may require only one well for the reservoir, whereas profit maximization results in each landowner having at least one well. Because the construction of these additional wells uses up real resources, the value of those resources represents an additional welfare loss. If some of the additional oil that is pumped is not sent immediately to market but rather is stored, the cost of storage is another welfare loss, for it is socially optimal to store oil in the ground. However, when a landowner is competing with other landowners for oil from the same reservoir, storage in the ground is costly, as some of it will be extracted by other landowners in the future. It may be more profitable to pump it early and store it above ground.

There are several other implications from the common pool problem. The fact that other firms are extracting at a fast pace reduces a firm’s profits from a well and thus reduces the potential gains from exploration, relative to an environment in which production is controlled in a socially optimal manner. Thus, the common pool problem can reduce the incentive to explore. One can also expect greater volatility in oil prices. When a new discovery is made, all landowners will rush to pump out oil before everyone else does. As a result, there is an upward shock in supply, which causes a downward shock in price. Once the extraction is completed, price then increases again.

Solutions to the Common Pool Problem

Although competition does not result in the social welfare optimum when there is a common pool problem, this fact need not imply a role for government intervention. As it turns out, there are several private mechanisms that may be able to solve the common pool problem. In some ways the simplest solution is to have a single individual own all the land over an oil reservoir (or at least the mineral rights). Of course, this is feasible only if a single landowner

can earn greater profits than the sum of the profits that would be earned by the multiple landowners; only then would an individual be able to buy up the land from the various landowners and turn a profit by doing so. When there is a common pool problem, this is indeed true because a single landowner can coordinate extraction on different lands and thereby earn a higher return. As a result, it is possible for an individual to pay each landowner an amount in excess of what they would otherwise earn and for the total payments made to the landowners to be less than the value of the combined land.

Alternatively, one could achieve the same objective through unitization. Unitization is when one of the owners develops the entire reservoir and the returns are shared by all parties. Of course, parties must agree ex ante as to how to share these returns. A third solution is for the landowners to agree privately to prorationing and thereby limit their production.

In light of all these private solutions, why was there then state intervention? These private solutions sound simple, but they can be rather difficult to implement. First, there are the transaction costs entailed in putting together any contract. If there are several hundred landowners, it will be quite costly for an individual to negotiate to buy several hundred properties. By similar reasoning, unitization and private prorationing will have large transaction costs when there are many landowners. A second problem is getting all of the landowners to agree. Each will probably have different information on and opinions about the value of the reservoir. In addition, some landowners are likely to try to hold out for a higher price for their land or a higher share of profits, in the case of unitization, or a higher share of production, in the case of private prorationing. Such behavior can prevent an agreement from being reached. The costly and difficult nature of many agents coming to an agreement can often prevent these private mechanisms from solving the common pool problem. As a matter of fact, economists have studied attempts at private prorationing in Oklahoma and Texas during 1926–1935. They found that success in contracting was lower, the less concentrated was landownership. This suggests that the costs of reaching an agreement were indeed influential.10

In light of these difficulties, there can be a rationale for government intervention in the production of crude oil. When a common pool problem exists and private solutions are costly to implement, social welfare may be increased by the government imposition of production restrictions.

**Effects of Prorationing**

The evidence on the effects of instituting prorationing is both weak and relatively unsubstantive, but it still provides some insight into the potential welfare implications of this form of state regulation. First, anecdotal evidence has shown that prorationing increased the

ultimate recovery of oil reservoirs. The productive lives of twenty fields in Arkansas, Louisiana, Oklahoma, and Texas were examined. Ten fields were developed before and ten after prorationing. For the pre-prorationing fields, the production rate in the fifteenth year was, on average, 8.6 percent of the peak production rate. In contrast, for the post-prorationing fields, the production rate was 73.9 percent of the peak rate in the fifteenth year.\textsuperscript{11} This evidence is consistent with the predicted effect of prorationing, which is that it reduces early extraction rates and thereby increases ultimate recovery. However, conclusions drawn from this analysis are tentative in that it does not control for differences in the fields and does not compare these rates to the social optimum.

The predicted effect of prorationing on drilling costs is ambiguous. On the one hand, because the maximum allowable production set by the state was directly dependent on the number of wells, there is an incentive to drill more wells. On the other hand, prorationing reduces production rates and thus reduces the need for more wells. The net effect on the number of wells is unclear. However, theory does predict that prorationing would increase the incentive to drill more costly wells, because deeper wells were allocated a higher allowable rate of production and stripper wells, which produce at a low rate and high cost, were exempt from prorationing.

The evidence suggests that prorationing did increase the cost of production. It is believed that prorationing resulted in 23,000 additional wells in the East Texas oil field in the early 1930s.\textsuperscript{12} As a result, the original maximum allowable production of 225 barrels per day set by the Railroad Commission in September 1931 had to be reduced to 37 barrels per day by December 1932 in order to maintain total production at a specified level.\textsuperscript{13} A more systematic study estimated the additional exploration and drilling costs due to prorationing. It found for 1961 that these annual costs were higher by $2.15 billion (in 1961 dollars) as a result of prorationing.\textsuperscript{14}

In evaluating prorationing, it appears to have raised drilling costs by increasing the incentive to drill wells. However, it is likely to have resulted in a more socially preferred rate of extraction as well as a higher degree of ultimate recovery. Unfortunately, the available evidence on the impact of prorationing is sufficiently sparse as to make it difficult to draw any definitive conclusion concerning its net welfare effect.

\textsuperscript{11} Zimmermann, Conservation, p. 286.


Mandatory Oil Import Program

Regulatory Practices

The Mandatory Oil Import Program was put in place in 1959. Although restrictions on oil imports were loosened in 1970, the MOIP was not suspended until April 1973. Officially, the rationale for limiting oil imports was that the United States was becoming too dependent on foreign oil and this dependence created a national security risk. A more likely reason was to prop up domestic oil prices and increase the profits of domestic oil-producing firms. Whether or not this result was intended, it was certainly an outcome of the program.

The MOIP initially restricted crude oil imports to 9 percent of projected domestic demand for oil. Because of the ambiguity of this criterion, the quota was changed to 12.2 percent of domestic oil production in 1962. Crude oil imports were subject to this quota, but residual fuel oil was not. It is notable that the MOIP applied only to oil refiners. Every oil refiner was given an import quota, regardless of its demand for oil imports prior to the MOIP.

Effects of Regulation

In figure 18.4, $D(P)$ denotes the domestic demand curve for oil and $S_d(P)$ the supply curve of domestic oil producers. The world supply curve, denoted $S_w(P)$, is assumed to be horizontal at the world price $P_w$. To consider the effects of the MOIP, we first need to derive the market equilibrium in its absence. With no restrictions on oil imports, the market-clearing price in the U.S. market would be the world price of $P_w$. At any price below that level, no oil would be supplied, as a firm would choose to sell exclusively in the non-U.S. market. At any price at or above that level, there is unlimited supply. Therefore, in the absence of the MOIP, the domestic price is $P_w$. Of the $D(P_w)$ units supplied to the domestic market, domestic oil producers provide $S_d(P_w)$ with oil imports of $D(P_w) - S_d(P_w)$.

Now consider the implementation of the MOIP. For simplicity, let us ignore the fact that residual fuel oil was exempt from the oil quota. In that case, oil imports were allowed to be 12.2 percent of domestic production. Let us further suppose, as was true, that

$$\frac{D(P_w) - S_d(P_w)}{S_d(P_w)} > 12.2,$$

so that the regulatory constraint is binding. In other words, at the unregulated market equilibrium, oil imports exceed the regulatory limit of 12.2 percent of

15. This statement is subject to the caveat of differential transportation costs. A domestic supplier would be willing to sell below the world price in the U.S. market because of the savings in transportation costs.

domestic supply. With excess demand at a price of $P_w$, such a price is inadequate to clear the domestic market. For prices at or above $P_w$, the supply curve for the domestic market is $S(P) = 1.122S_d(P)$, as domestic suppliers supply $S_d(P)$ and, though foreign supply is unlimited at those prices, oil imports are restricted to $0.122S_d(P)$ by the MOIP.

As depicted in figure 18.4, the equilibrium price under the MOIP is $\hat{P}$, which exceeds the world price. Notice that the MOIP has achieved its objective of reducing oil imports, as they have fallen from $D(P_w) - S_d(P_w)$ (or distance $cf$) to $D(\hat{P}) - S_d(\hat{P})$ (or distance $ba$). In assessing the welfare effects of limiting oil imports, first note that consumers are worse off by the sum of rectangle $\hat{P}aeP_w$ and triangle $bcd$. Because demand decreases from $D(P_w)$ to $D(\hat{P})$, consumers lose surplus measured by triangle $bcd$. In addition, consumers have to pay a higher price of $\hat{P}$ on its remaining demand. However, this higher price is paid only for domestic oil so that its expenditure increases by $(\hat{P} - P_w)S_d(\hat{P})$ or rectangle $\hat{P}aeP_w$. Although consumers are made worse off, domestic oil producers receive higher profits because of the increase in

**Figure 18.4**
Equilibrium under the Mandatory Oil Import Program
demand for domestically produced oil. Oil import quotas increase domestic oil producers’ profits from triangle $P_w/g$ to triangle $P_ag$. Thus, domestic producer surplus rises by $P_agP_w$. Summing up the change in consumer surplus and domestic producer surplus, the welfare loss from the MOIP is measured by the sum of triangles $aef$ and $bcd$. Triangle $bcd$ measures the forgone consumer surplus from the fall in demand from $D(P_w)$ to $D(P)$, whereas triangle $aef$ measures the value of wasted resources from having domestic oil producers supply an additional amount of $S_d(P) - S_d(P_w)$ rather than have it supplied more efficiently by importing oil at a price of $P_w$.

Studies have been conducted to estimate the welfare effect of limiting oil imports. To determine the price differential $P - P_w$, one can use the resale price for import quota vouchers because the resale price should reflect the differential between the marginal cost of producing domestically produced oil and the price of importing another barrel of oil. Using this method, $P - P_w$ was estimated to be $1.174$ per barrel for 1969. To give an idea of the size of the differential, the average world price was around $2.10$ per barrel, so that import quotas raised the domestic price by over 50 percent. With this estimated price differential, one can derive a measure of the welfare loss from the MOIP if one has information on the domestic supply and demand curves. One estimate put the cost to consumers in 1960 at $3.2$ billion, from which it steadily rose to $6.6$ billion by 1970.

Crude Oil Price Controls

Regulatory Practices

There are two important aspects to the government regulation of the oil industry beginning in 1971. First, a multitier pricing system was used that established different prices for oil according to the vintage of the well and other characteristics. The main reason for pursuing such a regulatory structure was to enhance incentives for domestic exploration by allowing higher prices for newly discovered oil. The second important aspect of regulatory practice was the entitlements program. This program determined how price-controlled oil was distributed among refiners.

18. Ibid. For comparable estimates, see Burrows and Dommencich, *Analysis.*
Multitier Pricing

In setting price controls, the EPAA delineated two types of oil, which were referred to as “old” oil and “new” oil. Old oil was all oil produced from an existing well at a production rate not in excess of the rate for May 1972. The price of a barrel of old oil was set at the price for May 15, 1973, plus $0.35. New oil included oil from new fields, oil produced in excess of 1972 levels from existing fields, and oil from properties that averaged less than ten barrels per day. New oil was free of price controls. Furthermore, for each barrel of new oil produced, one barrel of old oil was released from price controls.

The pricing system became a bit more complicated under the EPCA in that a three-tier system was instituted. The lower tier applied to oil produced below some base production control level (BPCL), whereas the upper tier applied to oil produced above the BPCL. However, all oil produced from wells developed after 1975 was classified as upper tier. The BPCL was initially set at a property’s average monthly production and sale of old crude oil during 1975. The initial price of lower-tier oil was set at $5.25 per barrel, in contrast to $11.28 per barrel for upper-tier oil. Finally, the third tier included imported oil, stripper-well oil, and some other special cases. Third-tier oil was not subject to price controls. In addition to these three levels, the composite price of oil could not exceed $7.66 per barrel. The composite price was allowed to grow at the rate of inflation plus 3 percent with the growth rate not to exceed 10 percent.

Entitlements Program

At the inception of EPAA, the differential between the price of old and new oil was relatively small. This quickly changed when OPEC quadrupled the world price of oil between August 1973 and June 1974. As the price of new oil was uncontrolled, a substantial difference between old and new oil prices emerged, which bestowed large windfall profits on refiners fortunate enough to be able to buy old oil. Because of the clear advantage given to some refiners, an entitlements program was instituted in November 1974 to allocate price-controlled oil.

To refine one barrel of price-controlled oil, a refiner needed to possess one entitlement. A refiner received entitlements equal to the number of barrels of price-controlled oil it would run if the percentage of price-controlled oil in its total crude oil input were the same as the national average. For example, the national average was 40 percent in December 1974, so that a refiner who processed 1,000,000 barrels of crude oil in that month would receive 400,000 \((= 0.40 \times 1,000,000)\) entitlements. If it wanted to refine 500,000 barrels of price-controlled oil, it would need an additional 100,000 entitlements, which it could acquire by buying them from other firms, as entitlements were transferable.
Windfall Profits Tax

Like EPCA, the Crude Oil Windfall Profits Tax was a three-tier system, but unlike EPCA, it did not control prices. Rather, it levied an excise tax on domestic oil sales. Tier one included oil produced from fields that were producing prior to 1979. The difference between the price of tier one oil and some base price was taxed at a rate of 70 percent. Tier two comprised stripper-well oil and oil produced from National Petroleum Reserves, and the differential between its price and a base price was taxed at a rate of 60 percent. Finally, tier three was made up of new oil (that is, produced from wells developed after 1978) and oil released from tiers one and two. The tax rate on the differential between its price and a base price was 50 percent.

Effects of Price Regulation

In the absence of oil price regulation and with the assumption of price-taking behavior, the marginal revenue curve faced by domestic crude oil producers would be approximately equal to the world price. By distorting marginal revenue, regulation resulted in insufficient production by domestic suppliers. On the demand side, the imposition of price ceilings, along with the entitlements program, resulted in the price faced by domestic oil refiners being below the world price. Refiners processed too much crude oil relative to the social optimum. In this section, the welfare implications of these supply and demand distortions are analyzed.

Distortive Effect on Domestic Supply

Due to the multitier system of price controls, the marginal revenue curve faced by domestic suppliers was a discontinuous function of output. Depicted in figure 18.5 is the marginal revenue curve faced by domestic suppliers under the EPAA. \( Q_s \) denotes the maximum production rate whereby a supplier is classified as a stripper. Recall that a stripper is not subject to price controls. Thus, if output does not exceed \( Q_s \), a supplier’s marginal revenue from another barrel is just the world price, \( P_w \). For production greater than \( Q_s \) but less than the base production control level (BPCL), a supplier received the controlled price for old oil, denoted \( P_o \). Oil produced in excess of BPCL is considered new oil, and its price is uncontrolled. However, recall that for each unit of new oil produced, a supplier gets to release one unit of old oil from price controls. Thus, by increasing production of new oil by one unit, a supplier earns \( P_w \) from selling it on the market and also increases revenue by \( P_w - P_o \) from being able to sell one unit of old oil at the world price rather than at the controlled price. Hence, marginal revenue for production in excess of BPCL equals \( 2P_w - P_o \). Of course, if production exceeded 2BPCL, then there is no old oil left to release, so that marginal revenue falls back down to \( P_w \).

In light of the complex nature of the marginal revenue curve under the EPAA, it is difficult to draw any general conclusions about the effect of regulation on domestic supply.
decisions. If a firm’s marginal cost is sufficiently steep so that the relevant production range is from $Q_s$ to $BPCL$, then the relevant marginal revenue is $P_o$. In that situation, regulation reduced supply, because the production of another barrel would only fetch the low regulated price of $P_o$. If instead the marginal cost function is less steep, so that the relevant production range is above $BPCL$, then marginal revenue is actually higher under regulation, because another barrel would be worth more than $P_w$, as it would allow one barrel of controlled oil to be sold at the world price. One would then expect a firm’s supply to be greater under regulation.

For the ensuing regulatory program of the EPCA, the distortive effect on domestic supply is much more clear-cut. A firm’s marginal revenue curve took the form depicted in figure 18.6. Once again, production below $Q_s$ is not subject to price controls, so that marginal revenue is $P_w$. If the property was producing prior to 1975, then production between $Q_s$ and $BPCL$ received the old oil price of $P_o$, while production in excess of $BPCL$ earned the upper-tier price of $P_u$. If the property was instead new, in that it did not begin producing until after 1975, it received the upper-tier price of $P_u$ for all production above $Q_s$. Since regulation under EPCA kept marginal revenue below the world price for all production in excess of $Q_s$, regulation would be expected to reduce domestic supply.
In the absence of the entitlements program, oil price regulation did not distort domestic demand. Given that price-controlled domestic crude oil was in excess demand, a refiner who considered processing one more barrel of crude oil would have to buy on the uncontrolled market at the world price. Thus, the marginal cost faced by refiners was that based on the world price for crude oil. Since this is the same as in the absence of price controls, the marginal cost of refiners was not distorted by price controls; their demand for crude oil would be unaffected. The only impact of price controls would be to transfer profits from domestic oil producers to those refiners who were fortunate enough to get the price-controlled oil.

Of course, this system lasted for only a few months under Phase IV price controls. With the passage of the EPAA, an entitlements program was introduced for allocating price-controlled oil. Recall that the number of entitlements a refiner received was equal to its total crude oil input multiplied by the national percentage of price-controlled oil of total oil. Let
this percentage be denoted \( z \), and suppose that a refiner considers buying one more unit of imported oil. The price it must pay to the foreign oil producer is \( P_w \). However, by having increased its crude oil input by one unit, the refiner gets a fraction \( z \) of an entitlement. It can now buy \( z \) less of a unit of imported oil and \( z \) more of a unit of price-controlled oil. If \( \bar{P}_d \) is the price of the latter, then the refiner saves \( Z(P_w - \bar{P}_d) \) on this exchange. It follows that the net cost of one more unit of imported oil is \( P_w - Z(P_w - \bar{P}_d) = (1 - Z)P_w + Z\bar{P}_d \). Therefore, under the entitlements program, the marginal cost of one more unit of imported oil to a refiner is less than the world price \( P_w \). Effectively, oil price regulation subsidized imported oil! We would predict domestic demand to be distorted upward as a result.

**Measuring the Welfare Loss**

Depicted in figure 18.7 are the domestic supply curve for crude oil, \( S_d(P) \), the domestic demand curve for crude oil, \( D_d(P) \), and the world supply curve for crude oil, \( S_w(P) \). For simplicity, we have assumed that the supply of imports is perfectly elastic at the world price. The effect of allowing \( S_w(P) \) to be upward-sloping is considered later.

In the absence of oil price controls, the equilibrium price would be \( P_w \), and refiners would demand \( Q^* \) units. At a price of \( P_w \), domestic oil producers would optimally supply \( Q_d \) units, so that imports would be \( Q^* - Q_d \). Now suppose a price ceiling of \( \bar{P}_d \) is imposed on domestic oil and an entitlements program is implemented, so that the marginal price faced by domestic refiners is \((1 - z)P_w + z\bar{P}_d\), which we will denote \( P_r \). Because of the lower price faced by refiners, their demand would increase from \( Q^* \) to \( Q^{**} \). Because domestic suppliers receive the lower price of \( \bar{P}_d \) they are only willing to supply \( \bar{Q}_d \) units. As a result, imports must increase from \( Q^* - Q_d \) to \( Q^{**} - \bar{Q}_d \).

We can now assess the welfare effects of regulation. Regulation induced domestic refiners to process too much oil by the amount \( Q^{**} - Q^* \). For those additional units, the social cost, as measured by the world price \( P_w \), exceeds the social benefit by the triangle \( abc \). In addition, the reduced domestic supply means that more resources are used to supply \( Q_d - \bar{Q}_d \). Thus, there is an additional welfare loss of triangle \( def \). The total welfare loss from regulation is then measured by the sum of the two shaded triangles in figure 18.7.

The sizes of these triangles were estimated for 1975–1980, and the estimates are shown in table 18.3. In 1975, triangle \( abc \) was estimated as $1,037 million (in 1980 dollars), whereas triangle \( def \) was estimated as $963 million. The total deadweight loss from distorted supply and demand decisions was $2 billion. From 1975 to 1980, the average annual welfare loss due to demand and supply distortions was about $2.5 billion so that the total welfare loss from EPCA was on the order of $15 billion. The exact welfare loss from regulation fluctuated over time because of the changing price of world oil. Also note that expenditure on imports increased considerably as a result of oil price regulation.

There are several caveats that need to be mentioned with respect to these estimates. First, these triangles measure the welfare loss from regulation under the assumption that the
Figure 18.7
Equilibrium under Crude Oil Price Controls
The marginal social cost of crude oil is properly measured by the world price. Thus, when regulation induces domestic consumption to increase from $Q^*$ to $Q^{**}$, the value of that additional consumption is less than the cost of the resources used as measured by $P_w$. Because $P_w$ represents the marginal social cost to the U.S. economy of another barrel of (imported) oil, triangles $abc$ and $def$ measure the welfare loss to the U.S. economy from demand and supply distortions. However, they do not measure the welfare loss to the world economy unless the world market is competitive. The world market was, in fact, not competitive, for OPEC was a price setter acting in a monopolistic manner. As a result, the price of oil in the world market was set too high relative to other energy sources. Because this excessive price implies that too little oil was being consumed, regulation that increased the consumption of oil need not have been welfare-reducing from the perspective of the world economy. The bottom line is that the estimated welfare losses are for the U.S. economy and not the world economy.

A second important caveat is that this analysis assumed that U.S. demand does not influence the world price. This assumption was implicit in specifying that the world price was fixed at $P_w$. On the contrary, the United States is a big player in the world energy market, so that its demand generally does influence the world price. To capture this effect, let us now specify a rising world supply curve so that changes in U.S. oil consumption affect the world price for oil.

Having made this modification in our model, we introduce two new effects. Because greater U.S. consumption raises the world price, this higher price will tend to reduce the increase in domestic consumption as a result of regulation. If $P_w$ is the world price in the absence of regulation, the world price with U.S. regulation will be above $P_w$ because of the upward distortion in domestic demand resulting from the effective subsidization of oil imports. However, U.S. consumption will not rise as much with a rising world supply curve because of the resulting higher world price. Thus, the welfare loss from demand distortions is lower if increased U.S. consumption raises the world price. Under the specification of a

<table>
<thead>
<tr>
<th>Year</th>
<th>Demand-Side Deadweight Loss</th>
<th>Supply-Side Deadweight Loss</th>
<th>Additional Expenditure on Imports</th>
</tr>
</thead>
<tbody>
<tr>
<td>1975</td>
<td>$1,037</td>
<td>$963</td>
<td>$11,550</td>
</tr>
<tr>
<td>1976</td>
<td>852</td>
<td>1,046</td>
<td>15,052</td>
</tr>
<tr>
<td>1977</td>
<td>654</td>
<td>1,213</td>
<td>16,496</td>
</tr>
<tr>
<td>1978</td>
<td>300</td>
<td>816</td>
<td>11,319</td>
</tr>
<tr>
<td>1979</td>
<td>627</td>
<td>331</td>
<td>17,644</td>
</tr>
<tr>
<td>1980</td>
<td>1,038</td>
<td>530</td>
<td>34,475</td>
</tr>
</tbody>
</table>

rising world supply curve, the welfare losses from demand distortions have been reestimated. For a range of reasonable values for the elasticity of the world supply curve, the welfare loss was lower by 13 to 61 percent. For example, if the world supply curve has a price elasticity equal to one, then the average annual welfare loss from demand distortions over 1975–1980 was $413 million, in contrast to $751 million when we assume a perfectly elastic world supply curve.\(^{20}\)

A second additional effect from regulation that arises with a rising world supply curve is that there is an increase in the transfer of wealth from the U.S. economy to foreign oil producers. This wealth transfer is due to the higher world price. Depending on the elasticity of the world supply curve, the average wealth transfer over 1975–1980 was between $1.625 billion and $8.115 billion per year.\(^{21}\) These are substantial welfare losses to the U.S. economy.

**Small-Refiner Bias and Productive Inefficiency**

In addition to the welfare losses from too much domestic demand and too little domestic supply, regulation also caused productive inefficiency by promoting the use of inefficiently small refiners. This bias was created by the entitlements program allocating additional entitlements to small refiners. The additional number of entitlements was decreasing with the production of a refiner and was eliminated at 175,000 barrels per day. In 1977, the small-refiner bias applied to 126 of 148 domestic refining firms. However, these 126 small refiners made up only 18 percent of U.S. refining capacity.\(^{22}\)

This small-refiner bias resulted in productive inefficiency because the minimum efficient size for a refiner was around 200,000 barrels per day.\(^{23}\) That is, average cost was decreasing until around 200,000 barrels, at which point it became relatively flat. Thus, by providing additional entitlements to small refiners, regulation provided an incentive for inefficiently small refiners to operate. The small refiner bias was found to be increasing over the period of regulation. Refiners with less than 175,000 barrels per day made up 14.4 percent of total refined product sales in 1972 and this had increased to 17.8 percent by 1975.\(^{24}\) Regulation not only resulted in inadequate domestic supply but also changed the composition of the domestic suppliers toward smaller, less efficient refiners. As a result of this bias, industry average cost was raised and social welfare was reduced.

---

20. Hubbard and Weiner, “Petroleum Regulation”.
21. Ibid.
24. Ibid.
Reduced Incentives for Exploration

It has been argued by many economists specializing in the oil industry that regulation reduced the incentive to explore for new oil reserves. On the one hand, since marginal revenue from new oil production was higher than the world price under the EPAA, one would think that regulation would have increased the expected profits from exploration and thus actually increased the incentive to explore. On the other hand, marginal revenue was constrained below the world price under the EPCA so that the effect on expected profits from exploration would be just the opposite.

A concern of firms is that if they discover new oil reserves, future regulation may reclassify it as old oil and thus restrict its price below the world price. This effect would tend to reduce the incentive to explore under either regulatory program. Furthermore, uncertainty about future regulation can increase the variability of returns from exploration. There is already considerable uncertainty associated with exploration, even in the absence of regulation. A firm is uncertain whether it will discover new oil reserves and, if it does, it is uncertain as to the revenues that will be generated. Revenue uncertainty is due to uncertainty both over the size of the reservoir and what the world price will be. A regulatory structure adds uncertainty over what the regulated price will be. Because increased variability is thought to reduce the incentive to explore, price regulation can reduce exploration efforts by the domestic oil-producing industry.

Administrative and Compliance Costs

With any regulation, whether or not there is a rationale for it, there is always a welfare loss associated with the resources used to implement it. In light of the complexity of oil price controls and the entitlements program, the administrative and compliance costs were quite substantial. By the late 1970s, the industry was incurring an annual cost around $500 million (in 1977 dollars) for reporting and administrative duties, while the cost to the FEA was around $40–50 million per year.25 Adding these costs to those associated with supply and demand distortions, the annual welfare loss in 1979 due to regulation was around $3.2 billion (in 1979 dollars).26

On this note, it is interesting to mention that when the Windfall Profits Tax was repealed in 1988, it was creating a welfare loss in spite of the fact that it was no longer collecting revenues! (Prior to that time, however, it did collect revenues totaling $77 billion over 1980–1988.)27 The industry estimated that it incurred on the order of $100 million annually

25. Ibid.
in reporting and complying with the law.\textsuperscript{28} Even when a regulation is not binding, there can be welfare losses associated with it.

### Price Regulation of the Natural Gas Industry

Like crude oil, natural gas is a hydrocarbon and is produced by drilling into an underground reservoir. In fact, gas is often found along with oil. Although oil has had a more significant historical role in the U.S. energy market, the first natural gas well in the United States was drilled in 1825, which predates the first oil well by more than 30 years. The standard unit of measurement for natural gas is 1,000 cubic feet, typically denoted Mcf.\textsuperscript{29}

#### Technological Structure

The production of natural gas is almost identical to that of crude oil. Furthermore, about 20 percent of natural gas in the United States comes from wells also producing oil.\textsuperscript{30} After extraction comes the transmission of natural gas from the wellhead to local distributors. The difficulty in its transmission is that natural gas is bulky. One million BTU of natural gas occupies about one Mcf at room temperature and atmospheric pressure, whereas the same amount of energy is stored in only 7.5 gallons of gasoline.\textsuperscript{31} Given its bulkiness, generally the only economical way in which to transport natural gas is by pipeline.\textsuperscript{32} However, early gas pipelines suffered from serious problems with leakage and, as a result, pipelines could not be of great length. Consequently, the natural gas industry was slower to develop nationally than the crude oil industry. A sign of the future national scope of the industry took place in 1931 when a pipeline was built between Chicago and the natural gas fields of Texas. Nevertheless, it was not until after World War II that pipeline construction grew rapidly. Only then did the natural gas industry become truly national in scope.

After production and transmission, the last segment to the natural gas industry is local distribution. A local distribution network connects individual residential and commercial users with the pipeline and thus has many of the properties of a local electric utility.

\begin{itemize}
  \item \textsuperscript{28} Ibid.
  \item \textsuperscript{31} Ibid.
  \item \textsuperscript{32} Actually, another method of transporting natural gas is to liquefy it to 1/600th of its normal size and then ship it.
\end{itemize}
Economic Structure of Transportation

Because a unique and essential aspect of the natural gas industry is its transportation system, let us briefly discuss its properties. First, both pipelines and local distribution systems are characterized by economies of scale, since there is a large fixed cost component and a relatively low marginal cost. Given that these investments are largely nonrecoverable, they require a long-term commitment from consumers. Hence, pipelines need local distribution networks to agree to buy gas from them for an extended length of time, and local distribution networks need a similar commitment from commercial and residential consumers. Although a local distribution network is a natural monopoly and a pipeline may also be the only supplier for that network, it is important to recognize that competitive forces are still at work, inasmuch as natural gas competes with other energy sources. Manufacturing plants can choose between gas, oil, and coal to run their plants. Though switching in the short run is often difficult, in the long run residential demand for natural gas depends on the price of other energy sources such as electricity and oil.

Regulatory History

During the early part of the Great Depression, the natural gas industry was chaotic. Markets in the eastern United States suffered from rising prices and significant shortages, whereas there was enormous excess supply in the Southwest. Increasing pipeline construction would have greatly alleviated this dilemma, but the Great Depression brought most pipeline construction to a halt. Amid this chaos, there were abuses by local public utility holding companies, as many acted essentially like unregulated monopolists and charged excessively high prices.

The Natural Gas Act of 1938

By the mid-1930s, state public service commissions and representatives of northern cities were lobbying the U.S. Congress to regulate the natural gas industry. In response, Congress passed the Natural Gas Act of 1938. This act gave the Federal Power Commission (FPC) control over the interstate transportation and sale for resale of natural gas in interstate commerce. However, the 1938 act stipulated that it did not apply to the production or gathering of natural gas. The FPC was specifically given control over entry in the transmission segment of the natural gas industry. FPC approval was required to build an interstate pipeline that delivered gas to a market already served by another gas line. That type of entry control is reminiscent of airline and trucking regulation.

33. Market forces are now the primary force in the distribution of natural gas, as we discuss in the section Transition from Regulation to Markets in the Transmission of Natural Gas.
Phillips Petroleum Co. v. State of Wisconsin

After the 1938 act, the FPC focused on reducing the risk associated with pipelines. This purpose was largely accomplished by limiting access of pipelines to end-use markets. Before a pipeline entered a market, the FPC required it to demonstrate that it had large reserves under contract. While giving it the right to control entry, the FPC interpreted the 1938 act as not giving it the right to control gas prices.

In a landmark case in 1954, the Supreme Court ruled in Phillips Petroleum Co. v. State of Wisconsin that the 1938 act did indeed grant the FPC the task of regulating wellhead rates for natural gas. From that point onward, the FPC was very active in regulating gas prices. As will be discussed in detail in the section on regulatory practices, a key feature of price regulation was the establishment of different rates for “old” and “new” gas. This was done along similar lines to the regulation of crude oil prices.

Natural Gas Policy Act of 1978

By the late 1960s, shortages were beginning to emerge in natural gas markets in the Midwest and Northeast. Then, to further exacerbate the situation, the oil price shocks of 1973–1974 hit. By the mid-1970s there was considerable disequilibrium in the natural gas market as government-set prices were creating significant shortages.

The response of the government was the Natural Gas Policy Act of 1978. This act called for the gradual decontrol of prices for new gas, defined as gas produced by wells discovered since 1977. The objective was to have new gas prices at market-clearing levels by 1985. However, the 1978 act continued to control old gas prices (that is, gas from wells discovered before 1977). Prices could only grow at the rate of inflation and, in addition, price control was extended to the intrastate market for the first time. Finally, government jurisdiction over the natural gas market moved from the FPC to the Federal Energy Regulatory Commission (FERC).

Because of the 1978 act, deregulation has indeed taken place. Prices for gas produced from deep wells were fully deregulated in November 1979 and, according to plan, new gas prices were decontrolled in January 1985. Old gas prices were also effectively deregulated in 1986 when the FERC issued Order 451. It collapsed fifteen vintages of old natural gas into one and set the ceiling price above the market-clearing price. In July 1989, President George Bush signed the Natural Gas Wellhead Decontrol Act of 1989, which fully deregulated gas prices. An overview of the regulatory structure from 1938 to 1985 is provided in figure 18.8.

Regulatory Practices

Rate-of-Return Regulation

The basic model used by the FPC in setting natural gas prices was rate-of-return regulation (see chapter 12). Prices were set with the objective of allowing natural gas producers to
In the early 1970s, the allowed rate of return was initially 12 percent and was then increased to 15 percent in 1974. For example, the FPC set a rate of $1.42/Mcf in 1976. This rate was calculated by allowing for each Mcf $0.37 for exploration and production costs, $0.03 for operating costs, $0.22 for royalties, $0.38 for taxes, and $0.42 for a return on investment.

Initially, the FPC determined rates on a case-by-case basis and, before 1954, it averaged 700 gas-rate filings per year. However, after the Supreme Court ruling in 1955, the number

---

**Figure 18.8**

Regulatory Jurisdiction of the Natural Gas Industry


35. Ibid.
of filings increased to around 11,000. The result was a massive backlog of cases. By 1960 there were still 3,278 producer rate increase filings awaiting decisions.

Given the burdensome nature of this approach to ratemaking, the FPC changed tactics in 1960 by dividing the United States into 23 geographic areas and requiring price to be uniform within each area. The initial price set was based on prices in 1956–1958. The first area rate case was the Permian Basin Area case. It took the FPC five years to decide it, and after all that it was taken to court. Three years later, in 1968, it was approved by the Supreme Court. Due to these delays, rates were basically frozen from 1960 to 1973.

Fourteen years after area ratemaking was accepted, the FPC had still not completed even half of the twenty-three area cases. In its ongoing pursuit of a more manageable system, nationwide ratemaking was accepted in June 1974. A rate of $0.42/Mcf was specified for all interstate sales. This rate applied only to wells completed after January 1, 1973.

**Multitier Pricing System**

One of the key properties of natural gas price regulation was the specification of different rates for different vintages of gas. New gas was defined as gas discovered or first committed to the interstate market after some specified date, whereas old gas was gas discovered before that date. In the Permian Basin Area case, new gas was priced at $0.165/Mcf and old gas was priced at $0.145/Mcf. The objective of the FPC was to keep new gas rates high to encourage exploration and old gas rates low to mitigate the transfer of wealth from consumers to producers. It was believed to be unnecessary to keep old gas prices at the market-clearing level. This is a point to which we will return shortly.

From November 1976 until the 1978 act, a five-tier system was in place, with prices ranging from $0.295 for gas produced from wells dating before January 1, 1973, to $1.42 for gas produced from wells dating after January 1, 1975. The number of tiers was increased in response to growing shortages in the interstate market. Finally, the system reached its acme in complexity with the 1978 act, when 28 different categories of gas were established for pricing purposes. New gas was defined as that from wells commencing production after January 1, 1977, and its rate was set at around $2.50/Mcf. Old gas prices were set close to pre-1978 rates. In addition, different rates were set for the interstate and intrastate markets.

**Effects of Price Regulation**

The important feature of gas price regulation was that both old and new gas prices were set below market-clearing levels. Since natural-gas producers were not required to meet demand,
the result of price regulation was substantial excess demand in the interstate market. A multi-
titier system of pricing improved incentives for exploration relative to setting a single price
for all gas, but it also increased drilling and exploration costs.

Divergence between Interstate and Intrastate Rates

The best way to show that interstate gas rates were set too low is to compare them with
intrastate rates. Given that the intrastate gas markets were unregulated, it is reasonable to
suppose that a competitive equilibrium was achieved there. Even though interstate gas rates
were regulated by the FPC while intrastate rates were uncontrolled, before 1970 the differ-
ence between interstate and intrastate rates was relatively small. However, from 1969 to 1975,
a considerable rift occurred due to growing interstate demand and the oil price shock. During
that time interstate prices for new contracts rose 158 percent, from $0.198/Mcf to $0.51/Mcf,
while intrastate prices for new contracts rose a whopping 650 percent, from $0.18/Mcf to
$1.35/Mcf.38 Figure 18.9 shows the growing differential between regulated interstate rates
and unregulated intrastate rates.

Diversion of Interstate Supply to the Intrastate Market

One interpretation of FPC policy is that it was based on the assumption that the industry
supply curve for natural gas was essentially vertical after some output rate, as depicted
in figure 18.10. If indeed the supply curve had this property, given growing demand for
natural gas due to rising energy demand and the increase in the price of a substitute (that is,
crude oil), one would anticipate gas prices to increase but the supply response to be minimal
once production of $Q^*$ is reached. In light of this model, one can understand FPC policy.
By keeping gas prices at $\bar{P}$ even when demand has shifted to $D_3(P)$, there is no welfare
loss, because supply is still $Q^*$, while the transfer of wealth from consumers to producers is
prevented.

Unfortunately, history has shown that the interstate supply curve for gas was indeed respon-
sive to the price. A more accurate representation is that shown in figure 18.11. In that case,
with a price ceiling of $\bar{P}$ as demand grows from $D_1(P)$ to $D_2(P)$ to $D_3(P)$, an increasing short-
age emerges because supply does not respond. With price ceilings in interstate markets, gas
producers optimally diverted supply to intrastate markets where they could receive a higher
price. This diversion was felt by interstate pipelines, as they were forced to meet less than
their contractual amounts to “interruptible” customers. Figure 18.12 shows the growing
number of curtailments by interstate pipelines. (A curtailment is an unfilled order.) By 1977,
curtailments had reached a level of 3.7 trillion cubic feet. Gas reserves dedicated to the inter-
state market also dropped drastically. Although 67 percent of reserves were committed to the

38. United States House of Representatives, report no. 94–732 (cited in Hubbard and Weiner, Petroleum Regula-
tion.) “Petroleum Regulation.”
interstate market from 1964 to 1969, less than 8 percent were so committed from 1970 to 1973.\textsuperscript{39}

### Allocative Inefficiency

By setting price ceilings below market-clearing levels while allowing producers to freely choose how much to supply, regulation resulted in inadequate amounts of natural gas being supplied to the interstate market. Consumers were forced to turn to more expensive and less efficient energy sources, such as oil and electricity. Potentially large welfare losses were incurred by the substitution of electricity for natural gas in providing heating to the residential market. In 1984, the price per million BTU faced by the residential market for electricity was $23.59, compared to just $6.26 for natural gas.\textsuperscript{40}

\textsuperscript{39} Ibid.

\textsuperscript{40} Pierce,
Figure 18.10
The Effects of FPC Policy with a Vertical Supply Curve
A second source of welfare loss is associated with inefficient rationing. As shown in the section on the theory of price ceilings, when there is excess demand for a good, additional welfare losses occur when the good does not end up in the hands of the consumers who value it the most. One can make a strong case that such losses occurred in the interstate gas market. Utilities and industries located in gas-producing states were able to receive ample supplies of gas, while residential consumers in the Northeast had to use expensive alternative energy sources like oil and electricity for heating. Because utilities and industries could have converted to coal, they had lower-cost alternatives than were faced by the Northeast residential market. Thus, efficiency would have been increased by reallocating gas to residential consumers in the Northeast.
Figure 18.12
Firm Curtailments by Interstate Pipelines, 1970–1977 (Intercompany Sales Eliminated)

Excessive Drilling Costs and Productive Inefficiency

As it often does, regulation creates unusual incentives that can lead to productive inefficiencies. Consider a natural gas producer whose well was producing prior to some date, so that all gas extracted from it was classified as old gas. If the gas could be reclassified as new gas, the producer’s revenue would rise from the higher price ceiling applied to new gas. In some cases, a gas producer could accomplish this reclassification by drilling a new well on an already discovered gas reservoir. As a result, regulation resulted in the drilling of wells that were superfluous from a production standpoint but were profitable from a revenue standpoint. The resources used to build these additional wells must be added as an additional source of welfare loss. An offsetting effect is that the higher price from getting old gas reclassified as new gas would be expected to bring forth additional supply to the interstate market.

Reduced Incentives for Exploration

As in the case of oil price controls, it is generally believed that gas price controls reduced the incentive to explore for new reserves. Because the price one could expect from finding new reserves was lower under regulation, the expected profits from exploration were reduced, so that less exploration would optimally take place. The evidence is consistent with this argument. Proved gas reserves rose quite steadily from 1948 to 1967, when they reached a high of 297 trillion cubic feet. However, under gas price controls, these reserves steadily declined and were 31 percent lower in 1980 than in 1970.\(^41\) This decline was particularly striking because the price of oil was way up. A higher price of a substitute for natural gas would tend to induce additional exploration.

Regulation specified higher price ceilings for gas from deep wells and thereby increased incentives to explore for high-cost gas. With the early elimination of price controls for gas from deep wells, there was in fact a deep-gas boom during 1978–1982. Because drilling costs increase roughly in proportion to the square of well depth, considerably more resources were used to drill deep wells. Generally, the incentive to explore was curtailed by regulation, but it was actually increased for higher-cost gas.

Nonprice Competition and Take-or-Pay Contracts

For a pipeline to be profitable, it is essential that it have access to natural gas producers. For this reason, pipelines often enter into long-term contractual relationships with producers. Because of the shortages generated by regulation, there was a heightened incentive to sign such long-term agreements. Pipelines competed with one another to get gas producers to sign agreements that would contractually guarantee supply to the pipelines. Since the

\(^{41}\) Hubbard and Weiner, “Petroleum Regulation.”
price they could offer producers was limited by regulation, pipelines turned to nonprice methods to compete. In particular, they competed through the take-or-pay provision of contracts.

A take-or-pay provision guarantees that a pipeline will purchase a minimum quantity each year at a specified price; it is equivalent to a fixed payment per year. As shown in table 18.4, the minimum purchase requirement as a percentage of deliverability or capacity was distinctly higher after the 1973 oil price shock than before it. This finding reflects the fact that the difference between the regulated price and the market-clearing price was much greater after 1973. With a higher differential, competition drove pipelines to offer more attractive terms, like a higher take-or-pay provision, in order to get gas supplies.

The increased take-or-pay requirement created considerable problems for pipelines because oil prices did not rise as originally expected, nor did natural gas prices. As a result, pipelines that entered long-term agreements with a large take-or-pay requirement found themselves faced with financial obligations often in excess of their assets. Though these obligations represented a serious postderegulation problem, they had the interesting implication of inducing substantive deregulation in the transportation of natural gas, a subject that will be discussed in a subsequent section.

**Path to Deregulation**

In setting up a plan for the gradual decontrol of natural gas prices, the 1978 act presumed that the price of oil would remain around a price $15 per barrel in 1978 dollars. Unfortunately, the price of oil shot up to $30 per barrel in 1979–1980 and resulted in a severe disequilibrium in the natural gas market.
The effects of continued regulation from 1977 to 1985, as opposed to fully decontrolling prices in 1977, have been estimated. These estimates reveal that continued price regulation reduced the revenues of gas producers by $106.3 billion over 1977–1985 but also reduced consumer expenditure by $98.7 billion. Rather interestingly, the impact on different consumer groups was quite varied. Continued regulation over 1977–1985 reduced gas expenditure for utilities by $41.3 billion, for commercial users by $3.0 billion, and for industrial users by $69.9 billion. In contrast, expenditure by residential users was actually higher by $8.6 billion as a result of gradual rather than full decontrol in 1977.

Transition from Regulation to Markets in the Transmission of Natural Gas

Industry Structure under the Natural Gas Act of 1938

As described earlier, the natural gas industry involves three functions: production, transmission (from producer to the local distribution system), and distribution (to the final customer). There are many feasible ways in which the industry could be organized. It could involve fully vertically integrated firms that perform all three functions. Alternatively, distinct firms could perform these functions. And, in that case, there are various arrangements that could exist between firms. A producer could sell gas to a pipeline or instead contract directly with a local distribution system (or even the final customer), in which case the producer or customer would hire the pipeline (and perhaps also the local distribution system) to transport it.

In giving the FPC jurisdiction over the interstate transmission of natural gas, the biggest impact of the Natural Gas Act of 1938 was to cause a particular organizational and contractual arrangement to emerge. Discouraging vertical integration and controlling entry, the FPC’s policies resulted in production, transmission, and distribution being provided by distinct firms. Furthermore, the FPC required a pipeline to own the gas that it transported or, as it is called, engage in “merchant carriage.” The process then involved a producer selling natural gas to a pipeline, which then transported it to a city market, at which point it sold the gas to the local distribution system. The transportation services of pipelines were forced to be bundled with their function as traders of natural gas.

A second important facet of the regulation of the transmission process was the control of entry. Prior to the construction of an interstate pipeline, the FPC required that the pipeline be certified. As part of the certification process, a pipeline had to show that it had gas reserves to supply its downstream customers typically for 15 or 20 years. Since the ownership of


43. This discussion is largely based on Arthur S. De Vany and W. David Walls, The Emerging New Order in Natural Gas: Markets versus Regulation (Westport, Conn.: Quorum Books, 1995).
reserves by pipelines was discouraged, this requirement caused pipelines to enter into long-term contracts with producers. The reserves of a producer were then dedicated to a small set of pipelines.

The implication of this regulatory-induced structure was a set of rigid and balkanized gas markets that acted largely independently of one another. A producer was tied to a particular pipeline, and the pipeline was tied to a particular set of local distribution systems. A surplus of gas in one market could not easily be moved to satisfy shortages in other markets. As a result, a national gas market failed to develop.

**Development of the Open Access Policy**

Owing to the extensive use of take-or-pay contracts and the unanticipated fall in natural gas prices, pipelines were left with the requirement to purchase a minimum amount of gas from producers at excessively high prices. As part of the renegotiation of the contracts between pipelines and producers in 1983 and at the approval of the FERC, these pipelines agreed to perform “contract carriage” or, in other words, transport gas owned by others and, more specifically, gas directly sold by producers to their customers (thereby avoiding the usage of the pipeline as a middleman).

This practice ultimately led to a change in FERC policy in October 1985. Order 436 allowed a pipeline to choose “open access” status. Declaration of this status meant that it was willing to transport gas on a nondiscriminatory basis to all customers. Pipelines responded by overwhelmingly choosing open access status so that, within two years, contract carriage made up 75 percent of all gas moving across state lines.\(^{44}\) Going further down this path, in 1992 Order 636 mandated the unbundling of gas supply from pipeline transportation.

It is worth noting that this structural change is representative of a general deregulatory movement to open access. The 1980s witnessed free entry into the long-distance telephone market with long-distance carriers having open access to the local telephone system operated by, for example, a regional Bell operating company. Also, the current deregulatory process in the generation and transmission of electric power has obvious similarities with what took place in the natural gas industry.\(^{45}\)

**Industry Structure under Open Access**

A vast structural change has emerged from the FERC’s decision to allow pipelines to be contract carriers. Rather than having producers rigidly tied to pipelines and pipelines rigidly tied

---


to local distribution systems, pipelines can deal with any entity that they can reach. This freedom has induced pipelines to create a well-connected network. Where pipelines pass close to one another, links have been added so that those pipelines can provide transportation services to an expanded array of producers and customers. This structural change has linked more buyers (local distribution systems and large customers like manufacturers) and sellers (producers of natural gas), and the ability of buyers and sellers to engage in direct transactions has spawned more efficient contractual arrangements. The evidence is that the result has been a more competitive and more efficient market for natural gas.  

Summary

Both the regulation of domestic crude oil prices and the regulation of interstate natural gas prices were characterized by price ceilings established at levels below what was required in order to clear the market. As a result, shortages emerged. However, due to differences in these two markets, the responses to these shortages were quite different. In the case of the crude oil market, consumers responded by increasing foreign imports. In the case of the interstate natural gas market, consumers were forced to switch to less efficient energy sources, such as electricity and oil. The gas shortage was further exacerbated by intrastate prices’ being unregulated. Gas producers had increased incentives to divert supply away from the interstate market and to the intrastate market.

Moving to other sources of regulatory-induced welfare losses, we found that price regulation caused productive inefficiencies in both the oil and gas markets. Oil price regulation created a bias for small refiners, so that there were increased incentives to operate refineries with capacities below minimum efficient scale. The small-refiner bias raised industry average cost and wasted valuable resources. Both oil and gas price regulation increased drilling and exploration costs by providing higher price ceilings to oil and gas produced from more costly wells. Producers were then provided with incentives to explore for reserves that entailed more rather than less resources for production. Finally, energy regulation reduced the incentive for oil and gas producers to explore for new reserves, resulting in even greater shortages in the future.

Although we have discussed the regulation of oil and gas separately, one would expect there to be important connections between the two markets, since oil and gas are substitutes. In particular, the oil price shocks of the 1970s played a central role in the magnitude of the regulatory distortions in the natural gas market. When the price of oil increased, demand for gas naturally increased. Of course, the regulatory-imposed price ceiling prevented gas prices from rising, so that supply did not increase so as to meet the greater demand. As a result,

welfare losses from gas regulation were much greater because of the movement in world oil prices in the 1970s.

Questions and Problems

1. Is there an economic rationale for a state government to control the rate at which crude oil is extracted?

2. Describe how the optimal extraction rate of crude oil is affected when
   a. Information is revealed that raises the expected price of crude oil in the future.
   b. The interest rate rises.

3. When crude oil price controls were in place, what would have been the welfare implications of a ban on oil imports? (Use figure 18.7.)

4. Assume the market supply curve is \( S(P) = 30P \) and the market demand curve is \( D(P) = 500 - 20P \).
   a. Derive the competitive equilibrium price.
   b. Suppose that the government institutes a price ceiling of 5. Derive the welfare loss from price controls.
   c. How much are producers willing to pay in order to get their legislators to remove the price ceiling?

5. How does the elasticity of the industry supply curve affect the welfare implications of price ceilings?

6. Let us consider the effects of the Emergency Petroleum Allocation Act (EPAA) on the domestic supply of crude oil. Assume the world price is 15, the price of old oil is set at 10, a domestic supplier is classified as a stripper if and only if its production does not exceed 10, and the base production control level is 20. Derive the effect of EPAA on the supply decision of a domestic supplier of crude oil when its marginal-cost curve is represented by
   a. \( MC(Q) = 5 + 2Q \)
   b. \( MC(Q) = 5 + 0.75Q \)
   c. \( MC(Q) = 0.02Q^2 \)

7. Assume that the domestic supply curve for crude oil is \( S(P) = 5P \) and the domestic demand curve for crude oil is \( D(P) = 500 - 20P \). Further assume that domestic oil refiners face a perfectly elastic supply of oil imports at a price of 16.
   a. Derive the domestic price, the quantity processed by domestic oil refiners, and the amount of imports at the competitive equilibrium. Now suppose that domestic crude oil suppliers face a price ceiling of 8. Further suppose that for each two units of crude oil purchased, a domestic oil refiner gets one entitlement to domestic crude oil.
   b. Derive the marginal price of crude oil faced by domestic oil refiners.
c. Derive the effect of regulation on the amount of crude oil processed by domestic oil refiners and the amount of imports.

d. Derive the welfare effect of regulation on U.S. consumers and producers.

8. Who gained and who lost by prorationing? By crude oil price controls? By natural gas price controls?

9. Why is the market for crude oil more international than the market for natural gas?

10. How did the combination of oil and gas price controls affect the incentives for exploration? (You should take into account the fact that a certain percentage of natural gas comes from wells also producing crude oil.)

11. The regulated price of natural gas for wells dating before January 1, 1973, was set lower than the regulated price for wells dating after January 1, 1973. Why did regulation create this price differential? Does it make a difference that natural gas is homogeneous?

12. In response to the doubling of the price of crude oil in 1979–1980, should the government have speeded up or slowed down the gradual decontrol of natural gas prices?
III HEALTH, SAFETY, AND ENVIRONMENTAL REGULATION
Introduction: The Emergence of Health, Safety, and Environmental Regulation

As the review of regulatory costs in chapter 2 indicated, environmental and other social regulations have become an increasingly prominent part of the regulatory mix. Indeed, the combined costs of the budgets of these regulatory agencies in 2005 was more than five times greater than that associated with economic regulations, because of both the growth in the newer social regulations and the rise of deregulation for the economic regulations.

If this book had been written before 1970, it is likely that there would have been no discussion at all of health, safety, and environmental regulation. Beginning in the early part of the twentieth century, there had been, of course, several efforts in the health, safety, and environmental regulation areas, but for the most part these effects were limited to the safety of food and drugs. Consumer advocates such as Ralph Nader became prominent influences on the national agenda in the mid-1960s, and the rising public concerns influenced national policy.

The decade of the 1970s marked the emergence of almost every major risk or environmental regulation agency. The U.S. Environmental Protection Agency, the National Highway Traffic Safety Administration, the Consumer Product Safety Commission, the Occupational Safety and Health Administration, and the Nuclear Regulatory Commission all began operation in that decade. Although in some cases these agencies absorbed functions that had been undertaken at a more modest level by other agencies, the advent of these regulatory agencies marked more than a consolidation of functions, inasmuch as each of the agencies also had its own legislative mandate that gave it substantial control over the direction of regulatory policies in its area.

The early years of operation of these regulatory agencies were fairly controversial. Expectations were high, and for the most part these expectations regarding potential gains that would be achieved were not fulfilled. Congressmen and engineers often predicted dramatic safety gains achievable at little cost, but these predictions often ignored the role played by individual choice. Firms could ignore regulations, and indeed had an incentive to ignore them when compliance was expensive. Consumers might have chosen not to take certain precautions, such as wearing seat belts. In addition, there was substantial resistance to the increased expansion of regulatory control over decisions that were formerly left to business. Many ill-designed regulations became the object of ridicule and vehement business opposition. Over time these regulatory efforts have become a more generally accepted component of the federal regulatory efforts. In addition, there has been an effort on the part of regulatory agencies to strike a better balance between the benefits achieved by these policies and the costs they impose.

Unlike the economic regulation areas, however, there has been no major push toward deregulation. Indeed, the recent emergence of concerns such as global climate change has increased the extent of this form of regulation. There is little doubt that actual market failures exist in the context of social regulation. In many cases, such as air pollution, no markets exist at all for the commodity being produced, and there is no market-based compensation
of the victims of pollution. Markets could never suffice in instances such as this, so that our objective in the social regulation area will always be sounder regulation rather than no regulation whatsoever.

Consider, for example, the current regulatory agendas of the various agencies. At the U.S. Environmental Protection Agency (EPA), the emphasis is on dealing with the major regulatory problems that will have long-run ecological consequences. These include the “greenhouse effect,” acid rain, and the depletion of the ozone layer in the atmosphere. The EPA is placing increased emphasis on environmental problems that will define the well-being of the world throughout the twenty-first century. None of these problems can be adequately addressed by free markets alone. At the Occupational Safety and Health Administration (OSHA), the task is to devise more effective enforcement mechanisms, to update and make more performance-oriented the technological standards for workplace design, and to address in a more comprehensive manner the health hazards imposed by workplace conditions. Finally, in the product safety area, regulation of automobiles and other products is continuing as before, but there is also a new element as the increased role of product liability has highlighted the importance of coordinating the various social institutions at work in promoting health and safety. The main question in this area is one over which no one institution has responsibility—how the responsibilities should be divided among federal regulatory agencies, the market, and the courts.

### Risk in Perspective

Before addressing the specific aspects of the efforts of the different regulatory agencies, it is helpful to put the scale of the problems they are addressing in perspective. Table 19.1 lists the various causes of death in the United States population.¹ The principal focus of the efforts of regulatory agencies is on accidents. All accidental deaths (unintentional injuries) make up only 4 percent of the total death rate, so that even a fully effective safety regulation effort would necessarily play a small part in reducing overall mortality rates. In the accident category, roughly half of all accidental deaths are due to motor vehicle accidents, which are a major component not only of nonworking accidents but also of work accidents. The second leading component is falls, which are concentrated almost entirely among the population over the age of seventy-five. These risks for the most part occur outside of the purview of government regulation. Similarly, drownings also tend to be age-specific, as a leading age group for drowning rates consists of children age, zero to four years. Boating and swimming accidents are monitored by various parties, such as lifeguards. Moreover, there are government

---

regulations of swimming pool slides, for example, that will influence drowning risks. However, for the most part these risks also fall out of the domain of regulatory policies. The categories of fires, burns, and poisoning are the subject of extensive government regulation, including safety cap and flammability requirements of various kinds. Most, but not all, accidents are potentially under the influence of government regulation. Even a fully effective regulatory effort could not be expected to eliminate accidents, however, because individual behavior often plays a critical role. Studies assigning responsibility for job accidents, for example, generally attribute most of these accidents, at least in part, to worker behavior.

In contrast, many of the leading causes of death are more in the domain of individual behavior rather than regulatory action. The main determinants of heart disease are individual diet and exercise. Similarly, many cancer experts believe that cancer risks are largely due to diet, smoking, genetics, and other person-specific factors. Environmental exposures generally rank low among the total determinants of cancer. Strokes likewise are not generally the result of risk exposures subject to government regulation, although air pollution is a prominent exception that will affect the propensity to have a stroke. Many of the causes of death that are less

2. A long-standing issue in health economics has been the role of individual consumption decisions, such as exercise and smoking, in influencing health status. For a discussion of these factors, see Victor R. Fuchs, *The Health Economy* (Cambridge, Mass.: Harvard University Press, 1986).
prominent than accidents, ranging from pulmonary disease to homicides, are also matters over which there is little regulatory control.

**Measuring Mortality Risks**

The statistics in table 19.1 indicate how various causes of death affect their probability that one will die from a particular cause. However, societal concern with these various deaths may vary depending on the length of life lost. Table 19.2 presents a ranking of twelve leading causes of death based both on the probability of death and on lost life expectancy. The first set of statistics pertains to the probability of death, and in this ranking cardiovascular diseases rank first, cancer ranks second, and “all accidents” rank fifth.

Matters are somewhat different if one looks at the duration of life lost. The second set of columns in table 19.2 indicates the rank order for the lost life expectancy. These statistics indicate the length of life lost conditional upon dying from a particular disease. Moreover, the estimates take into account the role of discounting, as the lost life expectancy is the discounted number of life years lost, recognizing the role of pertinent time lags between the time of exposure to the risk and the time of death. Focusing on lost life expectancy, cardiovascular disease drops to twelfth, cancer drops to eighth, and the category of “all accidents” ranks sixth in importance.

The final set of statistics in table 19.2 multiplies the probability of death by the lost life expectancy to obtain the expected life years lost. These rankings move auto accidents and

<table>
<thead>
<tr>
<th>Rank</th>
<th>Probability of Death</th>
<th>Lost Life Expectancy</th>
<th>Expected Life Years Lost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cardiovascular disease</td>
<td>0.4478</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>Neoplasms (cancer)</td>
<td>0.2184</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>Pneumonia/influenza</td>
<td>0.0415</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>Obstructive pulmonary conditions</td>
<td>0.0396</td>
<td>11</td>
</tr>
<tr>
<td>5</td>
<td>All accidents</td>
<td>0.0315</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>Diabetes</td>
<td>0.0214</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>Auto accidents</td>
<td>0.0142</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>Liver disease/cirrhosis</td>
<td>0.0099</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>Suicide</td>
<td>0.0099</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>Homicide</td>
<td>0.0068</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>Perinatal conditions</td>
<td>0.0045</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>Congenital anomalies</td>
<td>0.0037</td>
<td>2</td>
</tr>
</tbody>
</table>

**Note:** Lost life expectancies (LLE) and expected years of life lost [E(YLL)] calculations incorporate 3 percent annual discounting and ten-year lags for diseases that may not occur immediately after exposure to stimulus.

“all accidents” much higher on the mortality risk scale and substantially narrow the absolute difference in the risk between prominent disease categories and accidents. Because government regulations extend lives but do not confer immortality, it is appropriate to consider the length of life at risk, and not simply the probability of death.

The Infeasibility of a No-Risk Society

These patterns do not suggest that regulation is unimportant, even though a fully successful regulatory program will not drastically alter death risks. What they do indicate is that our expectations regarding the overall effects of regulation should be tempered by an appreciation of their likely impact, even under a best-case scenario. Moreover, it should also be recognized that many of the most prominent sources of death risks are matters of individual choice, such as diet. Therefore, it seems inappropriate to adopt an extremist approach to government regulation in which we pursue a no-risk strategy while at the same time permitting risks of greater consequence to be incurred through individual action. Moreover, we need to place greater emphasis on policies such as hazard warnings and nutrition information that will foster better risk-averting decisions.

An interesting perspective on the variety of risks that we face is provided by the data in table 19.3. Information provided in that table lists a variety of activities that will increase one’s annual death risk by one chance in 1 million. This group includes some of the most highly regulated risks, as the risk of accident of living within five miles of a nuclear reactor for fifty years and the risk of smoking 1.4 cigarettes are tantamount to the risks one would face by riding ten miles on a bicycle, eating forty tablespoons of peanut butter, drinking Miami drinking water for one year, or eating a hundred charcoal-broiled steaks.

The risks that we have chosen to regulate are not altogether different from those that we incur daily as part of our normal existence. This similarity does not mean that the government should abandon regulation or that we should stop eating peanut butter. But it does suggest that we will not have a risk-free environment no matter what regulations we pursue. Moreover, given the many risk-taking choices we make daily within the domain in which we institute regulations, it would not be appropriate to require that they produce a risk-free environment.

The basic issue is one of balance. Society should pursue regulations that are in our best interests, taking into account both the beneficial aspects of regulations and the costs that they impose. These regulations may be stringent, or they may require that we have no regulation at all. The mere presence of a risk within the domain of a regulatory agency is not in and of itself a reason to institute a regulation. The key ingredient is that there must be some market failure to warrant government intervention. Moreover, there should be some ability of a regulatory policy to influence the risk outcome of interest.
The necessity of making difficult tradeoffs is especially pertinent with respect to homeland security. After the September 11, 2001, attack on the World Trade Center and the Pentagon, there was support for greater vigilance to combat terrorism. Reducing terrorism risks is certainly not a controversial objective. But how we reduce these risks may entail other costs and compromise other fundamental concerns. What is most notable about this policy context is that many policy measures require that we sacrifice civil liberties to reduce the risks of terrorism. Two objectives that some claim can never be compromised, civil liberties and safety, were at odds. Because the perceived terrorism threat had risen from the level we thought it was before 9/11, it would be desirable to strike a different balance than before, with a greater level of precaution and fewer civil liberties.

Figure 19.1 illustrates the policy choice. The pre-9/11 curve indicates the achievable combinations of civil liberties and expected terrorism losses. If the government increases the stringency of airport screenings and surveillance, it can reduce the expected terrorism losses.

### Table 19.3
Risks That Increase the Annual Death Risk by One in 1 Million

<table>
<thead>
<tr>
<th>Activity</th>
<th>Cause of Death</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoking 1.4 cigarettes</td>
<td>Cancer, heart disease</td>
</tr>
<tr>
<td>Drinking 0.5 liter of wine</td>
<td>Cirrhosis of the liver</td>
</tr>
<tr>
<td>Spending 1 hour in a coal mine</td>
<td>Black lung disease</td>
</tr>
<tr>
<td>Spending 3 hours in a coal mine</td>
<td>Accident</td>
</tr>
<tr>
<td>Living 2 days in New York or Boston</td>
<td>Air pollution</td>
</tr>
<tr>
<td>Traveling 6 minutes by canoe</td>
<td>Accident</td>
</tr>
<tr>
<td>Traveling 10 miles by bicycle</td>
<td>Accident</td>
</tr>
<tr>
<td>Traveling 150 miles by car</td>
<td>Accident</td>
</tr>
<tr>
<td>Flying 1,000 miles by jet</td>
<td>Accident</td>
</tr>
<tr>
<td>Flying 6,000 miles by jet</td>
<td>Cancer caused by cosmic radiation</td>
</tr>
<tr>
<td>Living 2 months in Denver</td>
<td>Cancer caused by cosmic radiation</td>
</tr>
<tr>
<td>Living 2 months in average stone or brick building</td>
<td>Cancer caused by natural radioactivity</td>
</tr>
<tr>
<td>One chest X-ray taken in a good hospital</td>
<td>Cancer caused by radiation</td>
</tr>
<tr>
<td>Living 2 months with a cigarette smoker</td>
<td>Cancer, heart disease</td>
</tr>
<tr>
<td>Eating 40 tablespoons of peanut butter</td>
<td>Liver cancer caused by aflatoxin B</td>
</tr>
<tr>
<td>Drinking Miami drinking water for 1 year</td>
<td>Cancer caused by chloroform</td>
</tr>
<tr>
<td>Drinking 30 12-oz. cans of diet soda</td>
<td>Cancer caused by saccharin</td>
</tr>
<tr>
<td>Living 5 years at site boundary of a nuclear power plant in the open</td>
<td>Cancer caused by radiation</td>
</tr>
<tr>
<td>Drinking 1,000 24-oz. soft drinks from banned plastic bottles</td>
<td>Cancer from acrylonitrile monomer</td>
</tr>
<tr>
<td>Living 20 years near PVC plant</td>
<td>Cancer caused by vinyl chloride (1976 standard)</td>
</tr>
<tr>
<td>Living 150 years within 20 miles of a nuclear power plant</td>
<td>Cancer caused by radiation</td>
</tr>
<tr>
<td>Eating 100 charcoal-broiled steaks</td>
<td>Cancer from benzopyrene</td>
</tr>
<tr>
<td>Risk of accident by living within 5 miles of a nuclear reactor for 50 years</td>
<td>Cancer caused by radiation</td>
</tr>
</tbody>
</table>

Before the 9/11 attack the optimal policy choice was at point A, with a high level of civil liberties and a low level of terrorism losses. After the attack, society has perceived a greater level of risk, as shown on the post-9/11 curve in figure 19.1. If the level of civil liberties remained unchanged, then the terrorism losses would soar to point B. As a result, it becomes desirable to increase the vigilance of airport screening and other security measures to get to some new optimal tradeoff at point C.

How far down on the post-9/11 curve we should go depends on two components of the policy choice problem. First, what is the shape of the post-9/11 tradeoff curve? If that curve is very flat and sacrificing civil liberties buys us very little added safety, then it will not be desirable to sacrifice many of our civil liberties. Second, what are society’s preferences with respect to being willing to trade off civil liberties for safety? Thus, even if sacrificing civil liberties will reduce terrorism risks substantially, if we place relatively little value on added safety compared to civil liberties, then it will not be desirable to move too far down from point B on the tradeoff curve.
The homeland security example embodies the two key components of risk policy decisions. The nature of the available tradeoff opportunities and the preferences with respect to these components jointly determine the optimal policy choice. Neither our preferences nor the opportunities alone dictate the best decision. That people may be willing to compromise civil liberties at all is illustrated by the tradeoffs people are willing to make between travel time and the intrusions of airport security. Suppose you were asked the following survey question:

One way of reducing terrorism risks to plane flights is better screening of passengers. The FBI has developed a profile of the chances that a passenger is a terrorist, taking into account the person’s age, race, gender, national origin, appearance, and baggage. Airlines could either screen all passengers, leading to additional delays in line, or they could screen passengers based on the profiling. People who are singled out will have to undergo an extra ten minutes of searches. You would not be singled out for such racial profiling. Would you favor terrorist risk profiling if the alternative was for you to wait in line an extra ten minutes so that all passengers could be screened? How would you feel if the wait was thirty minutes or sixty minutes?

Table 19.4 presents survey results that indicate how people’s reluctance to support a racial profiling policy diminishes with the extent of the wait. For a ten-minute wait, a majority of respondents oppose profiling and prefer that all passengers be screened equally. However, once the wait reaches an hour in line, about three-fourths of the sample favors targeting specific groups of passengers for screening. People are willing to trade off civil liberties against travel time, which is a much less compelling tradeoff than civil liberties against safety.

Table 19.4 presents survey results that indicate how people’s reluctance to support a racial profiling policy diminishes with the extent of the wait. For a ten-minute wait, a majority of respondents oppose profiling and prefer that all passengers be screened equally. However, once the wait reaches an hour in line, about three-fourths of the sample favors targeting specific groups of passengers for screening. People are willing to trade off civil liberties against travel time, which is a much less compelling tradeoff than civil liberties against safety.

The 9/11 attack also has raised public fears about airplane travel. Airline travel has traditionally been among the safest transportation modes, notwithstanding many people’s fear of flying and the recent experience with terrorist attacks. In 2001, the year of the 9/11 attack on the World Trade Center, there were 531 U.S. civil aviation accidental deaths. Even with this unprecedented catastrophe, the overall fatality risk per mile was still below that of automobiles in that year. Indeed, more people died on U.S. highways in September 2001 than in airplane crashes. In other recent years, travel on large airlines has been extremely safe,

Table 19.4
Attitude Toward Use of Terrorism Risk Profiles

<table>
<thead>
<tr>
<th>Delay in Line Due to Screening Time</th>
<th>Percent Who Favor Screening Affecting Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 minutes</td>
<td>44.7</td>
</tr>
<tr>
<td>30 minutes</td>
<td>55.3</td>
</tr>
<tr>
<td>60 minutes</td>
<td>73.9</td>
</tr>
</tbody>
</table>


3. The overall death toll of the 9/11 attacks, including those on the ground, was about 3,000.
with zero deaths in 2002, ninety-two deaths in 2000, twelve deaths in 1999, and one death in 1998.\(^4\)

Yet, despite this remarkable record of safety, public fears of airplane crash risks remain high. A number of factors are at work. Airplane crashes are highly publicized, dramatic events involving risks out of our control. By comparison, there is very little publicity given to how much people tend to travel by plane, which is the denominator in the risk probability calculation. Airplane risks also involve small probabilities, which people are prone to overestimating.

**Wealth and Risk**

Our demand for greater regulation in our lives has developed in part from the increased affluence of society.\(^5\) As our society has become richer, we have begun to place a greater value on individual health status and on efforts that can improve our physical well-being. These developments are not new. Sanitation levels and the cleanliness of food preparation, for example, were much greater in the early part of the twentieth century than they were several hundred years ago.

The influence of increased societal income is evident in the accident trends sketched in figure 19.2. These accident statistics give the accident rate per 100,000 population, and as can be seen, a general declining trend has taken place. Overall accident rates are down, as are accidents at home, at work, and in other public places, other than those involving motor vehicles. The main exception to the pattern of striking decline is that of motor vehicle accidents. Although the motor vehicle death rate is lower now than it was seventy years ago, the change has not been great. This comparative stability is not a reflection of a failure to improve the quality and design of cars. Indeed, dramatic improvements have taken place with respect to automobile safety. Rather, the accident statistics are on a population basis, which does not take into account the level of driving intensity. If one were to examine figures based on the risk per mile driven, then the real safety gains that have been achieved would be more apparent.

The existence of a downward accident trend provides a key element in terms of the historical perspective one should adopt in interpreting accident statistics. Regulatory agencies routinely announce this year’s accident rates, which, if all has gone well, are lower than they were the previous year. The agency officials then tout these statistics as evidence of the efficacy of the agency’s activities. Although this annual ritual may be effective political salesmanship, it ignores the fact that in all likelihood there would have continued to be a decline in the accident rate, as there was throughout the twentieth century. The main evidence of the

---

effectiveness of an agency should be a shift in the accident trend to a risk level below what it otherwise would have been.

The pattern of decline in accident rates also highlights the importance of technological progress. As a society we have benefited from a variety of technological improvements in our environment, and these have been a tremendous boost to our well-being. New technologies are often highly controversial and lead to demands that the risks be stringently regulated, but one should also recognize that new technologies have brought about many substantial benefits over the long run.  

**Figure 19.2**

Trends in Accidental Death Rates in the United States, 1930–2002


Irrationality and Biases in Risk Perception

Whereas the negative relationship between risk levels and increased societal wealth reflects a rational economic response to risk, situations involving risk and uncertainty are also well

---

known for the irrational decisions that they may generate. In particular, the risk statistics examined indicate for the most part that we are dealing with a low probability of events. Large risks, such as the risk of being killed in a car crash, are roughly one in 6,000 per year. Many other risks that we face are much smaller in magnitude; for example, the data in table 19.3 indicated that we had to drink a thousand 24-ounce soft drinks from banned plastic bottles in order to incur a cancer risk of one in 1 million. Small probabilities such as this are very difficult to think about. Most people have little experience in dealing with risks such as one in a million or one in 100,000 on a sufficiently regular basis that would enable them to obtain enough experience in dealing with such events so as to make sensible decisions. In some cases, the risk may not be understood at all, as the individual may be ignorant of the risk and consequently unable to make any decision with respect to it. For risks that are called to individuals’ attention, there are also difficulties that arise because of biases in risk perception.

Figure 19.3 illustrates the relationship between perceived risks (on the vertical axis) and actual risks (on the horizontal axis) of death. Individuals tend to overestimate the risks associated with lower-probability events, such as botulism, tornadoes, and floods. In contrast, there is a tendency to underestimate the risks associated with higher-risk events, such as cancer, heart disease, and stroke. Not all of these areas of misperception may represent errors in judgment. W. Kip Viscusi, Jahn Hakes, and Alan Carlin found that differences in the length of life lost associated with these different causes of death accounted for much of the apparent bias in risk perceptions. Similarly, Daniel Benjamin and William Dougan found that much of the difference in assessed life expectancy loss could be accounted for by the particular age of the respondent, which should affect risk beliefs.

Even after making such adjustments, some misperceptions of risk will remain, implying that market decisions may not be optimal. However, additional regulation will not be required in all cases. For example, if risk perceptions are already too high, then the level of safety provided by the market will be excessive, as economic agents will be responding to exaggerated risk perceptions.

The overestimation of low-probability events also has substantial implications for government policy. To the extent that there is an alarmist reaction to small risks that are called


to our attention, and if these pressures in turn are exerted on the policymakers responsible for risk regulation, society may end up devoting too many resources to small risks that are not of great consequence.

An interesting policy question is how the government should respond, if at all, to public misperceptions. Suppose, for example, that the public greatly overestimates the risks associated with hazardous waste sites. Should the government respond to these fears because, presumably, in a democratic society the government action should reflect the interest of the citizenry? Alternatively, one might view the government as taking a more responsible role in that it should attempt to educate the public concerning the overly alarmist reactions that it has. Moreover, ideally the government should be a steadying force in such contexts. If the general public underestimates the risk, one presumably would not expect the government to be idle and to let citizens incur risks unknowingly. An important function of the government is to acquire more scientific information than is feasible for an individual to obtain, to com-
municate this information effectively to the public, and to issue regulations that are needed to control the real risks that are present.

A related result that has emerged in the risk-perception literature is that highly publicized events often are associated with substantial risk perceptions, even though the risks involved may not be great. This finding is not a sign of individual irrationality. Typically the events themselves rather than frequency statistics are publicized. We learn that a number of people have recently been killed by a tornado, but we are not given a sense of the frequency of these events other than the fact that coverage of tornado victims in the newspaper occurs much more often than coverage of asthma victims. Because of the sensitivity of risk perceptions to the amount of publicity as well as the level of the risk, the pressures that will be exerted on risk regulation agencies will not necessarily be in line with the direction that fosters society’s best interest. We will take as society’s objective the promotion of societal welfare based on the true risk levels, not the risk levels as they may be perceived by society more generally.

Even if we are equipped with this knowledge of how people err in their risk beliefs, it is sometimes difficult to tell whether we are overreacting to seemingly small probabilities. In 2002 a sniper in Washington, D.C., created extraordinary public fears that led football teams to cancel outdoor practices and thousands to change their daily routine. The daily risks were seemingly small—surely under a one-in-a-million chance of being killed, based on the sniper’s past record. But what if one day the sniper decided to kill dozens of people, not just one? Then the odds would become much more threatening. The chance that one would be shot by the sniper consequently was not well understood, since we have a small sample of observations that might not reflect accurately the extent of the risk.

Just as sniper risks pose challenges for our personal decisions, government regulators must routinely confront policy decisions in which the risks may be great but are not well understood. Reacting to worst-case scenarios will distract us from real risks that we face, but failing to attend to emerging hazards that are not yet well understood poses dangers as well.

Policy Evaluation

As a discussion of the various social regulation agencies will indicate, the stated objectives of these agencies are often quite narrow. The Clean Air Act, for example, forbids the EPA to consider cost considerations when setting air pollution standards, and the Supreme Court has explicitly prohibited OSHA from basing its regulations on a benefit-cost analysis. Nevertheless, some balancing is required; otherwise the cost associated with the regulations could easily shut down the entire economy. Moreover, as a practical matter, cost considerations and the need for balancing enter in a variety of ways. Agencies may phase in a particularly onerous regulation. Alternatively, the agency may choose not to adopt regulations that threaten the viability of an industry or employment in a local area.
Regulatory Standards

By far the most stringent standards that promote a balancing of societal interests are those that have been imposed through regulatory oversight mechanisms of the executive branch.9 The Ford administration instituted a requirement that the cost and inflationary impact of regulations be assessed. Under the Carter administration these requirements were extended to require that agencies demonstrate the cost effectiveness of their regulatory proposals. The Reagan administration went even further, requiring that agencies demonstrate that the benefits of the regulation exceed the costs imposed except when doing so would violate the legislative mandate of the agency. Moreover, even when there is such a legislative mandate, the agency must still calculate benefits and costs and submit these results to the Office of Management and Budget (OMB) for its review. The Bush, Clinton, and George W. Bush administrations have continued these policies.

The nature of the policy evaluation tools being discussed is not unique to the risk and environmental area. Procedures for assessing the benefits and costs of government policies have been in operation in the government for decades. Both the Army Corps of Engineers and the U.S. Department of Interior have been using benefit-cost analysis to govern their design and evaluation of water resources projects for several decades.

Benefit-Cost Analysis

The importance of using economic analysis to assess risk regulations can be illustrated by examining some statistics on the efficacy of differing levels of stringency of arsenic regulation in the workplace. Table 19.5 provides pertinent statistics for three different levels of stringency of a standard—loose, medium, or tight. As the second column of the table indicates, lower levels of exposure to arsenic are associated with tighter standards. With increased tightness come added costs. The third column in table 19.5 gives one measure of the cost where it has been put into cost-effectiveness terms, in particular the cost-per-unit benefit achieved. In this case the measure is in terms of the cost per statistical life saved by the policy. This average cost-per-life-saved figure ranges from $1.25 million to $5.63 million, reflecting a substantial but by no means enormous variation in the efficacy of the policy.

In contrast, estimates of the marginal cost per life saved such as appear in the final column of table 19.5 indicate that successive tightening of the standard becomes very expensive in terms of the lives that are saved per dollar expended. In the case of tight levels of the standard, the marginal cost imposed per life saved is $68.1 million, which as we will see for the results in chapter 20 is out of line with what a reasonable benefit value for life is. In this case,

the substantial acceleration in the marginal cost-per-life values was not as apparent when the agency focused on average cost-per-life figures.

In some cases, the fallacy of focusing on averages rather than marginal changes is even more dramatic. In a policy decision involving one of the authors, OSHA was considering differing levels of stringency for a standard that would control ethylene oxide exposures, which occur primarily among hospital workers who are involved in the cleaning of surgical equipment. OSHA’s regulatory analysis included calculations of the average cost per case of cancer prevented, but the officials responsible for the calculation had not noted that the last incremental tightening of the standard produced no reduction in cancer cases whatsoever. The marginal cost per case of cancer prevented in this case was actually infinite for the tightest level of the standard being considered. In contrast, the average cost per case of cancer remained fairly stable, thus disguising the inefficiency being created.

The Role of Heterogeneity

Examination of the benefits and costs of regulatory actions is important not only with respect to the stringency of any particular regulation, but also with respect to distinctions across different situations in terms of the stringency of the regulations. For example, pollution standards are generally set at a more stringent level for new enterprises than for existing enterprises, in part because the cost of compliance when introducing a new technology is believed to be less than the cost of adding pollution control devices to existing technology. There is the potential problem, however, of “new source bias.” Although we want regulatory standards for new pollution sources to be more stringent so that the marginal costs are equalized across different kinds of facilities, we could err too much in this direction, as we will discuss in subsequent chapters.10 Very often, in addition, one might want to make a distinction across industries in terms of the level of the standard because differences in industry technology imply different marginal cost curves. If the marginal benefit curve were the same in all

industries, higher marginal cost curves would shift the optimal level of environmental quality to a lower level, implying a less stringent regulatory regime.

The statistics presented in table 19.6 illustrate how in practice one might approach such differentiation of a regulation. The health impact of concern here is the effect of noise pollution in the workplace on workers’ hearing. In particular, the cases being examined are those involving a 25-decibel hearing threshold loss after twenty years of exposure to the noise. The statistics for the seventeen different industries represented indicate an order of magnitude variation in the cost per worker protected by differing levels of the standard. Suppose that one had to pick between setting the standard level at 80 decibels and at 90 decibels. In addition, for concreteness, let us assume that we are able to determine that a case of hearing loss has a benefit value of $200,000. A 90-decibel standard is justified in this case for all industries listed above “petroleum and coal products” in table 19.5. Moreover, a more stringent 80-decibel standard is warranted for all industries listed above “printing and publishing” in the table. Thus the tighter 80-decibel standard can be justified in situations where the cost of protecting workers tends to be less. (In this instance there are also other policy options, such as protective equipment, so that hearing loss can be prevented for all workers.)

In general, policymakers should attempt to exploit cost variations such as this and take them into account when designing regulatory policies. Differentiated standards that recog-

### Table 19.6

<table>
<thead>
<tr>
<th>Industry</th>
<th>90 Decibels</th>
<th>80 Decibels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electrical equipment and supplies</td>
<td>19</td>
<td>39</td>
</tr>
<tr>
<td>Rubber and plastics products</td>
<td>38</td>
<td>68</td>
</tr>
<tr>
<td>Stone, clay, and glass products</td>
<td>53</td>
<td>96</td>
</tr>
<tr>
<td>Paper and allied products</td>
<td>62</td>
<td>78</td>
</tr>
<tr>
<td>Food and kindred products</td>
<td>75</td>
<td>179</td>
</tr>
<tr>
<td>Chemicals and allied products</td>
<td>80</td>
<td>132</td>
</tr>
<tr>
<td>Transportation equipment</td>
<td>87</td>
<td>111</td>
</tr>
<tr>
<td>Tobacco manufactures</td>
<td>104</td>
<td>200</td>
</tr>
<tr>
<td>Printing and publishing</td>
<td>108</td>
<td>215</td>
</tr>
<tr>
<td>Electric, gas, and sanitary services</td>
<td>137</td>
<td>189</td>
</tr>
<tr>
<td>Furniture and fixtures</td>
<td>150</td>
<td>151</td>
</tr>
<tr>
<td>Fabricated metal products</td>
<td>192</td>
<td>188</td>
</tr>
<tr>
<td>Petroleum and coal products</td>
<td>215</td>
<td>257</td>
</tr>
<tr>
<td>Primary metal industries</td>
<td>218</td>
<td>372</td>
</tr>
<tr>
<td>Textile mill products</td>
<td>227</td>
<td>395</td>
</tr>
<tr>
<td>Lumber and wood products</td>
<td>228</td>
<td>303</td>
</tr>
<tr>
<td>Machinery, except electrical</td>
<td>233</td>
<td>245</td>
</tr>
<tr>
<td><strong>Weighted average</strong></td>
<td><strong>119</strong></td>
<td><strong>169</strong></td>
</tr>
</tbody>
</table>

nize these cost differences will produce greater net benefits to society than those that do not. Uniform standards miss opportunities to promote safety and environmental quality in situations where it is cheap to do so, and impose substantial burdens in situations where the cost of providing environmental quality and safety is higher. Society should take advantage of differences in the ability to produce environmental quality and safety, just as we take advantage of other productive capabilities with respect to other goods and services that an economy provides.

Uncertainty and Conservatism

Most situations of risk regulation involve these elements but are not well understood. Typical regulation of carcinogens, for example, is based on laboratory tests involving animals. To make the leap from animals to humans, one must adjust for differences in the size of the dosage of a particular chemical, differences in body weight and size, and differences in surface area, as well as possible differences in human as opposed to animal responses. Moreover, the results with different animal species may imply differing levels of riskiness that must then be assessed in such a manner that we can draw meaningful inferences for humans. Even in situations where we can reach a consensus on these issues, there is often a debate as to how one should model the risk relationships. For example, should one use a linear dose-response relationship or a nonlinear model? The fact that uncertainty exists does not imply that the risks are unimportant or should be ignored, but it does create an additional element that must be addressed in the course of risk regulation.

Even in the case of relatively common risks about which much is known, the range of uncertainty may be substantial. Often governmental agencies use some upper bound value of what the risk might be. By erring on the side of conservatism, in effect government agencies distort the true risk levels by differing amounts. We may be in a situation where a lower risk is the subject of more stringent regulation, not because it imposes a greater expected health loss, but because less is known about it.

The Role of Risk Ambiguity

If one were dealing with a single-trial situation in which one rarely incurred risk, the precision of the risk judgment would not enter. Uncertainty should not be a concern in the case of one-period decisions.

This principle can be traced back to the well-known Ellsberg paradox.\footnote{The original Ellsberg paradox is discussed in the article by Daniel Ellsberg, “Risk, Ambiguity, and the Savage Axioms,” \textit{Quarterly Journal of Economics} 75 (1961): 643–69.} Suppose that you face two different urns, each of which contains red balls and white balls. Urn 1 contains fifty
red balls and fifty white balls. Urn 2 contains an unknown mixture of red and white balls. Each urn contains a hundred balls, and you cannot see the contents of either urn. Suppose that you will win a prize if you can correctly guess the color of the ball that will be drawn. Which urn would you pick from and what color would you name?

The correct answer is that you should be indifferent between the two situations. Drawing from urn 1 offers you a known 50 percent chance of winning the prize. Because you did not know the composition of urn 2, you face an equivalent 50/50 chance of winning, by drawing from that urn, irrespective of the color of the ball you pick.

Most people confronted with the choice between the two urns believe that urn 1 is preferable because it offers a “hard” probability of success. However, the chances of success can be converted to an equally “hard” probability in the case of urn 2. Suppose you were to flip a fair coin before naming the color of the ball, and you had picked red if the outcome was heads and white if the outcome was tails. In that situation, you also face the same kind of precise probability of a 50/50 chance of success as you do with urn 1.

Although you should be indifferent between the two urns in the case of a single trial, this would not be the case if there were multiple trials. Suppose that you will be engaged in ten draws from the urn, where you have to replace the balls after each trial. From urn 1 you know you have a 50/50 chance of success on each trial, where this is a precisely understood probability. In contrast, from urn 2 the probability begins initially as a 50 percent chance of naming the correct ball, but with successive trials this probability will increase. For example, if the first three balls that you draw are red, then you will have acquired some information about the composition of that urn. In particular, the odds that there is a disproportionate number of red balls in that urn will have gone up, so that on successive trials your expected chance of receiving a prize after naming a red ball will be greater than 50/50. Even though this probability will remain a “soft” probability, it will still be in your interest to draw from the uncertain urn.

The same type of principle embodied in this example is true more generally. For one-shot decisions, the precision of the risk is not a matter of consequence. But in sequential decisions in which learning is possible and in which you can revise your decisions over time, it is preferable to have a situation of uncertainty rather than to have a precisely understood risk. In situations of uncertainty we can alter a course of action if the risk turns out to be different than we had anticipated originally.

In regulatory contexts, what this result implies is that the stringency of our regulation may depend in large part on uncertainty, but we will not necessarily respond in a conservative manner to this uncertainty. If you must take action now to avoid an environmental catastrophe, then uncertainty is irrelevant. The mean risk should be your guide. However, if we can learn about how serious the problem is and take effective action in the future, then it will generally be preferable to make less of a regulatory commitment than one would if this were a one-shot decision.
The conservatism approach to regulatory analysis runs the danger of confusing risk analysis with risk management. Ideally, the scientific analysis underlying regulatory policies should not be distorted by biases and conservatism factors. Policymakers should be aware of the true risks posed by different kinds of exposures so that we can make comparative judgments across different regulatory alternatives. Otherwise, we run the danger of distorting our policy mix and focusing attention on hazards that offer few expected payoffs but are not well understood.

The Role of Political Factors

Although benefit-cost analysis provides a convenient normative reference point for the determination of social regulation policies, in practice other political forces may be more instrumental. In particular, the same kinds of economic interests that influence the setting of economic regulations in a manner that does not maximize social efficiency also are at work in determining the structure of risk and environmental regulations. The Stigler/Peltzman/Becker models have applicability to social regulation as well.

Economic Models of Environmental Policies

One class of regulatory policies that has received very intense scrutiny by economists such as Robert Crandall, Peter Pashigian, Joseph Kalt, and Mark Zupan is the determination of environmental policies. In each case they have documented that a driving force behind the congressional voting over key environmental provisions that govern regulatory policy has been dictated by the economic stakes involved. A chief source of these differences is regional. Representatives of districts from the declining areas of the Northeast have in particular used regulatory policies to limit the degree to which the emerging economic regions of the United States could compete with them.

Consider, for example, the results presented in the analysis of congressional voting by Crandall, which are summarized in table 19.7.\textsuperscript{12} Pashigian derived results that are more detailed and similar in character.\textsuperscript{13} What Crandall found is that the current levels of air pollution and water pollution have little effect on the way in which a congressional delegation voted. The ones that were of greater consequence were more closely related to the economic interests involved.

The delegations with a larger share of Republicans in their districts were more likely to vote against environmental and energy policies, where the variable of interest in the

\textsuperscript{12} This discussion is based on Robert Crandall, \textit{Controlling Industrial Pollution: The Economics and Politics of Clean Air} (Washington, D.C.: Brookings Institution, 1983).

analysis was the percentage of the delegation voting “right” on environmental and energy issues. This result is consistent with the greater orientation of the Republican party toward business interests, which are more closely linked to the economic impacts of such regulation. Members of Congress representing districts with a large percentage of the national parks also are likely to vote no, perhaps because further expansion in these parks will take more land out of the economic base. The income level is positively linked with votes for environmental and energy issues, reflecting the fact that environmental quality is a normal economic good for which one’s preferences will increase with one’s income status.

The two key variables are the final ones in table 19.7. States with substantial income growth are more likely to vote against environmental controls, because these controls will restrict the expansion of industry in these states. In contrast, the states in the Frost Belt of the North Central, New England, and Middle Atlantic states are more likely to vote for strict environmental controls, because these controls will hit hardest on the newly emerging sources in other regions of the country. In particular, because the main structure of EPA policy imposes more stringent requirements on new sources of pollution rather than existing sources, more stringent environmental regulation generally implies that there will be a differential incidence of costs on newly emerging industries and regions with substantial economic expansion. As a consequence there are important distributional issues at stake when voting on the stringency of such environmental policies.

The result was that the air pollution regulations that were promulgated imposed a variety of requirements with substantial redistributional impacts. For example, the legislation required that firms install scrubbers to decrease their sulfur dioxide pollution. An alternative means of achieving the same objective would have been for firms and power plants to rely more on the western low-sulfur coal. However, this option would have hurt the coal-producing regions of the Midwest and Northeast, and, as a result, the representatives from these regions opposed giving firms the more performance-oriented standard. By requiring

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Sign</th>
<th>Explanatory Variable</th>
<th>Estimated Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Air pollution</td>
<td>?</td>
<td>No significant effect</td>
<td></td>
</tr>
<tr>
<td>Water quality</td>
<td>?</td>
<td>No significant effect</td>
<td></td>
</tr>
<tr>
<td>Party</td>
<td>?</td>
<td>Significant negative</td>
<td></td>
</tr>
<tr>
<td>Natural lands</td>
<td>?</td>
<td>Significant negative</td>
<td></td>
</tr>
<tr>
<td>Income</td>
<td>+</td>
<td>Significant positive (1 out of 9)</td>
<td></td>
</tr>
<tr>
<td>Income growth</td>
<td>-</td>
<td>Significant negative</td>
<td></td>
</tr>
<tr>
<td>Frost Belt</td>
<td>+</td>
<td>Significant positive (3 out of 9)</td>
<td></td>
</tr>
</tbody>
</table>
that firms meet a technology-based standard that did not permit the substitution of types of coal to meet the pollution target, they could protect the economic interest of the eastern coal producers.

Although there have been a number of other analyses indicating a variety of social regulations as being subject to the influence of such political factors, there remains a question in the literature as to the extent to which economic self-interest is the main determinant of regulatory policy. In an effort to explore these determinants more fully, Kalt and Zupan have developed a model in which they analyze two different sets of influences—economic self-interest as reflected in the standard capture theory models of regulation and ideology, which are more closely associated with the more normative approaches to regulation. The focus of their analysis was on the voting for coal strip-mining regulations. The requirement to restore strip-mine land to the premining state imposed substantial costs. The annual costs associated with this policy were believed to be $1.4 billion, where surface miners would bear roughly two-thirds of the cost and consumers would bear roughly one-third. The environmental gains were believed to be $1.3 billion per year. Overall, this policy would lead to a transfer from surface mine producers and coal consumers to underground producers and environmental consumers.

**Voting Patterns**

Under the capture theory, factors such as altruism, public interest objectives, and civic duty are believed to be insignificant determinants of voting patterns. In contrast, the ideological models of voting use the social objectives of the political actors as the objective function that is being maximized in the voting process. A potential role for ideological voting enters, because the market for controlling legislators meets infrequently. We vote for representatives every two years and for senators every six years. Moreover, the voters have very weak incentives to become informed, and it is often possible for representatives to shirk their responsibility to represent the voters’ interests in such a situation.

Table 19.8 summarizes the series of factors considered in the Kalt and Zupan analysis. These results suggest that both the capture and ideology models illuminate aspects of the determinants of the anti-strip-mining vote. In particular, members of Congress are more likely to vote in favor of anti-strip-mining regulations the higher the level of coal reserves in their state, the greater the environmental group membership in their state, the greater the unreclaimed value of the land that is stripped and not restored, and the greater the value of underground coal reserves. They are more likely to vote against anti-strip-mining regulations if there is a high regulation-induced mine cost for the state, a high amount of surface-coal reserves, or a high concentration of consumer groups in the state. Because it is the

---

The final column of estimates adds a capture theory variable, which is the representative’s pro-environmental voting record. This variable has a positive effect on voting in favor of anti-strip-mining legislation, and the magnitude of this effect is substantial. Kalt and Zupan interpret this result as indicating that ideology has an independent influence above and beyond capture. Indeed, the addition of this variable increases the percentage of the variation explained by the equation by 29 percent.

Interpretation of this measure as reflecting simply capture concerns also raises some difficulties, however. It may be that the pro-environmental voting record serves to reflect the influence of omitted aspects of the capture model. In particular, the fact that a legislator has voted in favor of environmental measures in the past may be a consequence of the influence of a variety of economic self-interest variables pertaining to the stakes that his voters have in such legislation. This variable may consequently serve as a proxy for a host of concerns that were not explicitly included in the equation. The pro-environmental voting record consequently may not reflect concerns restricted to ideology, but may reflect the past influence of capture variables that will also be at work with respect to the anti-strip-mining vote.

Interpretation of these models is also difficult because of the role of logrolling. If members of Congress exchange votes, agreeing to vote for measures they do not support in return for support of legislation in which they have a substantial interest, then the congressional voting patterns may be a misleading index of the impact of the political forces at work.

To the extent that there is a controversy in the economics literature over these issues, it is not with respect to whether capture models are of consequence, but the extent to which these

### Table 19.8
Factors Affecting Voting Patterns for Strip-Mining Regulation

<table>
<thead>
<tr>
<th>Explanatory Variables</th>
<th>Capture Model</th>
<th>Capture and Ideology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pro-environmental vote</td>
<td>NA</td>
<td>Significant positive</td>
</tr>
<tr>
<td>Regulation-induced mine cost</td>
<td>Significant negative</td>
<td>Significant negative</td>
</tr>
<tr>
<td>Surface reserves</td>
<td>Significant negative</td>
<td>Significant negative</td>
</tr>
<tr>
<td>Underground coal reserves</td>
<td>Significant positive</td>
<td>Significant positive</td>
</tr>
<tr>
<td>SPLIT rights for strippable land</td>
<td>Negative</td>
<td>Positive</td>
</tr>
<tr>
<td>Environmental group membership in state</td>
<td>Significant positive</td>
<td>Positive</td>
</tr>
<tr>
<td>Unreclaimed value of land (stripped not restored)</td>
<td>Significant positive</td>
<td>Significant positive</td>
</tr>
<tr>
<td>Coal consumption in state</td>
<td>Negative</td>
<td>Significant negative</td>
</tr>
<tr>
<td>Surface coal mine groups</td>
<td>Negative</td>
<td>Positive</td>
</tr>
<tr>
<td>Underground coal mine groups</td>
<td>Positive</td>
<td>Positive</td>
</tr>
<tr>
<td>Environmental groups</td>
<td>Significant positive</td>
<td>Negative</td>
</tr>
<tr>
<td>Consumer groups in state</td>
<td>Significant negative</td>
<td>Negative</td>
</tr>
</tbody>
</table>

surface-coal industry that will lose from the regulation and the underground coal industry that will benefit, these influences follow the pattern one would expect.
influences are at work. The extreme versions of these models claim that the capture theory accounts for all of the pattern of voting and regulation. Even if such economic self-interests are only major contributors to the outcomes and not the sole contributors, they should nevertheless be taken into account as an important factor driving the determination of regulatory policy.

Summary and Overview of Part III

Determining the optimal environmental policy typically involves only a straightforward application of benefit-cost analysis. Perhaps the main role for economists is in defining what these benefits are, particularly since environmental benefits typically are not explicitly traded on the market.

As a practical matter, environmental policies tend to be governed by a host of political factors that bear little relationship to the normative guidelines that might be prescribed by economists. Economic analysis nevertheless has a role to play with respect to these forces as well, as it illuminates how the different payoffs to the political actors have motivated the environmental policies that have emerged over the past two decades.

In subsequent chapters we will explore a representative mix of issues arising in the environment and safety area. This examination is not intended to be exhaustive, but we will address many of the problems that arise in this whole class of regulatory policies.

Chapter 20 begins with a discussion of the task of setting a price on the regulatory impacts. For the most part, social regulation efforts deal with situations in which there are no existing markets available to set a price. We can draw marginal benefit curves and total benefit curves as convenient abstractions, but ultimately we would need to know how much benefit society does place on environmental quality or reduced birth defects before we can make any policy judgments that are more than hypothetical. Over the past two decades, economists have devoted considerable attention to devising methodologies by which we can establish these prices. Indeed, the principal focus has perhaps been on the area that one might have thought would be least amenable to economic analysis, which is the economic value of reducing risk to human life. The implications of chapter 20 are pertinent to all subsequent discussions, inasmuch as they provide the best reference point that can be used for assessing the appropriate stringency of the regulatory policy.

Chapter 21 focuses on the specific cases of environmental regulation. The regulatory strategy there has been largely to set standards and to issue permits that allow firms to engage in a particular amount of pollution. The major difficulty is that compliance with these standards may be quite costly, and there is a need to set a balance between the economic costs to the firm and the benefits that will be achieved for society. Environmental protection efforts are utilizing increasingly imaginative means to strike such a balance, and we explore such approaches in this chapter.
In chapter 22 we turn to a series of product safety regulations. These regulations also deal with risks that are, at least in part, the result of market transactions. The diverse forms of product regulation include auto safety regulation, food and drug regulations, general product safety regulations by the Consumer Product Safety Commission, and the impact of the liability system on product safety. This diversity has created a need for coordination among these institutions. Moreover, in the product safety area in particular, the role of “moral hazard” has played a central role in economic analyses. Mandating seat belts may have seemed like an attractive regulatory policy, but if nobody uses the seat belts we will not experience any gain in safety. Other more complex behavioral responses are also possible, and we will examine these as well. Perhaps the main message of the product safety analysis is that the outcomes of regulatory policy are not dictated by regulatory technology but instead involve a complex interaction of technological and behavioral responses.

Job safety regulations, the focus of chapter 23, also involve markets that are in existence. In particular, workers incur job risks as part of the employment relationship, and in many cases these risks are understood by the worker who is facing the risk. Moreover, in return for bearing the risk, the worker will receive additional wages, as well as workers’ compensation benefits after an injury. The presence of a market makes regulation of job safety somewhat different in character from environmental regulations.

Moreover, the main regulatory issues in the job safety area seem to be quite different as well. Because of the presence of a market before the advent of regulation, the initial wave of job safety regulation met with considerable resistance. In addition, perhaps the weakest link in the regulatory effort in the job safety area is the lax enforcement effort. In this area in particular, implementation aspects of regulatory policy loom particularly large.

Questions and Problems

1. Contrast the kinds of market failure that lead to regulation of automobile safety as opposed to regulation of automobile emissions. In which case does a market exist that, if it were operating perfectly, would promote an efficient outcome? What kinds of impediments might lead to market failure?

2. Officials of the Environmental Protection Agency frequently argue that we should not discount the benefits of environmental regulation. Their argument is that it is acceptable to discount financial impacts, but it is not acceptable to discount health effects and environmental outcomes that occur in the future. What counterarguments can you muster?

3. What are the alternative mechanisms by which the government has intervened in the health, safety, and environmental area? How does the choice of mechanisms across agencies such as OSHA, the EPA, and the FDA vary with the regulatory context? Do you see any rationale, for example, for the difference in regulatory approach?

4. If future generations were able to contract with us, they would presumably bargain for a higher level of environmental quality than we would choose to leave them without such compensation.
To what extent should society recognize these future interests? Should our views be affected at all by the fact that future generations will be more affluent and will probably have a higher standard of living? How will this difference in wealth affect their willingness to pay for environmental quality, as opposed to that of current generations? What will be the equity effects in terms of income distribution across generations?

5. Should society react differently to voluntary and involuntary risks? Which risk would you regulate more stringently?

**Recommended Reading**

Establishing the appropriate degree of social regulation requires that we set a price for what the regulation produces. In the case of environmental regulation, we need to know what the value to society of additional pollution reduction will be before we can set the stringency of the standard. In the case of health and safety regulations, we need to know the value of preventing additional risks to life and health.

Although one can sidestep these issues in part by relying on cost-effectiveness analysis in which we calculate the cost per unit of social benefit achieved, such as the cost per expected life saved, the most that can be done with cost-effectiveness analysis is to weed out the truly bad projects. Ultimately, some judgment must be made with respect to the amount of resources society is willing to commit to a particular area of social regulation. In practice, this trade-off may be implicit, as government officials may make subjective judgments with respect to whether a policy is too onerous. Implicit overall judgments come close to setting an implicit value on life, health, or pollution, but often these judgments may result in serious imbalances across policy areas.

One reason for these imbalances is that taking trade-offs into consideration in an ad hoc manner may be a highly imperfect process. The Occupational Safety and Health Administration (OSHA), for example, attempts to avoid regulatory actions that will lead to the shutdown of a particular firm. The Environmental Protection Agency (EPA) has similar concerns, as it has made an effort to phase in pollution requirements for the steel industry. When EPA policies would have serious repercussions for local employment, it has sought the advice of the residents in the affected area. Often the compromise that is reached is that the requirements will be phased in over a long period of time, an approach that will reduce the costs of transition and that can better be accommodated, given the normal process of replacing capital equipment over time. This practice of phasing in requirements has also been followed for automobile regulation, where pollution control requirements and major safety innovations, such as air bag requirements, have been imposed with fairly long lead times so that the industry can adjust to the standards.

The focus of this chapter will be on how society can establish a more formal, systematic, and uniform basis for establishing trade-offs between the resources expended and the benefits achieved through social regulation efforts. For most economic commodities, this would be a straightforward process. The U.S. Bureau of Labor Statistics gathers price information on hundreds of commodities, so that finding out the price of a market-traded good is a fairly trivial undertaking. In contrast, social regulation efforts for the most part deal with commodities that are not traded explicitly in markets. Indeed, from a policy standpoint, it is in large part because of the lack of explicit trade that we have instituted government regulation in these areas. Victims of pollution do not sell the right to pollute to the firms that impose these pollution costs. Future generations that will suffer the ill effects of genetic damage likewise do not contract with current generations, the operators of genetic engineering experiments, or the firms that expose pregnant women to high levels of radiation. Nevertheless, to
the extent that it is possible, we would like to establish a market reference point for how much of a resource commitment we should make to preventing these outcomes, so that we can get a better sense of the degree to which various forms of social regulation should be pursued. We will use the valuation of the risks to life as the case study for considering how the government can value the benefits associated with regulations affecting health and the environment.

Two approaches have been used. The first is to estimate the implicit prices for these social risk commodities that may be traded implicitly in markets. Most important is that workers receive additional premiums for the risks they face on the job, and the wage trade-offs they receive can be used to establish an appropriate trade-off rate. A second general approach is to ask people through an interview context how much they value a particular health outcome. This methodology may have greater problems with respect to reliability, but it has the advantage that one can obtain trade-off information regarding a wide range of policy outcomes.

Policy Evaluation Principles

Suppose that this evening you will be crossing the street, and that you have one chance in 10,000 of being struck by a bus and killed instantaneously. We will offer you the opportunity to buy out of this risk for a cash payment now. For purposes of this calculation, you can assume that your credit is good and that, if necessary, you can draw on either your parents’ or your future resources. To put the risk in perspective, a probability of death of one chance in 10,000 is comparable to the average fatality risk faced each year by the average worker in the construction industry. How much would you be willing to pay for eliminating this risk?

This kind of thought process is exactly what the government should go through when thinking about how far to push various social regulation efforts. In particular, the main matter of concern is society’s total willingness to pay for eliminating small probabilities of death or adverse health effects. Thus we are not interested in the dollar value of your future earnings that will be lost, although this of course will be relevant to how you think about the calculation. In addition, we are not interested in how much you are willing to pay to avoid certain death. The level of the probability risk involved with certain death dwarfs that associated with small risk events by such an extent that the qualitative aspects of the risk event are quite different. It is noteworthy, for example, that society views suicide with disfavor, but the taking of small risks, such as the decision to drive a compact car rather than a larger car that offers greater safety, is generally viewed as being acceptable.

Let us now take your response to the willingness-to-pay question that we have asked and convert it into a value of life. What we mean by the value-of-life terminology is the value that you would be willing to pay to prevent a statistical death. To emphasize this dependence on probabilities, we will use the term value of a statistical life. This amount is straightforward to calculate. To calculate this magnitude, one simply divides your willingness-to-pay response by the level of the risk that you are reducing, or

\[
\text{Value of statistical life} = \frac{\text{Willingness to pay}}{\text{Size of risk reduction}}. \tag{20.1}
\]

This equation gives the amount you would be willing to pay per unit of mortality risk. For the specific values given in the example we considered, the value-of-statistical-life number can be calculated as

\[
\text{Value of statistical life} = \frac{\text{Willingness to pay}}{1/10,000} \tag{20.2}
\]

or

\[
\text{Value of statistical life} = 10,000 \times \text{Willingness to pay}. \tag{20.3}
\]

An alternative way of thinking about the value of life is the following. Consider a group of 10,000 people in a sports arena, one of whom will be killed at random. We will assume that such a death risk has no desirable features. There will be one expected death. If each person were willing to contribute $500 to eliminate the risk, then the amount that could be raised to prevent one expected death would be 10,000 multiplied by $500, or $5 million. This calculation is identical to that in equation 20.3.

Your value of life implicit in the response you gave is consequently 10,000 times the amount of your response. Table 20.1 gives different value-of-statistical-life estimates, depending on the level of your answer. If there is no finite amount of money that you would be willing to pay to prevent this risk, and if you were willing to devote all of your present and future resources to eliminate it (presumably retaining enough for minimal subsistence), then it would be safe to say that you place an infinite value on risks to your life, or at least a value that is very, very large. Any finite response below this amount implies that you would be willing to accept a finite value of statistical life or make a risk-dollar trade-off when confronted with a life-extending decision. When viewed in this manner, making a risk-dollar trade-off does not appear to be particularly controversial. Indeed, one might appear to be somewhat irrational if one were willing to expend all of one’s future resources to prevent small risks of death, particularly given the fact that we make such trade-offs daily, as some of the risk statistics in chapter 19 indicated.

For the finite value of life responses, a willingness to pay $1,000 to prevent a risk of death of one chance in 10,000 implies a value of statistical life of $10 million. A response of $500
to prevent the small risk implies a value of statistical life of $5 million. Similarly, at the extreme end, a zero response implies a value-of-statistical-life estimate of zero. Table 20.1 summarizes the relationship between various willingness-to-pay amounts and the value of life.

When presented with this survey task, most students tend to give fairly low responses, at the lower end of the range of the table. When we examine the implicit values of statistical life of workers based on the wages they receive for the risks they face in their jobs, we will show that their values of statistical life are much greater than those often given by students responding to the one in 10,000 death-risk question. The median estimate of the value of statistical life for a worker in the United States is about $7 million.

Three explanations come to mind for the low responses students often give to these risk reduction questions. First, dealing with low-probability events such as this one is a very difficult undertaking. Second, even if the risk probability is understood, the threat may not be credible. It is, after all, only a hypothetical situation. Third, there is a tendency to think in terms of one’s immediate resources rather than one’s lifetime resources when answering this question. The current budget of a typical college student is substantially below that of an average blue-collar worker, but the student’s ultimate lifetime earnings will be greater.

### Table 20.1
Relation of Survey Responses to Value of Statistical Life

<table>
<thead>
<tr>
<th>Amount Will Pay (dollars) to Eliminate 1/10,000 Risk</th>
<th>Value of Statistical Life (dollars)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infinite</td>
<td>Infinity</td>
</tr>
<tr>
<td>Above 1,000</td>
<td>At least 10,000,000</td>
</tr>
<tr>
<td>500–1,000</td>
<td>5,000,000–10,000,000</td>
</tr>
<tr>
<td>200–500</td>
<td>2,000,000–5,000,000</td>
</tr>
<tr>
<td>50–200</td>
<td>500,000–2,000,000</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Willingness-to-Pay versus Other Approaches

The procedure used to value risks to life, health, and environmental outcomes more generally is exactly the same as is used in other contexts in which we assess the benefits of a government program. In particular, the benefit value is simply society’s willingness to pay for the impact of the program. This outcome may be in the form of a lottery, as in the case where the probability of an adverse event is reduced through a beneficial risk-regulation effort.

---

2 This principle is the same as in all benefit contexts. See Edith Stokey and Richard J. Zeckhauser, *A Primer for Policy Analysis* (New York: W. W. Norton, 1978).
Although reliance on the willingness-to-pay approach may seem to gain us little in terms of enabling us to assess benefit values in practice, it does offer a considerable advantage in terms of preventing one from adopting a benefit assessment procedure that is not economically sound.

The economic pitfalls that may be encountered are apparent from considering some of the alternative approaches to valuing life that have been suggested. For the most part, these approaches rely on various human capital measures related to one’s lifetime earnings. However, the kind of approach that is useful in assessing the value of training or education may be wholly inappropriate for assessing the implications of life-extending efforts. The first human capital measure one can consider is the present value of one’s lifetime earnings. This might be taken as a good gross measure of one’s value to the GNP, and it is an easy number to calculate. The fallacy of using this measure is apparent in part from the fact that the elderly and people who choose to work outside of the labor force would fare particularly badly under such a procedure. In addition, although one’s income level is clearly going to influence one’s willingness to pay for risk reduction, it need not constrain it in a one-to-one manner. Thus, when dealing with a small risk of death, such as one chance in 10,000, one is not necessarily restricted to being willing to spend only 1/10,000 of one’s income to purchase a risk reduction. One could easily spend 5/10,000 or more for small incremental reductions in risk. Difficulties arising from budgetary constraints are encountered only when we are dealing with dramatic risk increments. Moreover, if one were faced with a substantial risk of death, one might choose to undertake unusual efforts, such as working overtime or moonlighting on a second job, if one’s survival depended on it.

A variant on the present-value-of-earnings approach is to take the present value of lifetime earnings net of the consumption of the deceased. This is a common measure used in court cases for compensating survivors in wrongful death cases, inasmuch as it is a reflection of the net economic loss to the survivors after the death of a family member. This type of calculation abstracts from the consumption expenditures of the individual who is deceased, and it is certainly the individual whose health is most affected who should figure prominently in any calculation of the benefits of pursuing any particular social regulation.

A final approach that has appeared in the literature is to look at the taxes that people might pay. Focusing on tax rates captures the net financial contribution one makes to society, but it has the drawback of neglecting the income contribution to oneself or one’s family.

Notwithstanding the inappropriateness of the various earnings approaches, this technique not only has appeared in the literature but also has been widely used by government agencies. Much of the appeal of the method is that it lends itself to calculation.

A major policy event that led to a shift in the approach taken was the OSHA hazard-communication regulation that was the subject of intense debate in the early 1980s.\(^4\) OSHA prepared its regulatory analysis, assessing the value of the risk reduction achieved by valuing these impacts according to the lost earnings of the individuals whose death or nonfatal cases of cancer could be prevented. OSHA justified this approach on the basis that it was much too sensitive an issue to value life, so that it would follow the alternative approach of simply assessing the costs of death.

Because of OSHA’s overoptimistic risk assessment assumptions, the U.S. Office of Management and Budget (OMB) rejected the regulatory proposal. OSHA appealed this decision to then-Vice President George Bush, who had delegated authority over regulatory matters. Vice President Bush viewed the controversy as hinging on technical economic issues, and W. Kip Viscusi was asked to resolve the controversy. The failure of the OSHA analysis to show that the benefits exceeded the costs could be traced to valuing lives based on the cost of death rather than on the basis of society’s willingness to pay to reduce fatality risks. After making this change, the estimated benefits of the proposed regulation far exceeded the costs. The day after this economic analysis reached the Reagan White House, OSHA was permitted to issue the regulation. Because willingness-to-pay amounts generally exceed the present value of lost earnings by roughly an order of magnitude, using an appropriate economic methodology greatly enhances the attractiveness of social regulation efforts and makes these regulations appear more attractive than they would otherwise be. Indeed, the substantial size of the benefit estimates that can be achieved using the willingness-to-pay measure, rather than its economic soundness, may be the principal contributor to the widespread adoption of this approach throughout the federal government.

There also appears to be less reluctance to address the life-saving issues directly. Two to three decades ago, raising the issue of the value of life appeared to be intrinsically immoral. However, once it is understood that what is at issue is the amount of resources one is willing to commit to small reductions of risk, rather than to prevent a certain death, then the approach becomes less controversial. Moreover, because the measure is simply the total willingness of society to pay for the risk reductions, it does not use economic pricing in any crass or illegitimate way, as would be the case with the various human capital measures noted earlier. Society has also become aware of the wide range of risks that we face, including those imposed by our diets and a variety of personal activities. The idea that it is not feasible to achieve an absolutely risk-free existence and that some trade-offs must ultimately be made is becoming more widely understood.

\(^4\) The debate over the hazard communication regulation and over the value of life itself was the object of a cover story in The Washington Post Magazine, June 9, 1985, pp. 10–13, 36–41.
Variations in the Value of Statistical Life

One dividend of going through the exercise summarized in table 20.1 is that individuals will give different answers to these willingness-to-pay questions. There is no right answer in terms of the value of life. Thus we are not undertaking an elusive search for a natural constant such as $e$ or $\pi$. Rather, the effort is simply one to establish an individual’s risk-dollar trade-off. Individuals can differ in terms of this trade-off just as they could with respect to other kinds of trade-offs they might make concerning various kinds of consumption commodities that they might purchase. It makes no more sense to claim that individuals should have the same value of life than it does to insist that everyone should like eating raw oysters.

A major source of differences in preferences is likely to be individuals’ lifetime wealth. People who are more affluent are likely to require a higher price to bear any particular risk. This relationship is exhibited in the substantial positive income elasticity in the demand for medical insurance, as well as in a positive relationship between individual income and the wage compensation needed to accept a hazardous job. The amount workers are willing to pay to avoid a given injury risk increases roughly proportionally with worker income, which is consistent with this pattern of influences. A recent review of sixty studies of the value of statistical life throughout the world found that there is an income elasticity of the value of statistical life of about 0.5 to 0.6.

Overall, there is likely to be substantial heterogeneity in individual preferences, and this heterogeneity will be exhibited in the choices that people make. Empirical evidence suggests that smokers are more willing to bear a variety of risks other than smoking in return for less compensation than would be required for a nonsmoker. Individuals who wear seat belts are particularly reluctant to incur job risks, a finding that one would also expect. Indeed, the estimates by Joni Hersch and W. Kip Viscusi indicate that, while the average worker values a typical lost-workday injury at $72,000, smokers value such an injury at $39,000, and workers who use seat belts value such an injury at $118,000. Thus, there is considerable heterogeneity in individual risk-money trade-offs. If one examined a distribution of job-related risks, such as that provided in table 20.2, one would expect that the individuals who are in the relatively safe occupations listed in the table would generally be more averse to risk than those


in the riskiest pursuits. In contrast, people who tend to gravitate to the high-risk jobs, who choose to skydive, or who smoke cigarettes are more likely to place a lower value on incurring such risks than do those who avoid such pursuits.

Although substantial differences such as those exist, from a policy standpoint the extent to which we would use such distinctions is not quite clear. Should we provide individuals with less stringent government regulations to protect them if they have revealed by other activities that they are willing to bear a variety of risks to their well-being? Viewed somewhat differently, should we override the decisions of people who find a particular wage-risk trade-off in the labor market attractive or who find the nuisance of wearing a seat belt to outweigh the perceived benefits to themselves? Although one should generally respect individual values in a democratic society, we may wish to distinguish situations in which individuals are believed to be irrational or where it is not feasible to educate people inexpensively with respect to the rational course of action. One danger of regulation of this type, however, is that we may impose the preferences of policymakers on the individuals whose well-being is supposed to be protected, an approach that may not necessarily be welfare-enhancing for those affected by the regulation.

The one area in which the differences in the value of statistical life should clearly be utilized is in assessing future impacts of regulatory programs. Because further benefits are deferred, discounting these benefits to bring them to present value reduces the current value of regulatory policies with long-run effects such as pollution control to reduce the depletion of the ozone layer around the atmosphere. If, however, we recognize that future generations are likely to be wealthier, then much of the role of discounting will be muted. Consider, for example, the situation in which the income elasticity of the value of the benefits is 1.0. Let the benefit \( n \) years hence be \( B \), the growth rate in income between now and the time when
the benefits are realized be \( g \), and the interest rate be \( r \). The following equation gives the present value of the benefits, which simply equal the dollar benefit value \( B \) multiplied by a growth factor, which is the spread between the growth rate in income minus the interest rate:

\[
\text{Present value of benefit} = \frac{B(1 + g)^n}{(1 + r)^n} = B(1 + g - r)^n. \tag{20.4}
\]

Thus the growth in income will mute to a large extent the influence of discounting when weighing the consequences of policies in the future.

One might raise the question whether one should discount at all or simply treat all policy outcomes in different years equally, irrespective of the time in which they transpire. This procedure has been advocated by the U.S. EPA because doing so will greatly enhance the attractiveness of its efforts, many of which have deferred effects. The fallacy of ignoring discounting altogether is apparent when one considers that in the absence of discounting, one would never take an action in which there will be a permanent adverse effect of any kind. The costs of such efforts will always be infinite, and such policies would never be pursued.

---

**The Labor Market Model**

Most of the empirical estimates of the value of life have been based on labor market data. The general procedure is to estimate the wage-risk trade-off that workers implicitly make as part of their jobs and to use implications of this trade-off as an estimate of the value of life.

As the starting point for the analysis, consider figure 20.1. Sketched in this diagram are two curves, \( EU_1 \) and \( EU_2 \), which are constant expected utility loci for the worker. This combination of wages and risk on each curve gives the worker the same expected utility. The required wage rate is an increasing function of the risk, which is true for a wide range of individual preferences. All that is required is that one would rather be healthy than not. It is not necessary that one be risk-averse in the sense of unwilling to accept actuarially unfair financial bets. Two expected utility loci offering constant expected utility are \( EU_1 \) and \( EU_2 \). Higher wage rates and lower risk levels are preferred, so that the direction of preference is toward the northwest.

Workers do not have all wage-risk combinations to choose from, but instead are limited to those that are offered by firms. Figure 20.2 illustrates how the available set of job opportunities is constructed. Each particular firm has a constant expected profits locus. Thus, one firm will have a locus \( MM \), where this isoprofit curve gives the locus of wage-risk combinations that give the firm the same level of profits. For example, if a firm lowers the risk level by investing in additional health and safety equipment, to maintain the same level of profits, the wage rate must go down. As a result, the wage that the firm can offer...
and maintain the same level of profits will be an increasing function of risk. The curvature of the MM isoprofit curve is dictated by the fact that additional safety reductions become increasingly difficult to achieve, so that as one moves to the left along the risk axis, the additional cost expenditures on the part of the firm become increasingly great. Consequently, the magnitude of the wage increase required for any particular risk reduction becomes greater. Curve NN is another example of an isoprofit curve for a different firm in the industry.

The outer envelope of the isoprofit curves for the entire industry provides the offer curve available to workers. Thus a worker’s task is to select the point along the offer curve VV that gives the worker the highest level of expected utility. Points below this curve will be domi-
Valuing Life and Other Nonmonetary Benefits

nated by points along it, since a point below $VV$ will be less desirable than a job that offers the same risk at a higher wage rate.

The nature of market equilibrium is illustrated in figure 20.3. Worker 1 achieves his constant expected utility at the point of tangency with the market opportunity locus $VV$, where his tangency point is at $X$. In contrast, worker 2 selects a higher risk-wage combination at point $Y$. Because of the aforementioned heterogeneity in individual tastes, the individuals

will generally sort themselves along the part of the wage offer curve that best suits their preferences.

How one interprets differences in the estimated value of statistical life depends on people’s opportunities, not just their preferences. Black workers in the United States, for example, have estimates of the value of statistical life of about half to two-thirds that of whites, depending on the particular analysis and sample. However, this difference in risk-wage trade-offs is not attributable solely to a greater willingness to bear risk. In particular, the market offer

---


curve $VV$ facing black workers is lower and flatter than that facing whites, reflecting their quite different array of labor market opportunities. Estimates of the value of statistical life in the labor market consequently reflect the economic forces of both supply and demand, which together will jointly determine the observed trade-off rate.

The task of empirical analysis in this area is to analyze the nature of the observed market equilibrium points reflected in data sets on worker behavior. Thus, if we observe points $X$ and $Y$, the estimation of a linear relationship between wages and risk would yield the curve $AA$ shown in figure 20.3. The slope of $AA$ gives the estimated wage-risk trade-off. In effect, what this curve does is indicate the terms of trade that workers, on average, are willing to accept between risk and wages. These terms of trade in turn can be used to extrapolate the implicit value that workers attach to a statistical death.

The details of the methodology vary depending on the particular data set used for the estimation. In general, the statistical approach involves the use of a large set of data on individual employment behavior. Table 20.3 summarizes the international evidence on the value of statistical life. The thirty studies from the U.S. labor market have a median value of $7 million. These studies have served as the primary basis for the value of statistical life used by U.S. regulatory agencies. Table 20.3 also presents similar estimates that have been generated for ten other countries. While the methodologies used in these different studies often differ, the results are consistent with expectations, given the income level differences.

Table 20.3
Labor Market Estimates of Value of Statistical Life Throughout the World

<table>
<thead>
<tr>
<th>Study/Country</th>
<th>Value of Statistical Life ($ millions)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median value from 30 U.S. studies</td>
<td>7.0</td>
</tr>
<tr>
<td>Australia</td>
<td>4.2</td>
</tr>
<tr>
<td>Austria</td>
<td>3.9–6.5</td>
</tr>
<tr>
<td>Canada</td>
<td>3.9–4.7</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>1.7</td>
</tr>
<tr>
<td>India</td>
<td>1.2–1.5</td>
</tr>
<tr>
<td>Japan</td>
<td>9.7</td>
</tr>
<tr>
<td>South Korea</td>
<td>0.8</td>
</tr>
<tr>
<td>Switzerland</td>
<td>6.3–8.6</td>
</tr>
<tr>
<td>Taiwan</td>
<td>0.2–0.9</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>4.2</td>
</tr>
</tbody>
</table>


The lowest estimates are for countries such as Taiwan, South Korea, and India. Japan and Switzerland have higher estimates of the value of statistical life.

While serving as chief economist for the World Bank, Larry Summers wrote a memo hypothesizing that poorer countries should be more willing to accept risks, such as being the site for hazardous waste storage. The memo generated substantial controversy. However, his economic assumption that the value of statistical life does vary across countries in the expected direction is borne out by the data in table 20.3. Because these data reflect implicit values of life based on market decisions, they imply that within these poorer countries, people are already striking a different balance between safety and money than in more affluent nations.

**Empirical Estimates of the Value of Life**

The general form of the estimation depends in part on the nature of the wage and risk information that is available, such as whether the data pertain to annual earnings or hourly wage rates. One form of estimating the equation is the following:

\[
\text{Annual earnings} = \alpha + \beta_1 \text{ Annual death risk} + \sum_{i=1}^{n} \gamma_i \text{ Personal characteristic}_i + \sum_{i=1}^{m} \psi_i \text{ Job characteristic}_i + \epsilon. \tag{20.5}
\]

The dependent variable in this analysis is the annual worker earnings, which is not as accurate a measure as the worker’s hourly wage rate, but for expositional purposes it facilitates our task of indicating how one constructs the value-of-statistical-life estimates in the equation. The explanatory variables include the annual death risk facing the worker. In general, this information is matched to the workers in the sample based on their responses regarding their industry or occupation.

The coefficient \( \beta_1 \) in equation 20.5 indicates how annual earnings will be affected by an increase in the annual death risk. If the annual death risk were 1.0, then \( \beta_1 \) would give the change in annual earnings required to face one expected death. Thus, for the equation as it has been set up here, \( \beta_1 \) is the value-of-statistical-life estimate. In particular, it represents the trade-off that workers exhibit between earnings and the risk of death.

Studies in the United States have used a variety of different fatality-risk data sets that provide risk estimates that can be matched to workers based on their industry or occupation. While some studies have used life insurance data or workers’ compensation rate data, the most reliable estimates are those based on fatality rate estimates using data from the U.S. Bureau of Labor Statistics and the National Institute of Occupational Safety and Health. Each
of these agencies attempts to generate a comprehensive census of all job-related fatalities rather than extrapolating from a smaller sample of fatalities.

The other variables included in equation 20.5 are designed to control for the other aspects of the worker and his job that will influence earnings. In general, the people who earn the highest incomes in our society also have fairly low-risk jobs. This observation, which can be traced back to the time of John Stuart Mill, reflects the positive income elasticity of the demand for health. By including a detailed set of other variables, including coverage of factors such as worker education and union status, one can successfully disentangle the premium for job risks as opposed to compensation for other attributes of the worker and his job.

Most U.S. value-of-statistical-life estimates using labor market data range from under $3.8 million to $9.0 million. This heterogeneity is not solely a consequence of the imprecision of the statistical measures but instead is due to the fact that these studies are measuring different things. The value-of-life estimates for samples of different riskiness are expected to be different because the mix of workers and their preferences across samples may be quite different. In addition, the degree to which different risk variables measure the true risk associated with the job may differ substantially across risk measures.

Even with the current state of econometric techniques and the substantial literature devoted to this issue, economists cannot yet pinpoint the value of statistical life that is appropriate in every particular instance. However, we have a good idea of the general range in which such values fall, and from the standpoint of making policy judgments with respect to the ballpark in which our policies should lie, this guidance should be sufficient.

---

### Value of Risks to Life for Regulatory Policies

For the most part, regulatory agencies have used estimates drawn from the value-of-statistical-life literature to value the benefits of regulations that reduce risks to life. Table 20.4 summarizes these unit benefit values from a wide range of regulatory impact analyses. Most of these estimates are in the $5 million to $6 million range reflected by larger labor market studies of the value of statistical life. The main exceptions are agencies in the U.S. Department of Transportation, for which the benefit values have been from $1 million to $3 million per statistical life. That agency has used benefit values on the low end of the labor market estimates, perhaps because historically that agency had used the compensation levels from auto accident cases to value lives, thus anchoring their analysis at a low number.

It is useful to examine the government policies that have actually been pursued in the social regulation area to see the extent to which they conform with an appropriate value of statistical life. While agencies diligently value life following norms, the amounts that are actually spent to reduce risks to life are often quite different and may bear little relationship to these
Table 20.4

<table>
<thead>
<tr>
<th>Year</th>
<th>Agency</th>
<th>Regulation</th>
<th>Value of a Statistical Life ($ millions, 2000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1985</td>
<td>Federal Aviation Administration</td>
<td>Protective Breathing Equipment (50 Federal Register 41452)</td>
<td>1.0*</td>
</tr>
<tr>
<td>1985</td>
<td>Environmental Protection Agency</td>
<td>Regulation of Fuels and Fuel Additives; Gasoline Lead Content (50 FR 9400)</td>
<td>1.7</td>
</tr>
<tr>
<td>1988</td>
<td>Federal Aviation Administration</td>
<td>Improved Survival Equipment for Inadvertent Water Landings (53 FR 24890)</td>
<td>1.5*</td>
</tr>
<tr>
<td>1988</td>
<td>Environmental Protection Agency</td>
<td>Protection of Stratospheric Ozone (53 FR 30566)</td>
<td>4.8</td>
</tr>
<tr>
<td>1990</td>
<td>Federal Aviation Administration</td>
<td>Proposed Establishment of the Harlingen Airport Radar Service Area, TX (55 FR 32064)</td>
<td>2.0*</td>
</tr>
<tr>
<td>1994</td>
<td>Food and Nutrition Service (USDA)</td>
<td>National School Lunch Program and School Breakfast Program (59 FR 30218)</td>
<td>1.7, 3.5*</td>
</tr>
<tr>
<td>1995</td>
<td>Consumer Product Safety Commission</td>
<td>Multiple Tube Mine and Shell Fireworks Devices (60 FR 34922)</td>
<td>5.6*</td>
</tr>
<tr>
<td>1996</td>
<td>Food Safety Inspection Service (USDA)</td>
<td>Pathogen Reduction; Hazard Analysis and Critical Control Point Systems (61 FR 38806)</td>
<td>1.9</td>
</tr>
<tr>
<td>1996</td>
<td>Food and Drug Administration</td>
<td>Regulations Restricting the Sale and Distribution of Cigarettes and Smokeless Tobacco to Protect Children and Adolescents (61 FR 44396)</td>
<td>2.7*</td>
</tr>
<tr>
<td>1996</td>
<td>Federal Aviation Administration</td>
<td>Aircraft Flight Simulator Use in Pilot Training, Testing, and Checking and at Training Centers (61 FR 34508)</td>
<td>3.0*</td>
</tr>
<tr>
<td>1996</td>
<td>Environmental Protection Agency</td>
<td>Requirements for Lead-Based Paint Activities in Target Housing and Child-Occupied Facilities (61 FR 45778)</td>
<td>6.3</td>
</tr>
<tr>
<td>1996</td>
<td>Food and Drug Administration</td>
<td>Medical Devices; Current Good Manufacturing Practice Final Rule; Quality System Regulation (61 FR 52602)</td>
<td>5.5*</td>
</tr>
<tr>
<td>1997</td>
<td>Environmental Protection Agency</td>
<td>National Ambient Air Quality Standards for Ozone (62 FR 38856)</td>
<td>6.3</td>
</tr>
<tr>
<td>1999</td>
<td>Environmental Protection Agency</td>
<td>Radon in Drinking Water Health Risk Reduction and Cost Analysis (64 FR 9560)</td>
<td>6.3</td>
</tr>
<tr>
<td>1999</td>
<td>Environmental Protection Agency</td>
<td>Control of Air Pollution from New Motor Vehicles: Tier 2 Motor Vehicle Emissions Standards and Gasoline Sulfur Control Requirements (65 FR 6698)</td>
<td>3.9, 6.3</td>
</tr>
<tr>
<td>2000</td>
<td>Consumer Product Safety Commission</td>
<td>Portable Bed Rails; Advance Notice of Proposed Rulemaking (65 FR 58968)</td>
<td>5.0*</td>
</tr>
</tbody>
</table>

* The published summaries of the regulatory impact analyses for these rules do not specify the year in which the reported dollars are denominated. We have assumed that the dollar year corresponds to the date of rule publication for purposes of converting all values into 2000 dollars. Note that the CPSC reported a value of statistical life of $5 million in both its 1995 and 2000 regulations; the difference in values reflects our deflating to 2000 dollars.

benefit values. Table 20.5 summarizes a variety of key aspects of major regulations. These regulations covered such diverse issues as cabin fire protection for airplanes, grain dust regulations for grain handling facilities, and environmental controls for arsenic/copper smelters.

There are three columns of data of interest in table 20.5. The first column of figures displays the cost per life saved by each of the programs. Some of these efforts, such as steering column protection for automobiles and other entries at the top of the table, are bargains. Their cost per expected life saved is well below $1 million. For concreteness, suppose that we took as an appropriate value of life a figure of $7.0 million. Then all regulations at the top part of the table, including the EPA’s ethylene dibromide in drinking water regulation, would pass a benefit-cost test. Similarly, all regulations at the bottom part of the table could not be justified on benefit-cost grounds.

The next column of statistics gives the cost per normalized life saved, where all lives are converted into accident equivalents based on the discounted number of life-years saved. Acute accidents save lives with much more substantial duration than do anticancer policies. The consequences of these adjustments are particularly great for the health-oriented regulations, which already tended to have low cost-effectiveness. Once one adjusts for the duration of life lost, all regulations beginning with the radionuclide regulation for uranium mines no longer pass a benefit-cost test. The effect of such adjustments is substantial. The EPA asbestos ban cost $132 million per life saved, but imposed a cost of $329 million per normalized life saved. Whether one should adjust for age and how one should adjust represents an economic controversy that will be examined in chapter 21 with respect to EPA policies.

The final column of statistics in table 20.5 gives the cost per life-year saved. The FDA is notable among regulatory agencies in focusing on life-years when assessing policies. Some of these estimated values are considerable, as the OSHA asbestos regulations in 1986 cost close to $11 million per year of life saved, which is well in excess of a reasonable value for preventing the loss of a lifetime of such years for a typical worker.

The wide distribution of different agencies in terms of the efficacy of their life-saving efforts is noteworthy. All regulations from the U.S. Department of Transportation—the National Highway Safety Administration (NHTSA) and the Federal Aviation Administration (FAA)—pass a benefit-cost test. Indeed, that department is exceptional in that it will not pursue any regulation that does not meet such a test of efficacy. In contrast, virtually every regulation by the EPA and OSHA fails a benefit-cost test because of the restrictive nature of their legislative mandates.

An additional message of table 20.5 is that in general, it is not necessary to pinpoint the exact value of statistical life that is appropriate for any government policy. For the most part,
### Table 20.5
The Costs of Various Risk-Reducing Regulations per Expected Life Saved

<table>
<thead>
<tr>
<th>Regulation</th>
<th>Year</th>
<th>Agency</th>
<th>Cost per Life Saved (millions of 1995 dollars)</th>
<th>Cost per Normalized Life Saved (millions of 1995 dollars)</th>
<th>Cost per Year of Life Saved (millions of 1995 dollars)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unvented space heater ban</td>
<td>1980</td>
<td>CPSC</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td>Aircraft cabin fire protection standard</td>
<td>1985</td>
<td>FAA</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td>Seat belt/air bag</td>
<td>1984</td>
<td>NHTSA</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td>Steering column protection standards</td>
<td>1987</td>
<td>NHTSA</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td>Underground construction standards</td>
<td>1989</td>
<td>OSHA</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td>Trihalomethane in drinking water</td>
<td>1979</td>
<td>EPA</td>
<td>0.2</td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td>Aircraft seat cushion flammability</td>
<td>1984</td>
<td>FAA</td>
<td>0.5</td>
<td>0.6</td>
<td>0.0</td>
</tr>
<tr>
<td>Alcohol and drug controls</td>
<td>1985</td>
<td>FRA</td>
<td>0.5</td>
<td>0.6</td>
<td>0.0</td>
</tr>
<tr>
<td>Auto fuel system integrity</td>
<td>1975</td>
<td>NHTSA</td>
<td>0.5</td>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>Auto wheel rim servicing</td>
<td>1984</td>
<td>OSHA</td>
<td>0.5</td>
<td>0.6</td>
<td>0.0</td>
</tr>
<tr>
<td>Aircraft floor emergency lighting</td>
<td>1984</td>
<td>FAA</td>
<td>0.7</td>
<td>0.9</td>
<td>0.0</td>
</tr>
<tr>
<td>Concrete and masonry construction</td>
<td>1988</td>
<td>OSHA</td>
<td>0.7</td>
<td>0.9</td>
<td>0.0</td>
</tr>
<tr>
<td>Crane-suspended personnel platform</td>
<td>1988</td>
<td>OSHA</td>
<td>0.8</td>
<td>1.0</td>
<td>0.1</td>
</tr>
<tr>
<td>Passive restraints for trucks and buses</td>
<td>1989</td>
<td>NHTSA</td>
<td>0.8</td>
<td>0.8</td>
<td>0.0</td>
</tr>
<tr>
<td>Auto side impact standards</td>
<td>1990</td>
<td>NHTSA</td>
<td>1.0</td>
<td>1.0</td>
<td>0.1</td>
</tr>
<tr>
<td>Children's sleepwear flammability ban</td>
<td>1973</td>
<td>CPSC</td>
<td>1.0</td>
<td>1.2</td>
<td>0.1</td>
</tr>
<tr>
<td>Auto side door supports</td>
<td>1970</td>
<td>NHTSA</td>
<td>1.0</td>
<td>1.0</td>
<td>0.1</td>
</tr>
<tr>
<td>Low-altitude wind shear equipment and training</td>
<td>1988</td>
<td>FAA</td>
<td>1.6</td>
<td>1.9</td>
<td>0.1</td>
</tr>
<tr>
<td>Metal mine electrical equipment standards</td>
<td>1970</td>
<td>MSHA</td>
<td>1.7</td>
<td>2.0</td>
<td>0.1</td>
</tr>
<tr>
<td>Trenching and excavation standards</td>
<td>1989</td>
<td>OSHA</td>
<td>1.8</td>
<td>2.2</td>
<td>0.1</td>
</tr>
<tr>
<td>Traffic alert and collision avoidance systems</td>
<td>1988</td>
<td>FAA</td>
<td>1.8</td>
<td>2.2</td>
<td>0.1</td>
</tr>
<tr>
<td>Hazard communication standard</td>
<td>1983</td>
<td>OSHA</td>
<td>1.9</td>
<td>4.8</td>
<td>0.2</td>
</tr>
<tr>
<td>Trucks, buses, and MPV side impact</td>
<td>1989</td>
<td>NHTSA</td>
<td>2.6</td>
<td>2.6</td>
<td>0.1</td>
</tr>
<tr>
<td>Grain dust explosion prevention standards</td>
<td>1987</td>
<td>OSHA</td>
<td>3.3</td>
<td>4.0</td>
<td>0.2</td>
</tr>
<tr>
<td>Risk Factor</td>
<td>Year</td>
<td>Source</td>
<td>Lower Bound</td>
<td>Upper Bound</td>
<td>Estimated Mortality Risk</td>
</tr>
<tr>
<td>----------------------------------------------------------------------------</td>
<td>------</td>
<td>--------------</td>
<td>-------------</td>
<td>-------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>Rear lap/shoulder belts for autos</td>
<td>1989</td>
<td>NHTSA</td>
<td>3.8</td>
<td>3.8</td>
<td>0.2</td>
</tr>
<tr>
<td>Standards for radionuclides in uranium mines</td>
<td>1984</td>
<td>EPA</td>
<td>4.1</td>
<td>10.1</td>
<td>0.5</td>
</tr>
<tr>
<td>Benzene NESHAP (original: fugitive emissions)</td>
<td>1984</td>
<td>EPA</td>
<td>4.1</td>
<td>10.1</td>
<td>0.5</td>
</tr>
<tr>
<td>Ethylene dibromide in drinking water</td>
<td>1991</td>
<td>EPA</td>
<td>6.8</td>
<td>17.0</td>
<td>0.8</td>
</tr>
<tr>
<td>Benzene NESHAP (revised: coke by-products)</td>
<td>1988</td>
<td>EPA</td>
<td>7.3</td>
<td>18.1</td>
<td>0.9</td>
</tr>
<tr>
<td>Asbestos occupational exposure limit</td>
<td>1972</td>
<td>OSHA</td>
<td>9.9</td>
<td>24.7</td>
<td>1.2</td>
</tr>
<tr>
<td>Benzene occupational exposure limit</td>
<td>1987</td>
<td>OSHA</td>
<td>10.6</td>
<td>26.5</td>
<td>1.3</td>
</tr>
<tr>
<td>Electrical equipment in coal mines</td>
<td>1970</td>
<td>MSHA</td>
<td>11.0</td>
<td>13.3</td>
<td>0.6</td>
</tr>
<tr>
<td>Arsenic emission standards for glass plants</td>
<td>1986</td>
<td>EPA</td>
<td>16.1</td>
<td>40.2</td>
<td>1.9</td>
</tr>
<tr>
<td>Ethylene oxide occupational exposure limit</td>
<td>1984</td>
<td>OSHA</td>
<td>24.4</td>
<td>61.0</td>
<td>2.9</td>
</tr>
<tr>
<td>Arsenic/copper NESHAP</td>
<td>1986</td>
<td>EPA</td>
<td>27.4</td>
<td>68.4</td>
<td>3.3</td>
</tr>
<tr>
<td>Hazardous waste listing of petroleum refining sludge</td>
<td>1990</td>
<td>OSHA</td>
<td>32.9</td>
<td>82.1</td>
<td>3.9</td>
</tr>
<tr>
<td>Cover/move uranium mill tailings (inactive)</td>
<td>1983</td>
<td>EPA</td>
<td>37.7</td>
<td>94.3</td>
<td>4.5</td>
</tr>
<tr>
<td>Benzene NESHAP (revised: transfer operations)</td>
<td>1990</td>
<td>EPA</td>
<td>39.2</td>
<td>97.9</td>
<td>4.7</td>
</tr>
<tr>
<td>Cover/move uranium mill tailings (active sites)</td>
<td>1983</td>
<td>EPA</td>
<td>53.6</td>
<td>133.8</td>
<td>6.4</td>
</tr>
<tr>
<td>Acrylonitrile occupational exposure limit</td>
<td>1978</td>
<td>OSHA</td>
<td>61.3</td>
<td>153.2</td>
<td>7.3</td>
</tr>
<tr>
<td>Coke ovens occupational exposure limit</td>
<td>1976</td>
<td>OSHA</td>
<td>75.6</td>
<td>188.9</td>
<td>9.1</td>
</tr>
<tr>
<td>Lockout/tagout</td>
<td>1989</td>
<td>OSHA</td>
<td>84.4</td>
<td>102.4</td>
<td>4.9</td>
</tr>
<tr>
<td>Asbestos occupational exposure limit</td>
<td>1986</td>
<td>OSHA</td>
<td>88.1</td>
<td>220.1</td>
<td>10.6</td>
</tr>
<tr>
<td>Arsenic occupational exposure limit</td>
<td>1978</td>
<td>OSHA</td>
<td>127.3</td>
<td>317.9</td>
<td>15.2</td>
</tr>
<tr>
<td>Asbestos ban</td>
<td>1989</td>
<td>EPA</td>
<td>131.8</td>
<td>329.2</td>
<td>15.8</td>
</tr>
<tr>
<td>Diethylstilbestrol (DES) cattle-feed ban</td>
<td>1979</td>
<td>FDA</td>
<td>148.6</td>
<td>371.2</td>
<td>17.8</td>
</tr>
<tr>
<td>Benzene NESHAP (revised: waste operations)</td>
<td>1990</td>
<td>EPA</td>
<td>200.2</td>
<td>500.2</td>
<td>24.0</td>
</tr>
<tr>
<td>1,2-dichloropropane in drinking water</td>
<td>1991</td>
<td>EPA</td>
<td>777.4</td>
<td>1,942.1</td>
<td>93.1</td>
</tr>
<tr>
<td>Hazardous waste land disposal ban</td>
<td>1988</td>
<td>EPA</td>
<td>4,988.7</td>
<td>12,462.7</td>
<td>597.4</td>
</tr>
<tr>
<td>Municipal solid waste landfills</td>
<td>1988</td>
<td>EPA</td>
<td>22,746.8</td>
<td>56,826.1</td>
<td>2,724.2</td>
</tr>
<tr>
<td>Formaldehyde occupational exposure limit</td>
<td>1987</td>
<td>OSHA</td>
<td>102,622.8</td>
<td>256,372.7</td>
<td>12,290.2</td>
</tr>
<tr>
<td>Atrazine/alachlor in drinking water</td>
<td>1991</td>
<td>EPA</td>
<td>109,608.5</td>
<td>273,824.4</td>
<td>13,126.8</td>
</tr>
<tr>
<td>Hazardous waste listing for wood-preserving chemicals</td>
<td>1990</td>
<td>EPA</td>
<td>6,785,822.0</td>
<td>16,952,364.9</td>
<td>812,673.3</td>
</tr>
</tbody>
</table>

rough judgments regarding the efficacy of a regulation can tell us a great deal. We know, for example, if OSHA arsenic regulations save lives at a cost of $127.3 million per life, that such efforts are out of line with what the beneficiaries of such an effort believe the value of such a regulation to be. Moreover, there is likely to be a wide range of other regulatory alternatives by OSHA or other agencies that are likely to be more cost-effective ways of saving lives. Unfortunately, OMB has limited influence, as it has never rejected a regulation with a cost per life of saved of under $100 million.

Although the range in the value-of-statistical-life estimates for the policies summarized in table 20.5 may seem to be substantial, in practice many government policies are proposed but not issued, because the value of life is even higher than many of the outliers in this table. For example, in 1984, the EPA proposed regulations for benzene/maleic anhydride that would cost $820 million per life saved. This regulation was rejected by the OMB as being too expensive. Calculating the costs, benefits, and appropriate reference values for the value of life often highlights gross policy distortions such as this one.

Survey Approaches to Valuing Policy Effects

There are many circumstances in which we do not have readily available market data that can be used to estimate either implicit or explicit prices. How much, for example, is it worth to prevent genetic damage or to save an endangered species?

In the absence of existing data on these issues, an approach that has been used in the benefit-valuation literature for several decades has been to run a survey in which individuals are polled with respect to these values. This approach is now the dominant methodology for assessing environmental benefits because of the paucity of good data on explicit or implicit environmental transactions.

The actual procedures that have evolved for doing so in effect attempt to replicate the hedonic market estimate approach used to analyze wage-risk trade-offs and similar factors using survey data. For example, such studies would not ask people how much they valued a job injury but would instead ask how much wage compensation they would require to face extra risk. Similarly, assessment of an environmental amenity would focus on purchasing a reduction in certain risks in the environment rather than certain outcomes. The term *contingent valuation* has been used to describe such studies because they represent values that are contingent on a hypothetical market existing. 12 Thus they represent a hybrid between the initial survey approaches used in the literature and the market-based valuation econometric studies that began in the 1970s.

The objective is to elicit benefit values by constructing survey questions concerning hypothetical situations. There are a variety of ways in which one could pose the valuation question. In each case one must first give individuals information regarding the risk or other outcome to be valued. The first approach would be to ask individuals how much that particular benefit would be worth to them. This is a one-step procedure. The second approach would be an iterative one in which the individual first answered the open-ended question, and then was asked whether he or she would be willing to pay a small amount more than the initial response. A third variant on this technique is that instead of asking open-ended questions, individuals could be given a series of bids, and they would then have to determine how high or low they would go. These bids could be given in either ascending or descending order. In the ascending case, an individual might first be asked whether he or she would be willing to pay $1 for improved air quality, and if the answer is yes, the respondent would be asked if he or she would be willing to pay $2 for improved air quality, and so on, until the individual is not willing to increase the bid. A fourth approach is to utilize paired comparisons in which an individual is given an alternative product or other binary choices to make. Using interactive computer programs, one can then give an individual a succession of options to pick from to locate the point of indifference.

All of these variations in terms of the methodology are largely ones of process rather than economic content. The underlying issue is how we can best frame the survey questions to elicit the true underlying economic values that individuals have. In the case of market outcomes we know from revealed preference that these values will be expressed in individual decisions, but in the case of surveys the values that we elicit may be sensitive to the manner in which we attempt to determine individual preferences.

More generally, considerable care must be exercised in the design of such survey studies so that they will give us reliable results. Often such studies rely on “convenience samples” such as groups of students, but our ultimate objective is to ascertain the willingness to pay of actual beneficiaries of the project, not the willingness to pay of students in the class, whose responses may be biased in part by substantial demand effects (they may give the answers that they expect their professor wants to see). Perhaps the major guidelines in assessing these studies is to determine the extent to which they replicate market processes in a meaningful manner.

When interview studies first came into use in the literature, economists feared that there would be a major problem in individuals’ misrepresenting their true values for strategic reasons. Advocates of pollution control efforts, for example, might give responses that indicate enormous willingness-to-pay amounts, knowing that they will not be taxed on the basis of their response and hoping that a high response will tilt the policy in their favor.

In practice, the strategic issue has not been a major problem with the survey studies. A more fundamental difficulty is that some individuals often may not give thoughtful or meaningful responses to the question, inasmuch as it does not involve a decision that they
actually make. Moreover, because many of the decisions involve risks, some of which are at very low probabilities, the results will not reflect their underlying values but instead will be contaminated by whatever irrationalities influence one’s decisions involving low-probability events.

Valuation of Air Quality

The nature of the performance of the survey approach varies from study to study, but some suggestions as to its likely precision are given by a study of air pollution valuation. Two approaches were used to value air quality. In the first, a hedonic rent-gradient equation for the Los Angeles area was estimated, analyzing the relationship of home sale prices to a variety of factors likely to influence house price (such as house age, area, school quality, public safety, and distance to the beach). In addition, this equation included measures of pollution in the Los Angeles area, including either total suspended particulates or nitrous oxide concentration levels. The authors found substantial housing price effects of pollution; controlling for other aspects of the housing market, higher pollution levels lowered the price of the house.

A survey approach was also used to assess the amount the individuals would be willing to pay in terms of a higher utility bill to achieve cleaner air. The expressed willingness to pay for different levels of air quality was roughly one-third of the market-based estimates. These results suggest that at least in this case, overstatement of valuations in surveys may not be a problem, although this conclusion may not be true more generally. In addition, there may not be an exact correspondence between survey valuation estimates and market estimates. Comparisons that have been done for worker-wage equations have yielded results that are more comparable to those obtained with market data, but in the job risk case one is dealing with a risk that is currently traded in a market and that individuals may have already thought about in this context, increasing the accuracy of the survey responses.

Exploratory Nature of the Survey Approach

Overall, survey approaches to establishing the benefits of social regulation represent an important complement to analyses using market data. This methodology should still be regarded as exploratory, however. Moreover, there will never be any general conclusions regarding the accuracy of such studies, because accuracy will vary from study to study, depending on the extent to which a realistic market context was created and the degree to which the individuals running the survey motivated the survey participants to give thoughtful and honest answers.

Sensitivity Analysis and Cost-Effectiveness

In the usual situation it will not be feasible to place dollar values on all outcomes of interest. In such circumstances one could undertake cost-effectiveness analysis to analyze the cost per unit outcome achieved, and indices such as this may often be instructive.

In addition, if there are multiple outcomes that one would wish to value but cannot, one can perform a sensitivity analysis assigning different relative weights to them to convert all of them into a common cost-effectiveness index. Table 20.6 summarizes calculations of this type that formed the basis for resolving the debate over the OSHA hazard communication regulation. The three health outcomes involved are lost workday job injuries, disabling illnesses, and cases of cancer. Suppose that, based on past studies on the relative valuation of cancer, we know that lost-workday job injuries have \( \frac{1}{20} \) of the value of a case of cancer. In addition, suppose that the main uncertainty is with respect to the value of disabling illnesses, where our task is to assess how severe this outcome is compared with injuries and cancer. The calculations in this table explore two different sets of weights, one in which lost-workday injuries and disabling illnesses are given the same weight, and a second in which disabling illnesses are viewed as being five times more severe than lost workday cases.

The first row of table 20.6 gives the net discounted costs less benefits of other kinds from the project, which total $2.6 billion. The second row gives the discounted (at 5 percent) number of lost-workday injury equivalents prevented, where these lost-workday equivalents have been calculated using the two sets of weights indicated above. Finally, the third row of the table gives the net discounted cost per lost-workday equivalent prevented. These estimates are in the range of $10,000 to $30,000, which is in line with the general estimates of implicit values of nonfatal injuries that have been obtained in labor market studies.

### Table 20.6
Cost-Effectiveness Measures for Hazard Communication Standard

<table>
<thead>
<tr>
<th>Lost-Workday Equivalents</th>
<th>Weight—1, 1, 20* Cost-Effectiveness</th>
<th>Weights—1, 5, 20* Cost-Effectiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Net discounted costs less monetized benefits</td>
<td>( 2.632 \times 10^9 )</td>
<td>( 2.632 \times 10^9 )</td>
</tr>
<tr>
<td>Total lost-workday equivalents (discounted)</td>
<td>( 9.5 \times 10^4 )</td>
<td>( 24.7 \times 10^4 )</td>
</tr>
<tr>
<td>Net discounted cost/lost-workday equivalent</td>
<td>$27,900</td>
<td>$10,700</td>
</tr>
</tbody>
</table>

* These are the relative weights placed on lost-workday cases (always 1), disabling illnesses (1 or 5), and cancers (always 20) in constructing a measure of lost-workday equivalents.

The approach used here is to establish one class of outcomes as the unit of metric and to put the other outcomes in terms of them when calculating a cost-effectiveness index that can capture all of the diverse impacts of a particular effort. In this case the metric is that of lost-workday equivalents, but in other situations the metric may be death equivalents prevented or number of birth defects prevented.

**Risk-Risk Analysis**

In the absence of a benefit-cost analysis for risk or environmental regulations, agencies will not be constrained regarding the stringency of these efforts. Because of the restrictive legislative mandates that these agencies have that often require that they reduce risk irrespective of cost, the result is that many regulations that are promulgated generate considerable costs, sometimes as high as $100 million per statistical life saved or more. Other than wasting societal resources, is there any harm from such profligacy?

Two classes of costs can be identified, where these come under the general heading of risk-risk analysis. First, there is a direct risk-risk trade-off arising from regulatory efforts. An automobile recall, for example, may require that consumers drive their cars back to the dealer for the repair. Because all motor vehicle traffic is hazardous, requiring that people undertake extra driving will expose them to additional risk that may be more hazardous than the defect being repaired, if it is minor. In addition, risk regulations stimulate economic activity, such as manufacturing efforts to produce pollution control equipment or construction efforts to carry away the waste at a Superfund site. All economic activity is dangerous, leading to worker injuries and illnesses. Roughly 4 percent of every dollar of production in industry is associated with the health and safety costs of that production. Regulations that stimulate substantial economic efforts to meet the regulatory objectives will necessarily create risks in the process of stimulating economic activity. Even if for some reason the regulatory agency chooses to ignore the dollar costs, a comprehensive tally of the risk consequences of the effort may suggest that it is counterproductive.

The newest form of risk-risk analysis that has emerged has drawn on the negative relationship between individual income and mortality. Regulatory expenditures represent a real opportunity cost to society as they take away resources from other uses, such as health care, that might enhance individual well-being. As a result, there is a mortality cost associated with


these regulatory efforts. The U.S. OMB raised this issue with OSHA, suggesting that some of the more expensive OSHA regulations may in fact do more harm than good through these mortality effects.

Although the theoretical relationships are not controversial, the exact value of the regulatory expenditure that will lead to a statistical death remains a matter of debate. One approach has been to examine studies that directly link changes in individual income with mortality. Analysis by OMB economists Randall Lutter and John Morrall indicates that a statistical life may be lost for an income decrease on the order of $10 million to $15 million.16 Another approach is to establish a formal theoretical link between the value of life from the standpoint of saving statistical lives and the amount of money spent by the government that will lead to the loss of a statistical life through the adverse health effects of making society poorer.17 This approach, developed by W. Kip Viscusi, leads to a value of $50 million in government expenditures that will lead to the loss of a statistical life. If we also take into account the adverse health effects of higher income on consumption patterns, there may be the loss of one statistical life for every $17 million drop in income.18

Using this estimate to assess the consequences of some of the regulations in table 20.5 yields some disturbing results. Due to the risk-risk tradeoffs involved, the high cost per life saved of the OSHA asbestos regulation leads to the loss of 1.5 lives for every expected life saved. The OSHA formaldehyde standard is even worse, costing 25 lives for every expected death prevented by the regulation. The EPA hazardous waste listing for wood-preserving chemicals has a lives lost to lives saved ratio of 3. What these and many other examples of counterproductive regulations indicate is that there are real opportunity costs of regulations and that inordinate expenditure levels as reflected in the regulations at the bottom of table 20.5 have adverse health consequences that outweigh the beneficial effects of the regulation.

This literature is still in its early stages. However, the general principle suggests that regulatory agencies should be cognizant of the harm that is done when they fail to take costs into account. The concern of economists with cost is not a professional bias, but ultimately has a link to individual welfare. Such links in turn involve our health and are just as real as the concerns that motivate the government regulations.

Establishing Prices for Health, Safety, and Environmental Regulation

Perhaps the most difficult policy issues arising in the social regulation area will always stem from the setting of appropriate prices for the outcomes achieved. Because social regulation efforts deal in large part with outcomes that are not the result of explicit market transactions, there will always be a need to establish the value of these efforts.

As a society, we cannot allocate unlimited resources to any particular area of concern, however important it may seem. Because additional gains to health, safety, and the environment come at a diminishing rate for additional expenditures of money, we would quickly exhaust our resources long before we ran out of opportunities for spending. The general economic approach to formulating a benefit assessment is not particularly controversial, but some of the empirical methodologies for establishing such values are still in their development stage.

As the discussion in subsequent chapters will indicate, in many instances the absence of a specific empirical estimate for the benefit value is not the most pressing policy problem. Rather, there is a more fundamental difficulty in that the importance of making trade-offs at all has not even been recognized. In these cases, substantial gains could be made by noting that we are not in an unconstrained situation and that there must be some balancing among the competing objectives.

Questions and Problems

1. This chapter's discussion of the value of life has focused on estimates from the labor market. Economists have also estimated risk-dollar trade-offs based on price data for risky products. Smoke detector purchases, differences in riskiness of cars, and seat belt use decisions are among the contexts that have been considered. Can you think of any other market situations in which, if you had perfect data, it would be possible to infer an implicit risk-dollar trade-off?

2. Environmental damage resulting from oil spills, such as that inflicted by the Exxon Valdez, is subject to quite specific environmental penalties. In particular, the companies responsible for the damage are required to pay an amount sufficient to compensate society for the environmental loss that has occurred. In economic terms, this compensation must be sufficient to put society at the same level of utility we would have had if it had not been for the accident. Can you think of methodological approaches for determining the appropriate compensation amount for oil spills such as the Exxon Valdez, which led to the death of thousands of fish and birds, as well as oil residues on thousands of miles of Alaskan beaches?

3. Would you use the same value of life to assess the regulatory benefits in situations in which risks are incurred voluntarily, as opposed to situations in which they are incurred involuntarily? For example, would you treat smoking-risk regulation policies and nuclear hazard-risk regulation policies the same from the standpoint of benefit assessment?
4. Suppose we were faced with two policy alternatives. Under one alternative we will be saving identified lives, in particular Kip, John, and Joe. Under a second policy option, we know that we will be saving three lives at random from the population, but we do not know whose lives they will be. Should we attach the same benefit value to each of these instances?

5. A variant on question 4 pertains to the girl trapped in a well. The mayor has to decide whether it is worth the $15 million in rescue costs to get her out of the well. Given your knowledge of the value of statistical life, what would you recommend? Is this situation different than in regulatory contexts?

6. Suppose there are two policy options. Policy 1 affects a population of 10,000, of whom 100 will die, so that the risk of death per person is one in 100. The second policy will likewise save 100 individuals, but from a population of 1 million, so that the individual risk is one in 10,000. From the standpoint of regulatory policy, should we exhibit any preference for one policy over the other?

7. One mechanism for obtaining contingent valuation bids is to ask the respondent how much he or she is willing to pay for some outcome and then to ask if the respondent would be willing to pay, for example, 10 percent more. This process continues until the respondent is no longer willing to increase the bid. Some researchers have argued that this approach will lead to a bias in terms of eliciting the true response. What direction do you believe the bias is in, and why do you believe such a bias would occur?
Environmental Regulation

The range of activities in the area of environmental regulation is perhaps the most diverse of any regulatory agency.\(^1\) The U.S. Environmental Protection Agency (EPA) has programs to regulate emissions of air pollution from stationary sources, such as power plants, as well as from mobile sources, such as motor vehicles. In addition, it has regulations pertaining to the discharge of water pollutants and other waste products into the environment. These pollutants include not only conventional pollutants, such as the waste by-product of pulp and paper mills, but also toxic pollutants.

In situations in which its regulations of discharges and emissions are not sufficient, the EPA also undertakes efforts to restore the environment to its original condition through waste treatment plants and the removal and disposal of hazardous wastes. Insecticides and chemicals are also within the general jurisdiction of the agency’s efforts. Moreover, the time dimension of the agency’s concerns is quite sweeping, because the environmental problems being addressed range from imminent health hazards to long-term effects on the climate of Earth that may not be apparent for decades.

In this chapter we will not attempt to provide a comprehensive catalog of environmental regulations, although we will draw on a number of examples in this area. The focus instead will be on the general economic frameworks that are available for analyzing environmental problems. The structure of these problems generally tends to be characterized by similar economic mechanisms for different classes of pollutants. In each case there is a generation of externalities affecting parties who have not contracted to bear the environmental damage. A similar economic framework is consequently applicable to a broad variety of environmental problems.

We will begin with an analysis of the basic economic theory dealing with externalities and then turn to variations in this theory to analyze the choices among policy alternatives. The issues we will address include current policy concerns. Should the EPA pursue various kinds of marketable permit schemes or rely on technology-based standards?\(^2\) In addition, there is increasing concern with long-term environmental risks associated with climate change. How should we conceptualize the economic approach to regulating these and other risks that pose new classes of environmental problems? Finally, we will review the character of the enforcement of environmental regulation, as well as the ultimate impact of environmental policy on environmental quality.


The Coase Theorem for Externalities

The fundamental theorem in the area of externalities was developed by Ronald Coase. The generic problem that he considered was that of a cattle rancher. Suppose that farm A raises cattle, but that these cattle stray onto the fields in farm B, damaging farm B’s crops. The straying cattle consequently inflict an externality on farm B.

What Coase indicated is that assessing these issues is often quite complex. Among the issues that must be considered from an economic standpoint are the following. Should the cattle be allowed to stray from farm A to farm B? Should farm A be required to put up a fence, and if so, who should pay for it? What are the implications from an economic standpoint if farm A is assigned the property rights and farm B can compensate farm A for putting up a fence? Alternatively, if we were to assign the property rights to the victim in this situation, farm B, what would be the economic implications of assigning the property rights to farm A?

The perhaps surprising result developed by Coase is that from an economic efficiency standpoint, the fencing outcome will be the same irrespective of the assignment of property rights. If we assign the right to let cattle stray to farm A, then farm B will bribe farm A to construct a fence if the damage caused to farm B’s crops exceeds the cost of the fence. Thus, whenever it is efficient to construct a fence, farm B will compensate farm A and contract voluntarily to purchase the externality so as to eliminate it.

Alternatively, if we were to assign the property rights to farm B, farm A could construct the fence to prevent the damage. If the cost of such a fence exceeded the damage being inflicted, farm A could contract with farm B to compensate farm B for the damage imposed by the straying cattle. In each case, we will obtain the same result in terms of whether or not the fence is constructed irrespective of whether we give farm A or farm B the property rights.

From an equity standpoint, the results are, however, quite different. If we assign the property rights to farm A, then farm B must compensate farm A to construct the fence, or alternatively farm B must suffer the damage. In contrast, if we were to assign the property rights to farm B, the cost of the fence construction or the cost of compensation for the damage would be imposed on farm A. The outcome in terms of whether the crops will be trampled or the fence will be constructed will be the same regardless of the property right assignment. However, the well-being of each of the parties and the cash transfers that take place will be quite different under the two regimes.


4. One observant student noted that the manure left by the stray cows on farm B may be a positive externality. For concreteness, we will assume the net externality is negative.
Economists generally have little of a conclusive nature to say about which situation is more equitable. Coase observed that we should not be too hasty in making a judgment of which property right assignment was most fair. From an equity standpoint one should take into account the reciprocal nature of the problem. In this situation, farm A inflicts harm on farm B. However, to avoid the harm to farm B we must harm farm A. The objective from an efficiency standpoint is to avoid the more serious harm.

**The Coase Theorem as a Bargaining Game**

What Coase did not explore in detail was the nature of the bargaining process that would lead to the efficient outcome that he discussed. To address these issues, it is useful to cast the Coase theorem problem within the context of a simple bargaining game. For concreteness, let us suppose that the property rights are being assigned to the pollution victims, so that it is the firm that must pay for the damage or control costs.

Table 21.1 summarizes the generic components of this and other bargaining games. The company in this situation has a maximum offer amount that it is willing to give the pollution victims for the damage being inflicted. The factors driving the maximum offer value are the expenditures that the firm would have to make to eliminate the externality of the cost that would be imposed on the firm by the legal rules addressing involuntary externalities. The maximum amount that the firm will be willing to pay will be the minimum of either the control costs or the penalty that will be imposed on the firm if it inflicts the externality.

From the standpoint of the individuals bearing the accident costs, the minimum amount they are willing to accept in return for suffering the impacts of the pollution will be that amount of compensation that restores their level of utility to what it would have been in the absence of pollution. We will refer to this amount as the minimum acceptance value.

There is a potentially feasible bargaining range if the maximum offer the firms are willing to make exceeds the minimum acceptance amount, which is the first inequality

| Table 21.1 |
The Coase Theorem Bargaining Game |
| Feasible bargaining requirement: |
| Maximum offer ≥ Minimum acceptance |

| Bargaining rent: |
| Bargaining rent = Maximum offer − Minimum acceptance |

| Settlement with equal bargaining power: |
| Settlement outcome = \( \frac{\text{Maximum offer} + \text{Minimum acceptance}}{2} \) |
| = Minimum acceptance + 0.5\text{Bargaining rent} |
listed at the top of table 21.1. If this condition is not satisfied, no bargain will take place, inasmuch as there is no feasible bargaining range. In such a situation, in which the minimum acceptance amount by the pollution victims exceeds the maximum amount firms are willing to offer, there will be no contractual solution. Firms will select the minimum-cost alternative of either installing the control device or paying the legally required damages amount. The absence of a feasible bargaining range does not imply that the Coase theorem is not true or that the market has broken down. Rather, it simply indicates that there is no room for constructive bargaining between the two parties. In such situations, the resolution of the bargaining game will be dictated by the initial assignment of property rights.

An essential component of the bargaining game is the bargaining rent. This rent represents the net potential gains that will be shared by the two parties as a result of being able to strike a bargain. As indicated in table 21.1, the bargaining rent is defined as the difference between the maximum offer amount and the minimum acceptance value.

This definition is quite general and pertains to other bargaining situations as well. For example, suppose that you were willing to pay $18,000 for a new Honda Accord, but the cost to the dealer of this car is $15,000. There is a $3,000 spread between your maximum offer and the minimum acceptance amount by the dealer, which represents the bargaining rent available. The objective of each of you is to capture as much of the rent as possible. You would like to push the dealer as close to the minimum acceptance amount as possible, and the dealer would like to push you to your reservation price. Much of the bargaining process is spent trying to ascertain the minimum offer and maximum acceptance amounts, because these values are not generally disclosed. Moreover, in the process of trying to learn these values, one may reveal considerable information regarding one’s bargaining skill and knowledge of the other party’s reservation price. A bid for the car that is substantially below the cost to the dealer, for example, does not indicate that one is a shrewd and tough bargainer, but rather usually suggests that one does not have a well-developed sense of the appropriate price for the car. In a situation in which the parties are equally matched with equal bargaining power, they will split the economic rent.

This symmetric bargaining weight situation provides a convenient reference point for analyzing the bargaining outcome. As indicated in table 21.1, if there is such symmetry the settlement outcome will simply be an average of the maximum offer and the minimum acceptance amount, which is equivalent to the minimum acceptance amount plus one-half of the economic rent at stake.

A Pollution Example

To illustrate these concepts, let us consider the pollution problem summarized in table 21.2. The upstream pulp and paper mill emits discharges that impose $500 of environmental damage. The citizens can eliminate this damage by constructing a water purification plant for
Finally, suppose that the company could eliminate this pollution through primary treatment at the plant for a cost of $100.

To see the impact that differences in the property right assignment make, consider first the situation in which the citizen victims of pollution are assigned the property rights. In this context, it is the company that must bribe the citizens for the right to pollute. The maximum amount the polluting firm is willing to pay for this pollution privilege is $100—the cost of installing a treatment facility. The citizens, however, have a reservation price of $300—the lesser of the costs of the water pollution treatment and the environmental damage. Because the maximum offer amount is below the minimum acceptance value, there is no profitable bargain that can be made by the two parties. The result will be that the company will install the pollution treatment system, and there will be no cash transfer between the parties.

The second situation considered is one in which the polluter has been assigned the property rights. In this situation, the maximum offer by the citizens to the firm will be $300. This amount exceeds the $100 cost of installing water pollution treatment for the company, which is the company’s minimum acceptance amount. As a result, there is a profitable bargain that can be arranged between the two parties, with a total bargaining rent of $200. The outcome will be that the citizens will pay the company $100 to install the pollution control device. Moreover, if the bargaining power of the two parties is equal, the citizens will also pay the firm an additional $100 as the company’s share of the bargaining rent.

Utilization of this bargaining game framework to analyze the Coasian pollution problems provides a more realistic perspective on what will actually transpire than did the original Coase paper, which assumed that the purchase price for the transfers will equal the minimum acceptance amount by the party holding the property rights. In each case, the pollution control

<table>
<thead>
<tr>
<th>Table 21.2</th>
<th>Property Right Assignment and the Bargaining Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Basic aspects of the pollution problem</strong></td>
<td></td>
</tr>
<tr>
<td>Primary treatment of effluent: $100</td>
<td>Water purification $&lt; costs: $300</td>
</tr>
<tr>
<td><strong>Bargaining with victim-assigned property rights</strong></td>
<td></td>
</tr>
<tr>
<td>Bargaining equation: Maximum offer by company = $100: $&lt;</td>
<td>Minimum acceptance by citizens = $300: $&gt;</td>
</tr>
<tr>
<td>Outcome: Company installs controls. No cash transfer.</td>
<td></td>
</tr>
<tr>
<td><strong>Bargaining with polluter-assigned property rights</strong></td>
<td></td>
</tr>
<tr>
<td>Bargaining equation: Maximum offer by citizens = $300: $&gt;</td>
<td>Minimum acceptance by company = $100: $&lt;</td>
</tr>
<tr>
<td>Outcome: Citizens pay company $100 to install controls and also pay company $100 share of rent if equal bargaining power.</td>
<td></td>
</tr>
</tbody>
</table>
outcome is the same, as the company will install the water treatment device. However, in the case where citizens do not have the property rights, not only will they have to pay for the water treatment, they will also have to make an additional $100 transfer to the company that they would not have had to make had they been given the property rights.

The difference in the equity of the two situations is substantial. The citizens must spend $200 if they do not have the property rights—$100 for the treatment cost and $100 to induce the company to install it. If the citizens have the property rights, the cost is $100 to the company for treatment. In each case, the water treatment is the same.

**Long-Run Efficiency Concerns**

What should also be emphasized is that this short-run equity issue is also a long-run efficiency issue. Ideally, we want the incentives for entry of new firms into the industry to be governed by the full resource costs associated with their activities. If firms are in effect being subsidized for their pollution by citizens paying for their pollution control equipment, then there will be too much entry and too much economic activity in the polluting industries of the economy. We will return to this point within the context of the debate over standards versus taxes. This long-run efficiency point is often ignored by policymakers and by economists who focus on the short-run pollution outcome rather than on the long-run incentives that the property right assignment may create.

**Transaction Costs and Other Problems**

One factor pertaining to the bargaining process that Coase noted is that there may be substantial transaction costs involved in carrying out these bargains. Although we can generate an efficient outcome through a contractual solution without the need for any regulation, achieving this outcome may be quite costly. If there is a large number of citizens whose actions must be coordinated, then the cost may be substantial. These coordination costs are likely to be particularly large in situations in which there are free-riders. Some individuals may not wish to contribute to the pollution control effort in hopes of obtaining the benefits of controls without contributing to them.

It has often been remarked that there is also a potential for strategic behavior. Some parties may behave irrationally in the bargaining process. However, by modeling the contractual components of the externality market in table 21.2 using an explicit model of the bargaining structure, we capture these aspects within the context of a rational game theory model. It may, of course, be true that people are irrational, but this is true of any economic context and is not a phenomenon unique to externality bargaining contexts. For example, people may mis-perceive the probability of a particular bargaining response or may not assess the reservation price of the other party correctly.

Perhaps the greatest caveat pertains to the degree to which we can distinguish discrete and well-defined assignments of the property rights. Even in situations in which there is a
property right assignment, there are often limitations on the use of these property rights. Moreover, when the courts must enforce these rights, there is often imperfect information. The courts, for example, do not know the actual damages the citizens may incur. Moreover, they may not know with perfect certainty the pollution control and treatment costs. There are also costs to acquiring this information, and within the context of most judicial settings there is substantial error in the information being provided to the court.

The net result is that in actual practice we do not generally turn the market loose and let people contract out of the externalities that are imposed. The victims in the eastern United States who suffer the consequences of the acid rain generated by power plants in the Midwest cannot easily contract with these electric power plants. Even more difficult would be attempting to contract with the automobile users in the Midwest to alter their behavior. The bargaining costs and free-rider problems would be insurmountable. Indeed, in many cases we cannot even identify the party with whom we might strike a bargain. Unlabeled drums of toxic waste in a landfill do not provide a convenient starting point for externality contracts.

Despite the many limitations of the voluntary contractual approach to externalities, the Coase theorem does serve an important purpose from the standpoint of regulatory economics. In particular, by assessing the outcome that would prevail with an efficient market given different assignments of the property rights, one can better ascertain the character of the impact of a particular regulatory program. To the extent that the purpose of government regulation is to eliminate market failures and to ensure efficiency, the implications of the Coase theorem provide us with frames of reference that can be applied in assessing the character of the different situations that will prevail under alternative regulatory regimes. These concerns will be particularly prominent with respect to market-oriented regulatory alternatives that involve the explicit pricing of pollution.

Smoking Externalities

An interesting application of the Coase theorem is to cigarette smoking. Environmental tobacco smoke has become an increasingly prominent public concern and a classic externality issue. Many nonsmokers find cigarette smoke unpleasant, and government agencies such as the EPA and OSHA have concluded that there may be some adverse health effects as well, though the extent of these effects remains controversial. Indeed, a 1998 U.S. district court decision in the Flue-Cured Tobacco case rejected the EPA study as a sound basis for policy because it had “cherry picked” its data rather than doing a more comprehensive and balanced analysis.

Whether the health risks of environmental tobacco smoke are large or small, real or imagined, is not essential for addressing these exposures and is not critical to how the Coase
Theorem will operate in this instance. What is important from the standpoint of the Coase theorem problem is that nonsmokers would be willing to pay a positive amount of money to avoid being exposed to environmental tobacco smoke. Similarly, smokers would be willing to pay to be able to smoke in public places where they generate environmental tobacco smoke. As in the case of the Coase theorem problem, the externalities are in many respects symmetric. Smoking will make the smoker better off and the nonsmoker worse off, whereas restricting smoking will make the smoker worse off and the nonsmoker better off. This is the classic Coase situation.

Applying the Coase logic, one might expect the nonsmokers in restaurants to walk over to the smokers’ tables and attempt to strike a bargain to get them to stop smoking. Doing so, however, is unpleasant and consequently costly. However, there are other economic mechanisms that can reflect these concerns. If the restaurant does not have a suitable policy with respect to smoking, customers can eat elsewhere. In effect, the market operation in this context will be through the price system. The smoking policy of the restaurant is a local public good in much the same way as the music, the lighting, and the overall restaurant environment one. In situations in which customers are free to patronize different restaurants, the major remaining concern presumably would be with those who have found that they have made a mistake after arriving at the restaurant for the first time and finding it difficult to go elsewhere. Workplaces have responded similarly to the concerns of workers, inasmuch as some of them have banned smoking and others have instituted smoking areas.

The government has also become active in this area, as hundreds of local governments have enacted various kinds of smoking restrictions. While some national regulations have been proposed, many state regulations have been enacted. As of 1999, all but eight states had enacted smoking restrictions for hospitals, but only four states had enacted any kind of restrictions on smoking in bars, such as separate smoking areas. Enclosed arenas have tended to be regulatory targets, while malls have not. The overall pattern is consistent with what the Coase theorem would suggest in that the areas where the harm is greatest would emerge as the first candidates for regulation. The difference is that the mechanism is not private Coasean bargains, which would be quite costly to organize, but rather coordinated regulatory action. Interestingly, even a majority of smokers support smoking bans for hospitals and indoor sporting events, so that for some forms of smoking restrictions there are common rather than conflicting interests. The reliance on regulatory solutions even in situations when both smokers and nonsmokers may support restrictions highlights the important role of regulations to implement desirable social policies in situations where there are costly impediments to individual Coasean bargains.

While private Coasean bargains to solve regulatory problems are not the norm, the Coase framework provides a useful approach to guide our thinking about which regulation contexts the market will be expected to work in and which it will not work in. Moreover, if it is believed that the market will not work, should one then inquire what are the efficiency effects on both parties? What are the losses to the parties from the current situation, and what will be the losses with regulation? This is the essential message of the Coase theorem that is pertinent to all such externality contexts.

A final set of externalities associated with smoking pertain to insurance. If smoking is risky, as is the scientific consensus, then presumably the adverse health consequences will have widespread consequences for insurance costs. Health costs will clearly be higher. However, because smokers will die sooner under this scenario, their early departure will save society pension and Social Security costs. A comprehensive tally of these effects appears in table 21.3. As is indicated by the summary of the insurance externalities in table 21.3, the cost per pack generated by smokers is particularly high for health insurance. However, there are offsetting savings arising from the higher mortality rates of smoking, chiefly the lower pension and Social Security costs. Because smokers die sooner, they are also less likely to get long-term diseases such as Alzheimer’s, thus diminishing some of their medical expenses later in life. On balance, smokers save money for society in terms of the net externality cost. This result does not mean that smoking is not consequential for the individuals whose lives are at risk or for the particular insurance programs whose costs are affected, nor does it mean that the death of smokers is a desirable social outcome. However, this result does suggest that many externalities often involve competing effects with fairly complex ramifications.

The medical costs associated with smoking led to a wave of lawsuits by the states that in 1998 produced a $206 billion out-of-court settlement by the cigarette industry with forty-six

<table>
<thead>
<tr>
<th>Costs</th>
<th>Discount Rate (3 percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total medical care</td>
<td>0.580</td>
</tr>
<tr>
<td>Sick leave</td>
<td>0.013</td>
</tr>
<tr>
<td>Group life insurance</td>
<td>0.114</td>
</tr>
<tr>
<td>Nursing home care</td>
<td>-0.239</td>
</tr>
<tr>
<td>Retirement and pension</td>
<td>-1.259</td>
</tr>
<tr>
<td>Fires</td>
<td>0.017</td>
</tr>
<tr>
<td>Taxes on earnings</td>
<td>0.425</td>
</tr>
<tr>
<td>Total net costs</td>
<td>-0.319</td>
</tr>
</tbody>
</table>

states and separate agreements for $37 billion with four other states. Setting aside the merits of the suits and the particular settlement amount, how would one wish to structure the costs imposed on the cigarette industry in order to provide appropriate economic incentives, assuming that this payment reflects the financial externality generated? Many opponents of the cigarette industry want the cost to be imposed through a lump-sum imposition on the companies, so that the costs will come directly out of corporate profits; however, a profit tax will not affect the marginal cost of production or the product price. Another possibility is to link the payment to the level of cigarette sales, which is the approach that has been taken. The net effect is that consumers will, in effect, pay for most of this cost through higher cigarette prices rather than having the cost paid directly by the corporations. Such incentive effects are exactly what we would want to promote in order to foster efficient economic behavior whereby the parties generating costs will be cognizant of the economic consequences of their actions. Boosting the price of cigarettes in this manner is exactly analogous to proposals that firms should pay pollution taxes to reflect the environmental damage they generate, as such taxes will induce more efficient behavior.

**Special Features of Environmental Contexts**

In environmental contexts, it should also be noted that the character of the markets that would emerge if we set up a market for pollution may be quite unusual. Most existing water pollution regulation is based on the assumption that the usability of water tends to follow a step function, such as the one indicated in figure 21.1.7 Initially, the water quality is quite high, and we will label the water pristine. After a certain level of pollution the water is no longer pristine, but you can still drink it. After another increase in pollution the usability of the water for drinking declines, but you can swim in the water with appropriate vaccinations. As the pollution level increases further, water is suitable for fishing but no longer for the other uses. Finally, with a very high level of pollution, even the fishing option disappears. At this high pollution level, there is no additional marginal cost being imposed on the citizenry from additional pollution if we assume for concreteness that all of the beneficial uses of the water have disappeared. The citizens could then sell an infinite number of pollution rights without suffering any additional damage beyond what they have already suffered. Moreover, within any particular step of the declining water-quality curve in figure 21.1 there is no loss to the citizenry, so that the marginal costs to them of selling additional pollution rights will be zero.

7. Figure 21.1 is a bit of a simplification of our current understanding of water-quality levels. Although the EPA formerly used a water-quality ladder similar to that shown in this diagram, it is now believed that such a rigorous step function does not in fact hold. Rather, the EPA considers the following four different dimensions of water quality: drinking, swimming, fishing, and aquatic uses. The scores on these various dimensions are correlated in the same direction as would be the case if a water-quality ladder existed, and it remains the case that water that is not safe for any of these uses will exhibit the kind of nonconvexity that we will discuss.
This character of environmental contexts—known formally as an example of *nonconvexities*—suggests that instead of always dispersing the risks, it may be profitable to concentrate the risks in a particular location. For example, are we better off siting hazardous wastes throughout the United States, or should they be concentrated in one area? If they are concentrated, society can adapt by prohibiting residential housing and commercial operations near the facility, so that a large environmental risk can be present without imposing substantial costs on society. In contrast, dispersing hazardous wastes on a uniform basis throughout the United States may appear more equitable, but it will impose larger risks to society at large because it is more difficult to isolate such a large number of individual risks.

The main difficulty with concentrating the risk in this manner involves the appropriate compensation of those who are unlucky enough to have been selected to be put at risk. The option of concentrating the risk is particularly attractive in theory, but in practice it implies that one group in particular will bear a substantial part of the costs. The NIMBY—not in my backyard—phenomenon looms particularly large in such contexts. It is these kinds of equity issues and the potential role for compensation of various kinds that are highlighted by application of the Coase theorem and the implications that can be developed from it.
Siting Nuclear Wastes

These issues concerning the siting of hazardous wastes are abstractions for pedagogical purposes. The nuclear waste repository debate has highlighted the practical importance of these efficiency and equity concerns. The government had invested $9 billion to develop the Yucca Mountain site in Nevada as the central repository for unspent nuclear fuel. The alternative was to scatter the wastes more diffusely across sixty-eight different sites. A central, safe location that embodied a large investment to ensure a low risk level had considerable appeal from an efficiency standpoint. In 2004 a U.S. court of appeals focused on the risk issues and ruled that the facility’s protections extending for 10,000 years were too short. Matters became muddied even further as the NIMBY concerns entered the political arena debate during the 2004 presidential campaign.

The problem that prompted this controversy was that some risk could emerge under longer time horizons than the 10,000-year period for which safety would be assured. A National Academy of Sciences panel concluded that in 270,000 years, a person standing just outside the fence could be exposed to sixty times the allowable radiation dose. This allowable dosage threshold is not linked to a specific risk probability, but there is the belief that the risk level is not zero. As a result, the panel recommended that the safety standard be extended for 300,000 years.

How might economists have approached these nuclear waste siting issues differently? A useful starting point would be to assess the technological risk trade-offs involved. The Yucca Mountain site may not be risk-free forever, but scattering nuclear wastes across the country poses more substantial, immediate risks. So when judging any waste siting policy, the comparison should be with the available policy alternatives, not a costless, risk-free world that does not exist.

The scenarios under which the Yucca Mountain site could become risky have a certain element of science fiction about them. After all, 270,000 years away is a pretty long time. A lot could happen on the technological front over that period, making it possible to address the nuclear waste risks more effectively and more cheaply. The cost of remedial measures to address problems that may develop surely will go down over time, so that risk estimates based on current capabilities will be too high.

Discounting also will all but eliminate these far-distant risks as a matter of concern. Suppose we adopt a modest discount rate of 3 percent. Then a dollar of benefits 270,000 years from now has a present discounted value of \( (1/1.03)^{270,000} \). To see the effect of discounting, consider the following example. Instead of having only one person exposed to radiation at the Yucca Mountain fence, suppose we crammed 300 million people up against the fence. Also assume a worst case of radiation exposure that leads all of them to experience fatal cases of cancer. (Note that the future risk could have been eliminated by not letting people live in close proximity to the site.) On a discounted basis, the result of having 300 million people
exposed to risk at the site would be the equivalent of a one in 100,000 chance of cancer today for a single person. Quite simply, any reasonable discounting of effects that are hundreds of thousands of years away will all but eliminate them from the analysis.

But what if we don’t worry about discounting, and suppose that for another $9 billion, Yucca Mountain could be made safe for 300,000 years. Wouldn’t that be the risk-reducing choice? Such an investment may not be safer from a net health standpoint. As the risk-risk analysis in chapter 20 indicated, there is an opportunity cost of expenditures so that, according to some estimates, there is one statistical death for every $20 million that we divert from the usual bundle of consumer purchases. Spending another $9 billion of taxpayer money will be that much less that is not spent to improve people’s standard of living and will lead to 450 expected deaths today. So, if we abstract from financial considerations and focus strictly on health, the question becomes whether the remote discounted value of the small risks 270,000 years from now outweigh the 450 immediate deaths.

The economists’ framework for conceptualizing these issues is consequently quite different from that of scientists. The questions posed are not in terms of the period of time over which Yucca Mountain will be completely safe. Rather, what is the magnitude of these risks? What are the costs and benefits of reducing the risks? How do these policy options compare with other available choices? Will changes over time alter the costs of risk reduction? And, what are the opportunity costs in terms of money and lives if we adopt a particular strategy?

Selecting the Optimal Policy: Standards versus Fines

Lawyers and economists generally have different answers to the question of how one should structure regulatory policy. In situations in which there is an externality that we would like to prevent, the answer given by lawyers is to set a standard prescribing the behavior that is acceptable. The usual approach by economists is somewhat different, as they attempt to replicate what would have occurred in an efficient market by establishing a pricing mechanism for pollution.

As we will see, each of these approaches can potentially lead to the efficient degree of pollution control, depending on how the standards and fees are set. In analyzing these pollution control options, we will assume that society approaches the control decision with the objective of providing for an efficient degree of pollution control. In tightening the pollution control standard, we should consequently not do so past the point where the marginal benefits accruing to society from this tightening no longer exceed the marginal costs.

In practice, the standard setting guidelines administered by the EPA are much more stringent. In the case of the Clean Air Act, for example, the EPA is required by law to set ambient air quality standards irrespective of cost considerations. Moreover, not only is the EPA
required to ignore costs, ensuring safety is not sufficient. The agency’s legislation requires it to provide a “margin of safety” below the zero-risk level. The result is that standards are generally set at excessively stringent levels from the standpoint of equating marginal benefits and marginal costs, but there are informal efforts to achieve balancing based on affordability.

Figure 21.2 illustrates the character of the compliance costs with the degree of pollution control. By making allowances with respect to the availability and affordability of technologies, the EPA and other risk regulation agencies attempt to limit the stringency of their regulations to a point such as $PC^*$, where the cost function begins to rise quite steeply. Such informal considerations of affordability may limit the most extreme excesses of regulatory cost impacts.

**Setting the Pollution Tax**

The shortcomings in the market that give rise to the rationale for government regulation stem not only from the character of the cost function but also from the relationship of these costs
to the benefits of controlling environmental externalities that would not otherwise be handled in an unregulated market context. Figure 21.3 indicates the nature of the market equilibrium in a situation in which the externality is not priced, but rather was inflicted involuntarily on the citizenry. The focus of this curve is on the marginal benefits and marginal costs of the production of gasoline, where the externality consists of air pollution. The market is governed by the relationship of the demand for gasoline, given by the marginal benefit curve $MB$. In setting the quantity level that will be produced, the market will be guided by the marginal cost curve reflecting the private marginal cost of gasoline, leading to a production of gasoline given by $Q_0$, whereas the socially optimal level of gasoline production is $Q^\ast$. The prevailing market price for gasoline is given by $P_0$. To achieve efficient pricing of gasoline, what is needed is an optimal tax that raises the price of gasoline to the amount $P^\ast$. Alternatively, this purpose can be achieved by constraining the quantity of gasoline produced to $Q^\ast$, where market forces will drive the price of gasoline up to the point $P^\ast$.

Focusing on either prices or quantities can each achieve the desired result. In the case of the quantity restrictions, the revenues accruing from the higher price of gasoline will go to
the companies producing gasoline, whereas under a tax scheme the taxes will go to the government.

The choice between taxes and quantity constraints is not simply a question of administrative feasibility. There are also important dollar stakes involved in terms of the transfers among the various market participants. Because market outcomes will produce too much of the externality, some form of government intervention is potentially warranted. If we adopt the usual approach in which we wish to establish the appropriate pollution control standard, the objective is to equalize the marginal benefits and marginal costs of pollution reduction.

The Role of Heterogeneity

Figure 21.4 illustrates the marginal cost curve for pollution reduction to two firms. Firm 1 has a higher control cost for pollution, as is reflected in its higher marginal cost curve $MC_1$. Firm 2 has a lower pollution reduction marginal cost curve given by $MC_2$. In situations in which the cost curves differ and where we can make distinctions among firms, the optimal solution is to have a differential standard in different contexts. Thus we should set a tighter
standard in the situation in which the marginal cost curve is lower, and we can achieve pollution control level $PC_2$, as compared with the looser standard of $PC_1$ for the higher-cost firm.

Distinctions such as this arise often among industries. It may be easier for some industries to comply with pollution requirements given the character of their technologies. If it is easier for chemical plants to reduce their water pollutant discharges than it is for dye manufacturers, then we should set the standard more stringently in that case to recognize the difference in the marginal costs of compliance.

Perhaps more controversial are the distinctions that regulatory agencies make among firms within a given industry depending on the character of their technology. For new facilities that can incorporate the new pollution equipment as part of the plant design, the marginal cost curve for compliance is generally less than it will be for an existing facility that must retrofit the pollution control equipment onto its existing technology. It is consequently optimal from an economic standpoint to impose stricter standards on new sources than on existing sources because of the differences in the marginal cost curves.

This economic principle has given rise to what many observers have identified as a “new source bias” in the policies of the EPA and other government agencies.\(^8\) A new source bias is efficient, but one must be careful in determining the extent to which one will have biased policies that set differential standards. For firms such as those in figure 21.4, one can justify the differing degrees of stringency indicated by the difference in marginal costs. The danger is that we often move beyond such distinctions because of political pressures exerted by the representatives from existing and declining industrial regions that are attempting to diminish the competition from the growth areas of the economy, as B. Peter Pashigian has shown.\(^9\) Economics provides a rationale for some new source bias, but it does not necessarily justify the extent of the new source bias that has been incorporated within the context of EPA policy.

The Role of Uncertainty

Setting the optimal standard is most straightforward when compliance costs and benefits arising from policies are known. In the usual policy context, there is substantial uncertainty regarding these magnitudes. Figure 21.5 illustrates the familiar case in which the cost uncertainty is likely to be greater than the benefits uncertainty. For most policies with comparatively small impacts on the nation’s environment, the marginal benefit curve will be flat. Firms’ marginal cost curves for pollution control are not flat but rather tend to slope upward quite steeply. Moreover, there may be considerable uncertainty regarding the degree of

---


compliance costs because the technologies needed to attain compliance may not yet have been developed. As is illustrated in figure 21.5, the optimal degree of pollution control ranges from $PC_0$ in a situation in which the marginal cost curve is given by $MC_0$ to the intermediate case of $PC_1$ for a marginal cost curve of $MC_1$, to a very high level of pollution control at $PC_2$ for a marginal cost curve $MC_2$. In situations in which the marginal cost curve can lie between $MC_0$ and $MC_2$, the standard consequently could have a very substantial range, depending on how we assess compliance costs.

If we assess these costs incorrectly, then we run the risk of imposing costs that may not be justified. On the one hand, if we set the policy on the basis of a marginal cost curve of $MC_1$, where the true marginal cost curve is governed by $MC_0$, then there will be a needless cost imposed by the regulation. The shaded triangle in figure 21.5 that lies above line $MB$ gives the value of the excess costs that are incurred because the regulation has been set too stringently. On the other hand, there could also be a competing error in terms of forgone benefits if the standard is set too leniently at $PC_1$ when the regulation should have been set at $PC_2$. If the true marginal cost curve is $MC_1$ and it is believed to be $MC_2$, there will be a loss
in benefits from inadequate regulation. This outcome is illustrated in figure 21.5 by the triangle that lies below line $MB$, between $PC_1$ and $PC_2$.

Although setting standards intrinsically must address this problem of uncertain compliance costs, if we were to set a pollution fine equal to the level of the marginal benefit curve in figure 21.5, then firms could pick their quantity of pollution control on a decentralized basis after the pollution had been priced. This approach not only accommodates differences at a particular point in time in terms of technologies, but also accommodates uncertainty regarding the present technology and uncertainty regarding future technological development. If the uncertainty with respect to cost is greater than with respect to benefits, as most regulatory economists believe, then a fee system is preferable to a standards system in such situations.

Pollution Taxes

The operation of a pollution tax approach to promoting optimal pollution control is illustrated in figure 21.6. In particular, suppose that we set the price of pollution equal to the marginal benefits given by the horizontal curve in that diagram. This optimal fine will lead the firm to

![Figure 21.6](image)

**Figure 21.6**
Setting the Optimal Pollution Penalty
install the pollution control equipment needed to achieve the level of pollution control given by $PC^*$. The amount of pollution reduced is indicated on the horizontal axis, as is the amount of pollution remaining. In addition, the shaded portion of figure 21.6 indicates the total fine that firms must pay for their pollution. From the standpoint of short-run efficiency, achieving the pollution control level $PC^*$ through a standard or the fine system is equivalent. From the standpoint of the firms that must comply with this standard, however, the attractiveness of standards is much greater than that of fines. With a standard, the only costs incurred are the compliance costs, whereas under the fine system firms must pay both the compliance costs and the fine for all of the pollution that remains above the optimal control point.

This difference in outcomes raises two classes of issues. The first is whether the fine has any role to play other than simply being a transfer of resources from firms to the citizenry. In terms of the short-run efficiency, the fine does not alter the pollution control outcomes. However, from the standpoint of long-run efficiency, we want all economic actors to pay the full price of their actions.\(^{10}\) If they do not, the incentive to enter polluting industries will be too great. In effect, society at large will be providing a subsidy to these polluting industries equal to the value of the remaining pollution. Imposition of fines consequently has a constructive role to play from a standpoint of providing correct incentives for entry into the industry and long-run efficiency, even though it will not alter the degree of pollution control by an existing firm.

A second observation with respect to the penalty proposals is that the imposition of costs on firms can be altered to make its impact more similar to that of a standard by making the fine asymmetric. In particular, if we impose a fine only for pollution levels below the standard $PC^*$, then the purpose of the fine is to bring firms into compliance with the standard. In situations in which firms choose to pay the fine rather than install the necessary control equipment, it may be an index that the original standard was not set appropriately given the firm’s particular cost curves. Thus, fines may provide a mechanism to introduce flexibility into an otherwise rigid standard system that does not recognize the heterogeneity in compliance costs that does in fact exist.

Cost Heterogeneity for Water Pollution Control

Figure 21.7 illustrates the considerable variation in compliance costs with water pollution control standards for firms in the tissue paper industry. Although most firms in the industry can comply with the standards for under $6 per ton of effluent, for some very high cost compliers the compliance costs could be four times as great. Rather than have to set standards that reflect the wide differences in compliance costs that may exist, offering the firms the

\(^{10}\) A calculation of the optimal pollution tax to recognize these long-run incentive issues is a nontrivial economic problem. For an analysis of it, see Dennis Carlton and Glenn Lourry, “The Limitations of Pigouvian Taxes as a Long-Run Remedy for Externalities,” *Quarterly Journal of Economics* 94 (1980): 559–66.
option to pay a penalty if they fall short of the standard may be a way to promote efficient pollution control in situations in which there is uncertainty regarding compliance costs. Firms will not have an incentive to misrepresent their compliance costs in such an instance because they must pay the penalty if they cannot meet the standard.

**Current Market Trading Policies**

Although there has been substantial support for various kinds of fee systems in the economics literature for at least two decades, policymakers have been slow to implement these concepts.\(^{11}\) Four types of emissions trading options that are available are summarized in table 21.4.\(^{12}\) In each case firms must apply to the EPA to be permitted to use these mechanisms, and the requirements on such systems are very stringent because there is a continuing suspicion among environmentalists of market outcomes that enable firms to buy their way out of meeting a pollution control standard.

11. It should be emphasized, however, that economists within these administrations have long advocated this approach. Most recently, see the Council of Economic Advisors, *Economic Report of the President* (Washington, D.C.: U.S. Government Printing Office, 1990), chap. 6.

12. For further discussion of these trading options see Crandall, *Controlling Industrial Pollution*, or Hahn and Hester, “Where Did All the Markets Go?”

---

**Figure 21.7**
Distribution of Water Pollution Control Expenditures in the Tissue Paper Industry

Table 21.4
Summary of Emissions Trading Activity

<table>
<thead>
<tr>
<th>Activity</th>
<th>Estimated Number of Internal Transactions</th>
<th>Estimated Number of External Transactions</th>
<th>Estimated Cost Savings (millions)</th>
<th>Environmental Quality Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Netting</td>
<td>5,000–12,000</td>
<td>None</td>
<td>$25–$300 in permitting costs; $500–$12,000 in emission control costs</td>
<td>Insignificant in individual cases; probably insignificant in aggregate</td>
</tr>
<tr>
<td>Offsets</td>
<td>1,800</td>
<td>200</td>
<td>Probably large, but not easily measured</td>
<td>Probably insignificant</td>
</tr>
<tr>
<td>Bubbles:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Federally approved</td>
<td>40</td>
<td>2</td>
<td>$300</td>
<td>Insignificant</td>
</tr>
<tr>
<td>State approved</td>
<td>89</td>
<td>0</td>
<td>$135</td>
<td>Insignificant</td>
</tr>
<tr>
<td>Banking</td>
<td>&lt;100</td>
<td>&lt;20</td>
<td>Small</td>
<td>Insignificant</td>
</tr>
</tbody>
</table>


Netting

The first of the mechanisms listed in table 21.4 is netting. Under the netting system a firm can alter its current plant and equipment in a manner that increases the pollution emissions from one source at the plant, provided that it also decreases the emissions from other sources so that the net increase that occurs does not equal that of a major source. These trades cannot take place across firms but are restricted to within firms. Such trades have occurred in several thousand instances. The estimated cost savings from having this flexibility range from $25 million to $300 million in terms of the permitting costs and from $500 million to $12 billion in terms of emission control costs. For this as well as for the other market trading systems listed, the adverse environmental effect is believed to be minimal.

Offsets

The second most frequent market trading activity is offsets. Under an offset option, firms will be permitted to construct new facilities in a part of the country that exceeds the EPA’s maximum permissible level of pollutants. However, before the company can build a plant in such an area, it must purchase pollution offsets from some existing facility in that area that provides for more than an equivalent reduction of the same pollutant. Moreover, the party selling these offsets must already be in compliance with EPA standards. Although there were 1,800 offset purchases by the mid-1980s, for the most part these involved internal market trades rather than external transactions.
Bubbles

The third policy option was introduced with great fanfare in December 1989 by the Carter administration. Under the bubble concept a firm does not have to meet compliance requirements for every particular emissions source at a firm. Ordinarily, each smokestack would have to comply with a particular standard. Instead, the firm can envision the plant as if it has been surrounded by an artificial bubble. The compliance task then becomes that of restricting the total emissions that will emerge from this bubble to a particular level. This option enables the firm to have some flexibility in terms of what sources it will choose to control. If there are two smokestacks, for example, as in the case of figure 21.8, the firm will choose to achieve the greatest pollution reduction from smokestack 1, as these costs will be lower than for pollution reduction in smokestack 2. There have been over a hundred such bubbles approved by the EPA, with cost savings to DuPont and other firms totaling $435 million.

Banking

The final option is banking. Under the banking policy, firms in compliance with their standards can store pollution rights over time, and then use these rights in the future as an offset against future pollution. The use of this policy option has been fairly infrequent.
The Future of Market Approaches

A major policy shift occurred in the 1990s. President Bush, for example, declared a commitment to increase reliance on market trading options, and some programs of this type were implemented. The EPA has not, however, replaced the thrust of its policy standards effort with a tradable pollution permit system.

Nevertheless, permits have attractive economic features, as firms with the highest compliance costs can purchase them, thus fostering an efficient degree of control of pollution.

The first advantage of tradable pollution rights is that they enable the EPA to equalize the opportunity costs of pollution control. Second, they encourage innovations to decrease pollution, whereas a rigid standard only encourages a firm to meet the standard, not to go any further. Pollution rights systems also create less uncertainty for firms that must make fixed capital investments. Changing technology-based standards over time poses a risk that a firm’s capital investments will become obsolete.

The disadvantage of pollution rights is that we must set the number of such rights. Establishing the quantity of such rights is not too dissimilar from setting an aggregate pollution level. It requires a similar kind of information, and it probably relies on more imperfect forms of information than would establishing a penalty scheme. However, a fee system for all pollution generated imposes such substantial costs that there is currently political opposition to this approach.

Other criticisms of pollution rights systems pertain to whether the market participants are really trading a uniform good. The impact of pollution depends on the character of the pollutants, the stack height, and similar idiosyncratic factors. These pollutants also may interact with other pollutants in the area so that their consequences may differ. There also may be decreased ability to enforce marketable permit systems, as compared with a situation where the EPA mandates a particular technology for which officials can readily verify compliance. This concern may be of less consequence because many EPA standards, such as its water discharge requirements, are in terms of discharge amounts that must be monitored and reported on a monthly basis to the EPA.

The final concern that has been raised relates to market power. Will some large players, such as public utilities, buy up all of the pollution rights? Thus far, such concerns have not been of practical consequences.

By far the greatest resistance to the marketable permit scheme is the general suspicion of markets among noneconomists. Their counterargument often takes the following form: “Should the government also sell rights to murder?” A more appropriate question to use is, which policy approach will be most effective in reducing pollution at less cost? Although the

EPA has attempted to increase their salability by labeling such systems as ones in which firms sell pollution reduction credits rather than purchase pollution rights, these efforts continue to remain limited and fairly experimental in nature.

Global Warming and Irreversible Environmental Effects

Whereas the environmental policies of the 1970s focused primarily on conventional air and water pollutants, and efforts of the 1980s turned to toxic chemicals and hazardous waste, attention in the 1990s shifted to the long-term character of Earth’s climate.

Chief among these concerns is global warming. The accumulation of carbon dioxide and other trace gases in Earth’s atmosphere in effect has created a greenhouse around Earth. This change in Earth’s atmosphere is expected to produce global warming from 1990 to 2100 on the order of one to six degrees Celsius, or two to ten degrees Fahrenheit. Scientists continue to debate the magnitude and timing of the effect. Some global warming is inevitable irrespective of current efforts to impose environmental controls because of the irreversible nature of the generation of the greenhouse gases. We have already taken the actions that will harm our future environment. The extent of the future warming is uncertain because of both the substantial uncertainty regarding climatological models and the uncertainty regarding factors such as population growth and our pollution control efforts in the coming decades.

Even more problematic is the effect that global warming will have on society. Although the temperature will rise by several degrees, for northern regions this trend will be a benefit, and for southern regions it will generally be a disadvantage. The warming in the winter will be beneficial and will occur to a greater extent than the warming in the summer, which will have an adverse effect. Russia and Canada may benefit from longer growing seasons. Some have even questioned the desirability of a temperature change. Will global warming, for example, be tantamount to getting on a plane in Boston and arriving in Los Angeles? U.S. retirement patterns suggest that warmer weather may in fact be preferable. Change of any kind will necessarily lead to the imposition of some adjustment costs. Climatologists also predict that there will be an increase in damage from natural disasters such as hurricanes. The average sea level will rise, and there may be droughts in interior lands.

Assessing the Merits of Global Warming Policies

Although a precise assessment of the optimal policy relating to global warming is not possible, one can frame the issues and obtain a sense of the types of concerns that are being addressed within the context of what will prove to be an ongoing policy debate.14

Figure 21.9 sketches the marginal cost curves for addressing global warming by controlling the emission of greenhouse gases. This has been the approach taken by economists such as William D. Nordhaus. The first of the three policy options is reducing chlorofluorocarbons, such as bans on the use of freon in refrigerators. The second policy option listed is the imposition of a global carbon tax, which will penalize usage of gasoline or coal to produce energy, thus recognizing the environmental externalities they impose. The third policy option listed is reforestation. Additional forests serve to reduce the global warming problem by converting carbon dioxide into oxygen.

Also shown in figure 21.9 are two marginal benefit curves, one designated "MB (low)" and another designated "MB (high)." The purpose of illustrating the two curves is to indicate how the policy might change depending on our uncertainty regarding the ultimate societal implications that global warming will have.

What is clear from this figure is that even in the case of the low marginal benefit curve, some actions are clearly worthwhile. Elimination of chlorofluorocarbons and the imposition of some global carbon tax is clearly efficient, even in the case in which the low-benefit

15. Nordhaus, "Global Warming."
scenario prevails. If benefits are at a higher level, then policies of reforestation and a steeper global carbon tax are also worthwhile.

Whereas in most environmental contexts it is the marginal costs that are more uncertain than the marginal benefits, in this long-run environmental context, benefits also pose substantial uncertainty. This uncertainty is at a very fundamental level. There is even a debate over whether on balance, global warming will be beneficial or adverse to our economy. However, even at the very low level of costs that are assumed in figure 21.9, some policy options such as chlorofluorocarbon reduction are optimal.

**How Should We React to Uncertainty?**

Although further study to resolve these uncertainties is clearly a desirable policy alternative, if we were in a situation in which we had to take an action today, an economic issue arises as to whether the substantial uncertainties imply that we should err on the side of caution or err on the side of reckless abandon.

As we continue to study the climate change issue, there will also be calls for policy action. One approach that has gained widespread support is the “no regrets” option. We should clearly adopt policies, such as energy conservation, that would be desirable irrespective of what we ultimately learn about the implications of climate change. Whether we should go beyond the “no regrets” policy is more controversial.

Some insight into resolving this problem is provided by examining the classic irreversible development decision situation. Figure 21.10 illustrates the basic irreversible investment paradigm. A developer must choose the degree of current development, where the benefits and costs of this development at the present time are known. There is, however, uncertainty regarding the degree to which environmental preservation will be valued in the future. There is some probability \( p \) that the preservation will have a high value, and there is some probability \( 1 - p \) that the preservation will have the same value that it does at the present time. In such a situation of uncertainty, how should one choose the extent to which one will develop the scarce resource, such as conversion of a national forest into a shopping center and suburbs?

In general, the answer is that one should err on the side of underdevelopment in such situations. Moreover, the greater the probability that preservation will have a high value and the greater the increase that this value will be, the more one should alter one’s current decision from what one would select based on a myopic assessment of the benefits and costs of the development policy.

This principle for underdevelopment does not generalize to every situation in which there are irreversible decisions to be made. For example, companies installing pollution control

equipment might rationally choose to overinvest in such equipment if they expect the standard to be tightened in the future. Much depends on the character of the problem and the nature of the uncertainty. However, for problems like global warming, where the main uncertainty is with respect to the potential increase in the benefits of controls above current levels based on the current benefits associated with pollution control, the general policy maxim is that conservatism is the best policy.

Moreover, it is noteworthy that this conservatism arises wholly apart from the presence of any risk aversion. Society does not choose to err on the side of caution because we are unwilling to engage in risks. Rather, the bias arises because the expected payoffs from development in the future may be much less than they are today, and we should take this possible change in values into account.

### Multiperson Decisions and Group Externalities

Externality problems become particularly complex within the context of group decisions. In the situation of an individual firm and the citizenry, one has to worry only about the actions of one economic actor. However, in actual practice many of the most important externalities arise from the decentralized decisions of a variety of actors. In these contexts, some coordination mechanism is often desirable to promote behavior that will be collectively beneficial to society.

### The Prisoner’s Dilemma

The standard situation in bargaining theory where uncoordinated action gives rise to an inferior outcome is that of the Prisoner’s Dilemma. Suppose that there are two partners in crime, each of whom has been captured by the police. The prisoners are held separately, preventing
cooperation. The police offer to lighten each prisoner’s sentence if he will incriminate the other. The prisoner must make a risky decision, based on what he believes the other is most likely to do. Following the standard scenario, the prisoners each choose their preferred strategy of talking to the police. Talking is a dominant strategy for each party, given any particular behavior on the part of the other prisoner. However, if both of the prisoners had agreed not to talk, they would have been better off than they will be after they both incriminate one another. The outcome is consequently Pareto inferior (that is, each gets a lower-valued payoff) when compared to the situation in which both of them remained silent.

### The N-Person Prisoner’s Dilemma

A variety of social situations also arise in which there are incentives for individual behavior that do not lead to optimal group outcomes. Figure 21.11 illustrates a multiperson Prisoner’s Dilemma, using a methodology developed by Thomas C. Schelling, where the particular context being considered is the purchase of a large or small car.\(^\text{17}\) We will suppose for concreteness that consumers prefer large cars to small cars. Thus, for any given number of small cars

---

\(^{17}\) The diagrammatic exposition that follows is based on the innovative work of Thomas C. Schelling, *Micromotives and Macrobahavior* (New York: W.W. Norton, 1978).
cars on the market along the horizontal axis, the consumer’s payoff received for using a large car exceeds that for a small car. The result is that because everybody has a dominant strategy to purchase a large car, we end up at the equilibrium, 0. This equilibrium is not a social optimum, however. In particular, if we could constrain everyone to purchase a small car, we could reach the outcome at point $S$, which has a higher value than 0. The reason why some constraint is needed is that this is not a stable equilibrium. Any individual driver has an incentive to break away and purchase a large car, leading to an unraveling until we reach the stable equilibrium at 0. Thus, some government regulation is required.

Applications of the Prisoner’s Dilemma

Group externalities such as this arise in a variety of contexts. In international whaling, exercising some restraint in terms of the number of whales that are caught in any year will maximize the value of the whaling population. Thus, even if one were simply concerned with the commercial value of the whales, some limitation on whaling is optimal. However, from the standpoint of the individual fisherman it is always optimal to catch as many whales as you can. If all of the whaling vessels follow their dominant strategy, as most of them have, the result is that the whaling population will be overfished and that we will have a dwindling number of whales. In this instance, the optimal strategy is to provide for some restraint but not a complete abolition of whaling activities. Achieving this moderation in the degree of whaling has proven to be a long-term international regulatory problem.

The international whaling example has proven to be more than a hypothetical case. In 1994 the United States government proposed that the Georges Bank fishing area off New England be closed so that the species could revive. This fishing ground, which had formerly been one of the richest in the Atlantic Ocean, was the source of fish such as cod and haddock. This fishing area also served as the principal source of livelihood for fishing villages in New England, such as Gloucester. Restraints on fishing proved to be ineffective, which led the federal government in 1994 to propose the more drastic step of closing these fishing grounds altogether so that the fishing stocks could revive. Unfortunately, the difficulty in monitoring and enforcing appropriate fishing restrictions has proven to be so great that the government was led to a much more costly and disruptive regulatory policy option that has led to the abandonment of a fishing fleet and the shutdown of a major industry throughout much of the New England area.

Similar classes of issues arise within the context of vaccinations. If a critical mass in society has received an inoculation, it is not optimal to get vaccinated because the risk of contracting the disease will generally be much less than the expected health loss due to an adverse reaction to the vaccine. We clearly need some coordinating mechanism to ensure that a sufficient portion of the population has received the vaccination, but given the fact that society has established such a vaccination requirement, each of us has an incentive to be exempted from the vaccination.
Similarly, home owners who are doing battle against Japanese beetles will be able to diminish their efforts if all of their neighbors use insecticides. However, it is essential to establish a sufficiently broad insecticide use to control the beetle population. The initial insecticide user may obtain little benefit unless a sufficient number of his neighbors also use the insecticides. At low and high levels of community-wide insecticide use, the individual incentive to use insecticides will be lacking. There is no voluntary incentive for an unassisted market process to begin generating the decentralized decisions needed to reach the social optimum.

The general result that pertains in situations in which there are group externalities is that some form of coordination is often worthwhile. This coordination often takes the form of explicit regulations. Hockey players are required to wear helmets, traffic rules require that we drive on the right side of the road, and daylight saving requirements establish uniform changes in the time schedule for everyone. Individually, the payoff of shifting to daylight saving time is quite low if no one else in society shifts, but if we can all coordinate our actions, we will all be better off.

The Enforcement and Performance of Environmental Regulation

Enforcement Options and Consequences

The promulgation of regulations does not ensure that firms will comply with them. As a result, the EPA and other regulatory agencies couple the issuance of regulations with vigorous enforcement efforts. In the case of major sources of air and water pollution, the EPA attempts to inspect the emissions source at least once per year. Moreover, in the case of water pollution discharges, the EPA requires by law that the firms submit a record of the nature of the discharge to EPA and that each firm report its compliance status with the pollution permit that it has been given.

The enforcement task with respect to conventional pollutants is generally viewed as being the simplest. Next in terms of the degree of difficulty is enforcement with respect to toxic chemicals. These chemicals are often more difficult to monitor than are conventional pollutants because of specific chemical testing that must be undertaken.

The nature of the source of the pollution also affects the feasibility of effective enforcement. Hazards that arise on a decentralized basis, such as toxic wastes, radon in consumers’ homes, and asbestos in buildings, often impose substantial enforcement problems because of the large number of pollution sources involved and, in the case of toxic chemical dumping, the difficulty of monitoring the party responsible.

Enforcement of environmental regulations pertaining to chemicals and pesticides varies in effectiveness depending on the nature of the regulation. The process of screening chemicals and regulating the chemicals that are being sold and used commercially is quite effective
because of the ability to monitor mass-produced consumer goods. The EPA also can readily monitor the hazard warnings attached to these products. Much more difficult to monitor is the manner in which the products are used. The disposal of chemical containers and the dilution of insecticides are among the decentralized activities that pose almost insurmountable problems. The best that the EPA can achieve in these instances is to provide risk information to foster the appropriate safety-enhancing action on the part of the product users.

In these various inspection contexts, the EPA has several enforcement tools that it can use. Not all of these involve fines, but they do impose costs of various kinds on the affected firms. The EPA can inspect a firm. It can request that the firm provide data to it. It can send the firm letters, or it can meet with the firm’s managers to discuss pollution control problems. Most of the EPA’s contacts with firms are of this character.

In terms of sanctions, there are two classes of financial penalties that can be levied. The first consists of administrative penalties that are usually modest in size and limited in terms of the circumstances in which they can be levied. The main sanction that the EPA has is not the penalties that it can assess, but rather the penalties that can be assessed through prosecution of the polluter by the U.S. Department of Justice. In severe, flagrant, or persistent cases of violations of EPA standards, the EPA frequently refers the case to the U.S. Department of Justice for civil or criminal prosecution. The costs associated with the prospective litigation, as well as the possibility that substantial fines may be imposed, often provides a compelling enforcement sanction.

**Hazardous Wastes**

Public opinion polls typically rank the cleanup of hazardous wastes as one of the most important environmental problems. Beginning in the 1980s, the U.S. EPA became much more concerned with toxic substances and hazardous wastes. This cleanup effort, known as the Superfund Program, has sought to eliminate the risks posed by these chemical waste sites, which chiefly consisted of cancer hazards to the surrounding population.

What is perhaps most striking about this environmental policy area is the substantial mismatch between the public’s concern with the environmental risks and the efficacy of the environmental cleanup effort. The source of the difficulty can be traced in part to the legislative mandates under which the EPA operates. There is no stipulation that the EPA balance the benefits to surrounding populations against the costs of cleanup, but instead the focus is on risk alone. Moreover, since the cleanup costs will be borne largely by the potentially responsible parties, which are private firms rather than the citizens affected by the hazard, there will be considerable political pressure for uncompromising cleanup remedies, such as removing the contaminated waste from the site and incinerating it.

The policy trigger for cleanup is that a site must be cleaned up if it poses a potential lifetime cancer risk of at least one in 10,000, and cleanup is at the EPA’s discretion provided the lifetime risk is at one in 1,000,000 or more. Recall from table 19.3 that many routine daily
activities pose a risk from a single event of one chance in a million. Eating forty tablespoons of peanut butter, traveling ten miles by bicycle, and smoking 1.4 cigarettes all pose a one-in-a-million fatality risk. If one were to undertake such activities over one’s lifetime rather than in a single episode, then the overall risk would be even greater and would dwarf that posed by many hazardous waste sites that have been targeted for cleanup.

Before deciding on the level of the hazard, the EPA must first ascertain who lives near the site and will be exposed to the risk. In addition to examining current populations, the EPA assumes there is a risk if there is some potential chance that a future population could be exposed to the risk, even if such a chance is unlikely. Supreme Court Justice Stephen Breyer, for example, noted that at one Superfund site involving a case in which he ruled, a modest cleanup effort could make the dirt at the site clean enough so that children could eat the dirt for 70 days per year. However, the EPA spent an additional $9.3 million to clean up the site so that children would be able to eat the dirt without risk for up to 245 days per year. What was noteworthy about the site is that no children lived near the site, which was a swamp. Similar unrealistic assumptions may affect the risk estimates at other sites, such as the North Carolina Superfund site at which it is assumed that a factory will be built in the future and that during their lunch break workers will swim in a nearby creek, exposing them to the contaminated water.

For the EPA to find a risk there need not be a population actually exposed to the hazardous waste site. If a person could potentially move to that area and have some potential for exposure in the future, then the EPA will treat the risk as being just as consequential as would be the case if there were a large exposed population. The net result is that whether there are in fact exposed populations plays no role in triggering an EPA cleanup, a fact which will have important consequences for the efficacy of cleanups.

Note that the trigger for cleanup is whether a real or hypothetically exposed future individual has reached a critical lifetime risk threshold. This focus on individual risks consequently ignores the size of the total population at risk. Densely populated areas in close proximity to a Superfund site receive the same policy weight as a single hypothetically exposed future individual. Because minority populations tend to be disproportionately concentrated near hazardous waste sites, the practical effect of this approach is to give inadequate weight to such sites in the priority setting process. In contrast, consideration of economic benefits in terms of the expected number of cancer cases prevented would give greater weight to these highly populated sites.

In calculating the individual cancer risk at a site, the EPA uses conservative risk estimates. To see how conservative biases enter the analysis, it is useful to consider the components of the calculation. Lifetime excess cancer risk is given by

The denominator terms are not controversial, as the EPA uses an average body weight assumption, while the averaging time component simply controls for the proper units in the calculation. The key components are the five elements in the numerator of the calculation. For each of these individual variables, the EPA uses a conservative assumption, typically the 95th percentile of the distribution. Thus, there is only one chance in 20 that the exposure duration could be as great as the assumed value. By using such upper bound values for each of the five parameters in the numerator, the result is a cascading of conservatism bias so that the resulting estimate is well beyond the 99.99th percentile of the true risk distribution.

Economists instead would generally recommend calculating the expected number of cancer cases based on the mean values of the risk. If there is political support for being very protective in the cleanup actions, that concern can be expressed through a high unit benefit value on the cancer cases prevented. The current EPA practice distorts regulatory priorities by shifting the policy emphasis toward dimly understood risks that may pose no threat to existing populations.

What will be the consequence of ignoring cleanup costs, mismeasuring the magnitude of the risk, and failing to account for the size of exposed populations? One would expect hazardous waste cleanup efforts to be very ineffective. Table 21.5 summarizes the cost-effectiveness of various Superfund cleanup efforts measured in terms of the cost per expected case of cancer prevented. After ranking the sites from the most cost-effective to the least cost-effective, James T. Hamilton and W. Kip Viscusi calculated the cost-effectiveness at these

Table 21.5
Summary of Superfund Cost-Effectiveness*

<table>
<thead>
<tr>
<th>Percentage of Remediation Expenditures, Ranked by Cancer Cost Effectiveness</th>
<th>Cumulative Percentage of Total Expected Cancer Cases Averted</th>
<th>Marginal Cost per Cancer Case Averted ($ millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>99.47</td>
<td>145</td>
</tr>
<tr>
<td>25</td>
<td>99.86</td>
<td>1,107</td>
</tr>
<tr>
<td>50</td>
<td>99.96</td>
<td>6,442</td>
</tr>
<tr>
<td>75</td>
<td>99.97</td>
<td>28,257</td>
</tr>
<tr>
<td>95</td>
<td>99.98</td>
<td>241,058</td>
</tr>
</tbody>
</table>

* Using the following assumptions: average exposure concentrations and intake parameters, 3 percent discount rate and no growth factors for cost, 3 percent discount rate for cancers, and a ten-year latency period for the development of cancer.

different levels. U.S. Supreme Court Justice Stephen Breyer and others have hypothesized that there is a 90–10 principle, whereby agencies expend 90 percent of their resources to clean up the last 10 percent of the risk. In the case of the Superfund Program, the drop-off in efficacy is much more stark. The 5 percent of Superfund cleanup efforts that are the most effective address the hazards that will eliminate 99 percent of the human health risks. As indicated by the statistics in table 21.5, less than 1 percent of the cancer cases are eliminated for the least effective 95 percent of the expenditures.

The drop-off in cost-effectiveness is enormous. By the fifth percentile the cost per case of cancer prevented is $145 million, and the median Superfund cleanup expenditure prevents cases of cancer at a cost of $6.4 billion per case. Even these estimates, high as they are, understate the actual cost per case of cancer prevented because they are based on conservative EPA health risk assumptions and conservative assumptions about the degree to which populations will in fact be exposed to the risk.

If EPA policy decisions are not responsive to economic efficiency, what is it that drives them? The factors that appear to be most influential are political. The voting rate for the county, for example, is particularly influential in determining whether a site is cleaned up and the stringency of cleanup.

The cleanup of hazardous waste has also been the focal point of the environmental equity movement, whereby many have suggested that minorities are disproportionately exposed to hazardous waste. Much of this problem can be traced to the fact that minorities have less political leverage than do more affluent white populations. It is also noteworthy that targeting cleanups based on the economic efficiency concerns would do more to advance environmental equity than the current politically based process. Minority sites have higher benefit-cost ratios, or lower cost levels per case of cancer prevented. Economic efficiency concerns are in fact supportive of environmental equity by making cleanup of hazardous wastes equally meritorious, irrespective of the political clout of the affected population.

**Contingent Valuation for the *Exxon Valdez* Oil Spill**

One of the most controversial areas on the frontier of environmental economics is the use of contingent valuation techniques to value environmental damages. Under this approach researchers design survey questions to elicit the values that people attach to scarce environmental resources for which no good market values exist. The debate over the soundness of the technique reached its peak with respect to the *Exxon Valdez* oil spill, for which damage levels in the billions of dollars raised the stakes of the economic debate over this

---

methodology considerably. Calculating the losses to fishermen, the costs of cleaning up the shoreline, and related financial allocations is fairly straightforward. However, how should one calculate the environmental damages suffered by the entire U.S. citizenry because of the Exxon Valdez oil spill? Controversies over the use of contingent valuation continue to rage and have led to a government report that sought to provide some guidance; contributors to the report included two Nobel laureates in economics, Kenneth Arrow and Robert Solow.20

To get some sense of the nature of this enterprise, consider the contingent valuation study undertaken by the state of Alaska as part of the litigation over the environmental damage. The implications of this survey were not simply of academic interest but served as critical inputs to the litigation process and would have played a substantial role in the court deliberations had the case not been settled out of court. The U.S. Department of Justice also undertook a series of contingent valuation studies, and the Exxon Corporation solicited numerous economists as well to assess the damages and to comment on the validity of the other parties’ assessments.

The survey developed for the state of Alaska reflects the general character of the contingent valuation approach.21 The objective was to determine how much the public should be compensated to offset the loss they suffered because of the spill. After asking respondents about their general views on a variety of policy issues, the survey asked respondents if they were acquainted with the Exxon Valdez oil spill, which occurred in Prince William Sound, Alaska, in March 1989. As a result of this incident, 11 million gallons of crude oil spilled into the water. At that point the respondents indicate whether they recalled hearing of the Exxon Valdez oil spill. Although the survey proceeds regardless of whether they had heard of this spill, some debate remains as to whether the valuations should matter if people have not heard of the spill and suffered a welfare loss. The alternative perspective is that the value assigned should be based on what it would be if people had full information regarding it.

The survey then undertook a substantial educational effort regarding the character of the spill. Respondents considered maps and photos indicating the area on Prince William Sound that was contaminated by the spill. In addition, they were shown photos of wildlife in the area, including sea ducks, murres, seagulls, and sea otters. After viewing a picture of the tanker sailing through the sound, respondents then considered a variety of maps indicating the extent of the spill, which affected about 1,000 miles of shoreline. They also viewed a series of photos showing the oiled shore and the cleanup activity. Notwithstanding these


21. The following discussion will be based on the Exxon Valdez C.V. Survey Questionnaire, National Opinion Survey Main Interview Questionnaire, administered for the state of Alaska by Westat. The principal researchers included Richard Carson, Robert Mitchell, and other economists.
efforts, there was a significant effect on wildlife. The survey informed the respondents that “22,600 dead birds were found” and that scientists estimate that “the total number of birds killed by the spill is between 75,000 and 150,000.” This death total included 5,000 bald eagles. Respondents also learned that 580 otters and 100 seals were killed by the spill. They received information about how long it would take for these populations to return to normal.

One of the critiques of contingent valuation studies is that respondents may not be sensitive to whether they have learned that 100 birds or 10,000 birds have been killed, as they may give the same willingness-to-pay answer to prevent either incident.

The unresponsiveness of the willingness-to-pay values to the extent of the environmental damage has been designated the “embedding” problem. It may be that respondents are not in fact expressing their preference for the particular environmental good specified in the survey but rather are simply voicing support for the environment more generally. Incorporating a detailed series of rationality tests in a survey can help test for whether this potential problem is in fact pertinent.

After learning of the damage caused by the Exxon Valdez spill, respondents are asked how much they would be willing to pay for an escort ship policy that would prevent such spills from occurring over the next ten years. Without such a program, there would be one spill expected on average, according to the survey. The price mechanism would be a one-time tax on both oil companies and on households, where the household tax would be levied through higher federal income taxes. Respondents then considered a variety of possible costs for the program, such as $60 per household in higher taxes, and were asked whether they would vote for such an effort. The median response for the households was a willingness to pay for the escort program on the order of $49 per household, or a total value for the United States of $2.8 billion. Some critics might think that this willingness to pay is inordinately large for a comparatively modest escort program.

Although the approach taken in the Alaskan survey is one possible survey methodology, there are others as well. Surveys can differ considerably with respect to the level of detail that is presented to the respondents about the spill. Moreover, how the effects of the spill are presented can be influential. If the respondents were to consider the percentage of birds in the local population that were affected as opposed to the absolute number, their view might be different. Moreover, there are a variety of policy contexts that could be used. One might, for example, ask how much one would be willing to pay to reverse the effects of the spill through an ambitious cleanup operation. For prospective scenarios, there is a wide range of policy options that one could suggest to respondents as being potentially effective. Moreover, there are payment mechanisms other than higher federal taxes that could come into play, such as higher gasoline prices. What is essential, however, is that the payment mechanism be credible and that respondents indicate their true willingness to pay for prevention efforts, rather than simply naming some hypothetical dollar figure to impress an interviewer who has spent half an hour showing them pictures of dead birds and sullied shorelines. Because no reliable
market prices exist for many natural resources and because these valuations are critical both in court cases and for policy decisions, the controversies over how outcomes should be valued will continue to rage for many years to come.

The Senior Discount for the Value of Life

In 2003 the EPA generated a national controversy with respect to the value it applied to the reduced risks for senior citizens. The context of this controversy was a proposed air pollution policy called the Clear Skies Initiative. With many air pollution efforts, the benefits are concentrated at the tails of the population, as children and the elderly are most at risk. Should the same benefit value be applied to each age group? The elderly have a much shorter life expectancy at risk, so some age adjustment might seem reasonable. Based on willingness-to-pay survey results, the EPA applied a 37 percent senior discount to their benefits, leading to an outcry from senior citizen groups such as AARP: “Seniors on sale, 37% off.” The EPA administrator resigned shortly after this controversy emerged.

Table 21.6 summarizes the reduced fatalities and the associated benefits. There were two different benefit estimates, one based on long-term exposures and one based on short-term exposures. If decreased expected fatalities are valued using the EPA’s uniform value of $6.1 million, then one obtains the benefit estimate in the constant value of life column. The EPA also showed alternative benefit estimates adopting a 37 percent senior discount, and this change decreases benefits by $13.5 billion based on long-term exposures and by $7.2 billion

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Reduced Annual Fatalities in 2010</th>
<th>Benefits of Reduced Mortality ($ billions undiscounted)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Constant Value of Life</td>
</tr>
<tr>
<td><strong>Base Estimates—Long-Term Exposure:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adults, 18–64</td>
<td>1,900</td>
<td>11.6</td>
</tr>
<tr>
<td>Adults, 65 and older</td>
<td>6,000</td>
<td>36.6</td>
</tr>
<tr>
<td><strong>Alternative Estimate—Short-Term Exposure:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Children, 0–17</td>
<td>30</td>
<td>0.2</td>
</tr>
<tr>
<td>Adults, 18–64</td>
<td>1,100</td>
<td>6.7</td>
</tr>
<tr>
<td>Adults, 65 and older</td>
<td>3,600</td>
<td>21.9</td>
</tr>
</tbody>
</table>

based on short-term exposures. The stakes involved in terms of the level of benefits were quite substantial.

To address this sensitive issue, it is useful to go back to first principles. The appropriate benefit value is the willingness to pay for the risk reduction. This amount could decline for those with a shorter life expectancy, but it also might remain high because of increases in wealth with age. What matters is this willingness-to-pay value, not the quantity of life per se. On a theoretical basis, the value of statistical life should rise and then eventually fall over the life cycle. The main open question is empirical. How much does this value decline for those with short life expectancies?

The empirical evidence is still emerging. One survey of respondents in the United Kingdom indicated that the willingness to pay dropped by 37 percent for senior citizens, while a survey in Canada showed a comparatively flat relationship with age. Labor market evidence on workers’ value of statistical life is consistent with the inverted U-shaped relationship, but estimates differ in terms of the steepness of this decline at the upper end of the age distribution. One set of estimates that account for changes in the level of consumption over time indicates that the value of statistical life for those in their sixties is below their peak lifetime value, but is still above comparable values for people in their twenties and thirties. Using these values in the final column of table 21.6 boosts the overall benefit values, even if there is a “senior discount” relative to one’s peak value of statistical life. While a definitive set of estimates of the appropriate senior discount has not emerged, economic estimates of the role of such heterogeneity in benefit values are likely to reach a consensus in much the same way as agreement has been reached on the general range of the average value of statistical life.

But what should we do about equity concerns? If we value the lives of seniors less, isn’t that unfair? Alternatively, isn’t it unfair to value young people’s lives at the same amount as seniors because doing so places a higher value on each year of life for seniors than each year of life for those at the early end of the age distribution? Viewed in this manner, the seniors’ claim to be treated “fairly” may seem less compelling. However, if senior citizens do have a high willingness to pay for risk reductions, it would be inefficient and unfair not to count their benefit values.

Evaluating Performance

The objective of regulatory policy is not simply to promulgate and enforce regulations, but also to improve environmental outcomes. Assessing the impact of regulations is complicated by the fact that we observe trends in environmental quality, but we do not know what these trends would have been in the absence of regulation. Nevertheless, examination of pollution trends reveals the kinds of progress reflected in more formal statistical analyses.

Table 21.7 summarizes the pollution trends from 1970 to 2001 for five principal categories of air pollution emissions. One category not shown is that of lead pollution, which has been all but eliminated by EPA regulation. Since 1970 all but one pollutant category has exhibited
Table 21.7
National Pollution Emissions Trends

<table>
<thead>
<tr>
<th>Year</th>
<th>PM-10</th>
<th>Sulfur Dioxides</th>
<th>Nitrogen Oxides</th>
<th>Carbon Monoxide</th>
<th>VOCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1970</td>
<td>13.0</td>
<td>31.2</td>
<td>26.9</td>
<td>204.0</td>
<td>34.7</td>
</tr>
<tr>
<td>1975</td>
<td>7.6</td>
<td>28.0</td>
<td>26.4</td>
<td>188.4</td>
<td>30.8</td>
</tr>
<tr>
<td>1980</td>
<td>7.0</td>
<td>25.9</td>
<td>27.1</td>
<td>185.4</td>
<td>31.1</td>
</tr>
<tr>
<td>1981</td>
<td>6.5</td>
<td>24.6</td>
<td>26.8</td>
<td>182.2</td>
<td>29.3</td>
</tr>
<tr>
<td>1982</td>
<td>5.2</td>
<td>23.2</td>
<td>26.4</td>
<td>177.7</td>
<td>27.8</td>
</tr>
<tr>
<td>1983</td>
<td>6.0</td>
<td>22.6</td>
<td>26.2</td>
<td>179.2</td>
<td>28.5</td>
</tr>
<tr>
<td>1984</td>
<td>6.2</td>
<td>23.5</td>
<td>26.7</td>
<td>176.6</td>
<td>29.2</td>
</tr>
<tr>
<td>1985</td>
<td>41.3</td>
<td>23.3</td>
<td>25.8</td>
<td>176.8</td>
<td>27.4</td>
</tr>
<tr>
<td>1986</td>
<td>40.5</td>
<td>22.5</td>
<td>25.4</td>
<td>173.7</td>
<td>26.8</td>
</tr>
<tr>
<td>1987</td>
<td>40.8</td>
<td>22.3</td>
<td>25.6</td>
<td>173.0</td>
<td>26.7</td>
</tr>
<tr>
<td>1988</td>
<td>42.8</td>
<td>22.7</td>
<td>25.6</td>
<td>174.4</td>
<td>27.0</td>
</tr>
<tr>
<td>1989</td>
<td>40.8</td>
<td>22.8</td>
<td>25.4</td>
<td>160.5</td>
<td>25.6</td>
</tr>
<tr>
<td>1990</td>
<td>27.8</td>
<td>23.1</td>
<td>25.5</td>
<td>154.2</td>
<td>24.1</td>
</tr>
<tr>
<td>1991</td>
<td>27.3</td>
<td>22.4</td>
<td>25.2</td>
<td>147.1</td>
<td>23.6</td>
</tr>
<tr>
<td>1992</td>
<td>27.1</td>
<td>22.1</td>
<td>25.3</td>
<td>140.9</td>
<td>23.1</td>
</tr>
<tr>
<td>1993</td>
<td>27.4</td>
<td>21.8</td>
<td>25.4</td>
<td>135.9</td>
<td>22.7</td>
</tr>
<tr>
<td>1994</td>
<td>28.6</td>
<td>21.3</td>
<td>25.3</td>
<td>133.6</td>
<td>22.6</td>
</tr>
<tr>
<td>1995</td>
<td>25.8</td>
<td>18.6</td>
<td>24.0</td>
<td>126.8</td>
<td>22.0</td>
</tr>
<tr>
<td>1996</td>
<td>22.9</td>
<td>18.4</td>
<td>24.8</td>
<td>128.9</td>
<td>20.9</td>
</tr>
<tr>
<td>1997</td>
<td>22.9</td>
<td>18.8</td>
<td>24.7</td>
<td>117.9</td>
<td>19.5</td>
</tr>
<tr>
<td>1998</td>
<td>22.9</td>
<td>18.9</td>
<td>24.3</td>
<td>115.4</td>
<td>18.8</td>
</tr>
<tr>
<td>1999</td>
<td>21.6</td>
<td>17.7</td>
<td>23.7</td>
<td>117.2</td>
<td>19.4</td>
</tr>
<tr>
<td>2000</td>
<td>24.7</td>
<td>16.3</td>
<td>23.2</td>
<td>123.6</td>
<td>19.7</td>
</tr>
<tr>
<td>2001</td>
<td>24.1</td>
<td>15.8</td>
<td>22.3</td>
<td>120.8</td>
<td>18.0</td>
</tr>
</tbody>
</table>

Percentage Annual Growth Rate

<table>
<thead>
<tr>
<th>Period</th>
<th>PM-10</th>
<th>Sulfur Dioxides</th>
<th>Nitrogen Oxides</th>
<th>Carbon Monoxide</th>
<th>VOCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1970–79</td>
<td>–5.6</td>
<td>–1.5</td>
<td>0.1</td>
<td>–0.9</td>
<td>–0.8</td>
</tr>
<tr>
<td>1980–89</td>
<td>17.6</td>
<td>–1.3</td>
<td>–0.6</td>
<td>–1.4</td>
<td>–2.0</td>
</tr>
<tr>
<td>1990–99</td>
<td>–2.5</td>
<td>–2.7</td>
<td>–0.8</td>
<td>–2.7</td>
<td>–2.2</td>
</tr>
</tbody>
</table>

Note: PM-10 refers to particulate matter less than 10 micrometers in diameter, and includes small particles of dust, dirt, soot, smoke, and liquid droplets often associated with fossil fuel combustion, fires, and natural windblown dust. VOCs are volatile organic compounds and are a precursor to ozone (ground-level smog). VOCs are emitted through fossil fuel combustion, as well as in chemical manufacturing, dry cleaning, and other activities using solvents.

steady progress. Even particulate matter (PM-10, usually arising from fuel combustion, industrial processes, and motor vehicles) has exhibited improvement since 1988. The other pollution categories displayed more consistent improvement, including sulfur oxide emissions (chiefly arising from stationary fuel combustion and industrial processes), nitrogen oxide emissions (arising primarily from highway motor vehicles and cold-fired electric utility boilers), carbon monoxide emissions (primarily arising from highway motor vehicles), and volatile organic compounds (primarily from fossil fuel consumption and chemical manufacturing).

It is also noteworthy that the estimated benefits for this regulatory success exceeded the estimated costs. This regulation was one of the few EPA regulations that passed a test of economic desirability, and the result was a dramatic improvement in lead pollution levels achieved at reasonable cost.

Although a precise test of the EPA’s impact on these various pollution measures has not yet been undertaken, it is clear that some progress has been made. Because one would have expected an increase in pollution levels with an expanding economy and a growing population, the fact that there was any decrease in the pollution, much less the dramatic declines that have occurred since 1970, is evidence of some payoff to society from the costs that have been incurred.

Summary

Environmental problems represent a classic situation in which there is an externality being imposed involuntarily. What is most noteworthy about this situation is that the optimal level of pollution is not zero. The fact that there is an externality that is being imposed without a voluntary contract does not mean that the activity should be prohibited. Whether we are talking about second-hand smoke or toxic waste disposal, the efficient level of pollution is generally not zero. However, the efficient level of pollution is also generally not going to be what arises within a market context, because the party generating the pollution has inadequate incentives to reflect the social cost imposed in its decisions.

Our review of the Coase theorem indicated that the main focal point should be the efficient pollution level, which is the level that would arise under a voluntary contractual situation if parties could contract costlessly. Examining pollution problems within the context of the bargaining problems used to illuminate the Coase theorem also sheds light on the distributional impacts involved. Assignment of property rights not only has distributional implication but also affects the long-run efficiency aspects of the system.

Similar concerns arise with respect to the choice of standards versus fines. Each of these approaches can provide for the same degree of short-run efficiency that can be achieved through a Coasian contractual outcome. However, standards differ from fines in terms of the
total costs that will be borne by firms and in terms of their long-run efficiency. Moreover, there are a number of other features that distinguish the relative attractiveness of these options. Further exploration of the potential role of market trading options is long overdue, but in some contexts standards may be preferable, so that it will not always be the case that a particular class of policy options will be dominant.

The same kinds of methodologies that we apply to analyzing conventional pollutants, such as air pollution, can also be applied to analyzing global warming, as well as to more complex externalities, such as the group decisions that lead to overfishing. Examination of these various contexts as well as the policies that have been developed to address them suggest that considerable insight can be obtained by assessing how efficient markets would deal with externalities, if such markets existed.

Questions and Problems

1. Consider the following basic problem regarding a driver and a pedestrian in an accident situation. The driver makes a decision regarding his degree of care, but the pedestrian has no such decision to make. Payoffs to each party are as follows:

<table>
<thead>
<tr>
<th>Driving Speed</th>
<th>Total Benefit to Driver</th>
<th>Expected Cost to Pedestrian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rapidly</td>
<td>170</td>
<td>160</td>
</tr>
<tr>
<td>Moderately</td>
<td>100</td>
<td>40</td>
</tr>
<tr>
<td>Slowly</td>
<td>90</td>
<td>10</td>
</tr>
</tbody>
</table>

Suppose that instead of an anonymous driver-pedestrian relationship we had a two-person society, one driver and one pedestrian.

a. If the driver could undertake voluntary bargains that would be enforceable, what driving speed would result?

b. If both parties have equal bargaining power, what is the predicted settlement amount (that is, amount of transfer from pedestrian to driver)?

2. Suppose that a pulp-and-paper mill discharges water pollutants that impede the value of the stream for swimming. It would cost the mill $5,000 to install pollution abatement equipment to eliminate pollution, and doing so would result in an additional $10,000 in swimming benefits to the residents downstream.

a. If the residents are assigned the property rights, and if each party has equal bargaining power, what will be the predicted outcome and the dollar transfer between the two parties?

b. If the firm is assigned the property right to pollute, what will be the predicted outcome and the income transfer between the two parties?
3. The U.S. Department of Transportation has just rerouted the interstate highway through your yard, so that you now have to sell your house. The government proposes that we compensate you an amount equal to the market value of your home. Is this fair? Is it efficient? Answer the same questions supposing that, instead of the government wishing to purchase your house, I have decided that I want to live in your house. Would it be possible for me to evict you and to pay the market value? Would your answer change if we could accurately determine your reservation price for selling the house so that we would ensure that you would experience no utility loss from such an eviction? How do you believe the functioning of society would change if such a compensation mechanism were instituted?

4. The discussion in the chapter regarding the desirability of taxes and regulatory standards focused primarily on the short-run issues. However, these different policies also have important dynamic implications, particularly regarding the incentives for innovation. Under which type of governmental approach will there be greater incentives to innovate in a beneficial way from the standpoint of decreased environmental and health risks?

5. Suppose that the government must undertake an irreversible policy decision regarding the extent of air pollution regulation. The government is making this decision in a situation of uncertainty, however. In particular, there is some probability \( p \) that the benefits will remain the same as they are this year for all future years, but there is some probability \( 1 - p \) that benefits will be less in all future years. If we take into consideration the multiperiod aspects, should we err on the side of overregulation or underregulation, as compared with what we would do within a single-period choice?

6. Figure 21.11 illustrates a multiperson Prisoner’s Dilemma for a situation in which the payoff curves for the two kinds of cars do not intersect. However, there may be externality situations in which the payoffs do intersect, inasmuch as the desirability of different activities may change in a differential manner for the two different decisions. If these payoff curves intersected, with the bottom payoff curve intersecting the top from below, what would be the nature of the market equilibrium that would prevail? Would this equilibrium be efficient?

7. Environmentalists argue that because the actions we take today will have an irreversible effect on climate change, we should take action now and err on the side of excessive restrictions. Some economists, however, have argued that because of the opportunity to acquire additional information, we should postpone a decision until we learn more about the merits of taking a regulatory action. Which strategy do you find more compelling and why?

8. Class Exercise: A useful class exercise is to develop a contingent valuation survey to determine society’s willingness to pay to preserve some local environmental amenity. Examples include rare species of birds or plants in the area, or freedom from the noise of Jet-Skis at a local lake. What information would you provide respondents about the environmental amenity? What is the payment mechanism that you would establish for people’s willingness to pay? Are there any tests that you can incorporate within the context of your survey to insure its validity, such as transitivity tests or tests for whether people are willing to pay more for broader environmental commodities that should have a larger value?
Product Safety

Although product safety concerns are not entirely new, they did not become a prominent part of the regulatory agenda until after the establishment of the social risk regulation agencies in the early 1970s. A pivotal event that led to the increase in public attention to product safety issues was the publication of Ralph Nader’s *Unsafe At Any Speed*. Nader charged that the automobile industry devoted insufficient resources to product safety, as was evidenced, for example, in the turnover risks posed by the Chevrolet Corvair. This compact, rear-engine car was marketed by Chevrolet in the early 1960s as a moderately priced compact that had some of the driving feel of a sports car. Its main disadvantage was that the car was highly unstable during cornering maneuvers, leading to a rash of deaths to Corvair owners. Among the victims was Ernie Kovacs, who at the time was the host of a popular television comedy show.

Emergence of Product Safety Regulations

The product safety era of the early 1960s was quite different from what it is today. There were no requirements that automobiles include safety belts, and in general they did not. Debates over passive restraint systems and air bags had yet to surface, as the primary concern was whether there ought to be any safety belt requirements at all.

Auto safety was not the only area where new regulations were emerging. In the mid-1960s, Congress instituted requirements pertaining to the hazard warnings that had to be included on cigarette packages. The initial requirements for protective packaging were also instituted in that period so as to make aspirin product and prescription drug packaging child-resistant.

Even the subsequent establishment of the social regulation agencies in the 1970s and the emergence of these new regulations did not lead to the same degree of sensitivity to safety concerns as at the present time. In the 1970s, for example, the Ford Motor Company marketed a subcompact that it called the Ford Pinto. This car was the brainchild of Lee Iacocca, who wished to develop a budget-priced car that would compete with cheap imports. The design of the Pinto was a hurried affair, with catastrophic results. The main safety defect of the car was the placement of the gas tank too near the rear of the car. As a result, the car was highly vulnerable to rear-end collisions. Ford was conscious of the potential risks and the extra $11 per car that would have had to be spent in order to eliminate the hazard, but it chose to stick with the cheaper design. The result was a series of fatal accidents involving the Ford Pinto, which exploded upon rear impact, causing severe burn injuries and deaths. The substantial damages that were ultimately awarded by the courts became part of the increased product liability price tag being imposed on the nation’s businesses.

Current Safety Decisions

Firms currently contemplating product safety decisions no longer look solely toward the market. Rather, their efforts are governed by a complex set of regulations and judicial precedents. In some cases, these regulations are quite specific. The U.S. Department of Transportation regulations for municipal buses are almost tantamount to a comprehensive bus design. The focus of this chapter will be on how such product safety regulations affect the various market participants. We will also address how society should approach product safety regulation to achieve an appropriate balance between the competing objectives.

Figure 22.1 summarizes the main mechanisms at work in influencing product safety. Let us begin with the decisions by the producer. The producer’s environment is governed by three sets of influences: the market through consumer behavior, government regulation, and tort liability. If the market were fully efficient, then there would be no need for social regulation or product liability litigation. In a perfect market, safer products will command a higher price. If consumers are unaware of the risks, however, market outcome will not be ideal. As was indicated in the discussion of the rationale for social regulation, assessing the degree and character of imperfect information is an area in which one should exercise caution. Even if consumers are not fully knowledgeable of all the implications of the product, that fact does not always mean that the market supplies too little safety. Indeed, the opposite result may

Figure 22.1
The Accident-Generating Process
pertain if consumers systematically overestimate the risk, as many consumers do for low-probability events called to their attention. Consequently, one must assess the particular context and nature of the risk involved before one can necessarily conclude that market incentives for safety will be adequate.

**Consumer Complaints**

One form of information that is frequently used as an index of informational failure is the presence of consumer complaints. If consumers file complaints concerning the performance of a product, can we necessarily assume that there is a market failure warranting some form of regulation? To the extent that consumers are filing complaints for products that fail to meet with their reasonable expectations, there is clearly likely to be some informational value to examining the prevalence of complaints. However, in general it is difficult to distinguish the extent to which consumer complaints reflect a market failure or simply consumers who are unlucky. For example, suppose that consumers know that there is a 75 percent chance that new Hyundai cars will run well, but there is a 25 percent chance that these cars turn out to be lemons. Because of the car’s low price, the consumers are willing to take this gamble. However, after the fact the 25 percent of the consumers who get stuck with a lemon are no longer facing an uncertain prospect. Rather, they must confront the certainty of definitely owning a bad car. Conditional upon knowing that they have purchased a lemon, this 25 percent of the consumers may voice regret with their purchase, but on an ex ante basis, before they knew the outcome of the product quality lottery, they may have been making a sound decision from an expected utility standpoint.

One aspect of information provision in which there is more likely to be a clear-cut market failure is with respect to information that is of a public good nature. Firms, for example, will have little incentive to investigate the safety properties of antilock brakes because disclosure of this information will benefit all manufacturers of cars with antilock brakes, not simply the firm undertaking the research.

**Factors Affecting Producer and Consumer Actions**

Based on the impacts of the market, regulation, and tort liability, the producer will choose the products it will make and their characteristics, leading to the product attribute outcome indicated in figure 22.1. Consumers are also making decisions. Based on the information they have received from the media, their experiences, firms, and government regulation, they will make a purchase decision. Moreover, they will also make a product use decision that may be influenced by the incentives created by regulation and tort liability. Government regulations, for example, mandate the use of safety belts in many states and prohibit the use of alcohol.

---

for minors. Similarly, tort liability creates incentives for consumer behavior because people who drive while intoxicated, operate an all-terrain vehicle negligently, and otherwise do not exercise appropriate care can be found guilty of contributory negligence, thus reducing or possibly eliminating their prospective court award after an accident.

**Product Performance and Consumer Actions**

The combined impact of product attributes and the consumer actions result in the risk outcome and financial loss component of figure 22.1. The main issue that has been stressed by economists and generally ignored by the product safety professionals is that product safety is not simply an engineering issue. Individual behavior is relevant both to market decisions and to safety-enhancing actions that individuals may take.

The performance of these products affects the experience base of information that consumers have in making subsequent purchases, which in turn will alter the market environment of the firm.

From the standpoint of regulatory policy, two aspects of these relationships are particularly noteworthy. First, government regulation is not the only economic influence affecting safety incentives. The market and tort liability also are of consequence. Regulation generally affects firms either through design standards that influence the technology or by addressing observed product defects, as in the case of safety recall actions. In contrast, tort liability operates ex post facto. The courts do not address products that are potentially risky but for which there have been no adverse outcomes. The focus instead is on observed defects. People must be injured to collect for bodily injury losses. Consequently, the timing of the institutions in terms of how they can potentially influence product safety is different. Regulations have a greater opportunity to operate in a more anticipatory manner.

The second key feature of figure 22.1 is that safety is the outcome of the joint influence of producers’ safety decisions and user actions. The task of regulatory policy is to ascertain how best to influence both of these determinants of safety outcomes rather than restricting our focus on technological solutions to safety.

**Changing Emphasis of Product Regulation**

There has been a shifting emphasis of the determinants of the product safety environment. The regulatory period of the 1970s concentrated on technological solutions to safety, such as mandated changes in automobile design. The decade of the 1980s marked a shift toward regulating consumer behavior, both through a wave of right-to-know policies and through requirements such as more stringent drunken driving rules and mandatory safety belt use requirements. In addition, there has been a change in the role of tort liability, as there had been an escalation in the role of tort liability through the 1970s, culminating in the explosion in liability insurance premiums in the mid-1980s.
Product safety issues had previously been an afterthought. These matters were in the domain of corporate public affairs offices, which dealt with product safety as part of their general public relations efforts. By the 1990s product safety had become a central corporate concern. Safety and environmental regulations affecting the automobile industry were blamed by some critics as the source of that industry’s collapse. Tort liability awards for workers in the asbestos industry led to the bankruptcy of a major American firm and the elimination of the asbestos industry. Other entire industries have also disappeared because of product safety concerns. The rising liability costs for private planes, which averaged over $100,000 per plane, have led aircraft companies such as Beech, Cessna, and Piper to all but eliminate their production of private airplanes. The manufacturing of diving boards for motel swimming pools is also a vanishing industry. The focus of the remainder of this chapter is on developing the economic tools needed to approach such regulatory issues in a sensible manner.

Premanufacturing Screening: The Case of Pharmaceuticals

Regulations that hit products at a particularly early stage are those that pertain to the premanufacturing screening of products. Firms selling medical devices cannot market these products without prior government approval. Pharmaceuticals, insecticides, and chemical products are all subject to extensive testing and labeling requirements before they can be marketed. Food products are also subject to premarket testing, although these inspection procedures are generally viewed as more lax than the other premanufacturing screening regulations already noted. Notwithstanding food safety regulation, for example, imported produce drenched in pesticides and meat from animals treated with large doses of hormones and antibiotics are staples in the typical American diet.

The most extensively analyzed premanufacturing screening effort is that of pharmaceutical regulation by the Food and Drug Administration (FDA). Before a firm is permitted to market a pharmaceutical product, it must establish the safety and efficacy of that good. Although the regulatory requirements have evolved throughout the century, a pivotal event was the 1962 Kefauver-Harris amendments to the Food, Drug, and Cosmetics Act. The major stimulus for this regulatory regime was the effect of the morning sickness drug thalidomide on pregnant women in England, many of whom had babies with serious deformities caused by the drug. This drug had not been approved for use in the United States, but now, four decades later, it has been revived as an anticancer medication.

Although restrictions on drugs with severe side effects that create an overall net health risk to society are clearly desirable, establishing an appropriate balance in the premanufacturing screening decision is a complicated task.

The principal benefits of more stringent screening pertain to the decreased risk of approving a drug that might have adverse effects. This more stringent screening process also imposes
costs. The first class of costs consists of the testing costs and the forgone opportunity to market a potentially profitable drug. There is an additional cost to society, which may be deprived of potentially beneficial drugs with life-extending properties because they are tied up in the testing process. These costs have been of substantial concern to those with terminal diseases such as AIDS, for which a potentially effective drug that is possibly risky appears to be a good gamble. The regulator’s task is to attempt to balance these competing concerns.

**Weighing the Significance of Side Effects**

Particularly in the case of pharmaceuticals, simplistic alternative policy objectives of eliminating all risks associated with pharmaceutical products are clearly inappropriate. Perhaps the main distinguishing feature of prescription drugs is that they pose potential hazards and, as a result, their use must be closely monitored by a physician. The FDA requires the information pertaining to the drug to be summarized in a label containing hazard warnings. These warnings are reprinted in an annual volume, the *Physicians’ Desk Reference*, distributed to all doctors throughout the United States. Inspection of almost any entry in the *Physicians’ Desk Reference* will indicate the presence of potentially severe adverse effects or complications that may result from prescription drugs. These health impacts range from renal failure to anaphylactic shock. The presence of such potential hazards with prescription drugs does not imply that FDA regulation has been remiss, only that there are inherent risky attributes of the product and that ultimately society must strike a balance between the benefits these drugs provide and the potential hazards they pose.

Recognition of the need for balance does not completely resolve all of the policy issues at stake. The FDA must also decide on the stringency of the testing criteria.

**Drug Approval Strategies**

Table 22.1 summarizes the nature of the trade-off.\(^4\) Suppose that the FDA is analyzing a new drug that is both safe and effective, but the properties are not yet known to it. Ideally it would like to approve such beneficial drugs, but there is the potential that it may reject a beneficial drug because of misleading test results. In addition, firms seeking approval for beneficial drugs may be discouraged by the costs associated with the lengthy approval process and may abandon a drug. Situations in which the FDA review process leads to the rejection of potentially beneficial drugs are designated Type I errors.

If the FDA were to adopt a more lenient drug approval strategy, then it would incur the competing danger of approving dangerous drugs that should not be marketed. Errors of this

---

\(^4\) The role of pharmaceutical regulatory policy with respect to such trade-offs is a principal theme of the book by Henry Grabowski and John Vernon, *The Regulation of Pharmaceuticals: Balancing the Benefits and Risks* (Washington, D.C.: American Enterprise Institute, 1983).
type are designated Type II errors. Ideally, the FDA wants to approve all beneficial drugs and reject drugs that are not safe and effective. Achieving both of these objectives simultaneously is generally not feasible, in large part because information about the drugs obtained through premarket testing is never fully informative. Moreover, because this information is costly to acquire, there are limits to the burdens the FDA can impose. As a result, there is always a need to strike a balance between the Type I and Type II errors.

Although a few critics have charged that the FDA has been too lax, the consensus in the economics literature is that the FDA has placed too great an emphasis on Type II errors. The FDA primarily seeks to avoid approving drugs with potentially adverse consequences, and it places insufficient weight on the Type I error of failing to approve beneficial new drugs. A political factor generating the motivation for this emphasis is the fact that the victims of Type II errors are more readily identifiable than the victims of Type I errors. In the case of Type II errors, specific people will suffer adverse consequences that can be linked to the drug. In contrast, Type I errors generally have a more diffuse probabilistic effect. One percent more of the 3,000 patients suffering from a variant of heart disease may die if a new drug does not appear on the market, leading to a total of thirty expected deaths. However, the particular people who will die because the drug is not available may not be identifiable ex ante. Rather, there is simply a treatment group population that will suffer a probabilistic loss in terms of expected health if the drug is not available.

The identifiability of the parties suffering the adverse consequences becomes quite different when the lobbying group for the new drugs does not consist of a diffuse set of patients who are potentially at risk because of the absence of the drug’s availability, but rather consists of a well-defined group with a clear-cut stake in the accelerated approval of such drugs. One such constituency that emerged in the late 1980s was that of AIDS patients,
who sought more rapid approval of AIDS-related drugs. The result of these efforts was an accelerated drug approval schedule for such drugs. We will consider the impact of this reform effort.

The FDA not only must set the criteria for whether it accepts or rejects a drug, it also must determine the degree of premarket testing that it will require. Because full information is prohibitive and the cost of information acquisition may be substantial, the extent of premarket testing must necessarily be bounded. Figure 22.2 summarizes the shape of the health and non-health costs of premarket testing. As the downward-sloping curve in the diagram indicates, the expected health costs from unsafe or ineffective drugs decline as the extent of premarket testing increases because the FDA is better able to avoid approving drugs that will turn out to have adverse consequences. However, minimization of health costs is not our sole objective, since testing also imposes costs. The cost of the research and development and of the lost market opportunities stemming from delay in the drug’s approval is indicated by the upward-sloping curve in the diagram. The sum of the cost to the firms and the health costs leads to the total cost function at the top of the diagram. For total cost levels associated with the amount of premarket testing indicated by $t^*$, costs are minimized, establishing this amount of testing as the optimal amount.
Accelerated Drug Approval Process

Although economists have long urged that drug approval time be accelerated, the FDA has done little to change its overall policy. However, in 1987 the FDA instituted an accelerated drug approval process for drugs that address life-threatening diseases such as AIDS. Table 22.2 summarizes the patterns of drug approvals and the time for approval for different classes of drugs. Consider first the rate of approval of drugs that represent legitimate innovations, in particular new chemical entities (NCEs). The number of such approvals ranges from eleven to twenty-six per year for the 1980–1989 period. The average approval time for such drugs is on the order of approximately three years.

Now let us consider a subgroup of these drugs designated as 1AA/1A. These drugs consist of those that are targeted for fighting diseases such as AIDS, as well as drugs that the FDA believes will have a substantial impact. Following the institution of the accelerated drug approval policy in 1987, there was an apparent increase in the rate of such drug approvals. The lag time for approval in 1988 was quite high even for these drugs, but this substantial lag time may reflect the fact that many of these drugs had been in the FDA drug approval pipeline for years prior to the new policy. Beginning in 1989, there was a substantial speed-up in the approval time for the 1AA and 1A drugs. One of the main factors leading to this apparent shift in FDA policy is the fact that accelerating the approval time of drugs that address life-threatening diseases not only has a well-defined constituency, but there is also less potential adverse health risk because of the high risk of mortality. Although there is a widespread consensus that this shift in policy was attractive, most FDA decisions are not as

Table 22.2
New Drug Approvals and Time to Approval

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of NCE* Approvals</th>
<th>Avg. Lag Time (months) from Submission to Approval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All Drugs</td>
<td>1AA/1A Drugs</td>
</tr>
<tr>
<td>1980</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>1981</td>
<td>23</td>
<td>2</td>
</tr>
<tr>
<td>1982</td>
<td>22</td>
<td>4</td>
</tr>
<tr>
<td>1983</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>1984</td>
<td>21</td>
<td>2</td>
</tr>
<tr>
<td>1985</td>
<td>26</td>
<td>3</td>
</tr>
<tr>
<td>1986</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>1987</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>1988</td>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td>1989</td>
<td>21</td>
<td>5</td>
</tr>
</tbody>
</table>

* NCE = new chemical entities.

Source: All figures based on calculations by the authors using chronology of new chemical entities developed by the University of Rochester Center for Study of Drug Development, July 10, 1990 (computer printout).
clear-cut. Even if officials are willing to commit to a particular trade-off, there is a major difficulty in ascertaining what these trade-offs are. In terms of figure 22.2, the cost to the firms can be estimated reasonably reliably, but there is often substantial uncertainty regarding the expected health costs to the population. In situations in which we do not know the health curve costs, FDA officials may be held responsible for their judgments regarding the entire shape of this curve if they adopt an aggressive drug approval policy. The incentives for bureaucratic risk aversion are clear.

The Behavioral Response to Product Safety Regulation

Seat belt usage has played a prominent role in the regulation literature. Use of automobile safety belts reduces the fatality risk to the passenger who buckles up, but entails costs of time and discomfort. Based on these trade-offs, Glenn Blomquist estimated the implicit value of statistical life implied by patterns of seat belt usage was $1.0 million. These estimates, however, only reflect risks to the passengers. Safety belts also raise potential externality concerns as well.

The usual approach to product safety regulation has been to alter the technology of the product in a safety-enhancing manner. If the behavior of the users of the product remains unchanged after the mandated safety device is instituted, then we will reap the benefits of the engineering controls. In the case of automobile safety, for example, engineering experts generated a variety of predictions of substantial gains in safety that would result from the wave of initial safety regulations. These experts predicted a 0.5 percent reduction in occupant death rates from dual braking systems, a 0 to 2.5 percent reduction in occupant death rates from improved windshields, a 4 to 6.5 percent reduction in occupant death rates from an energy-absorbing steering column, a 7 to 16 percent reduction in occupant death rates from lap seat belts, and a 0.25 to 1 percent reduction in occupant death rates from shoulder belts. The benefits derived from these improvements all hinge on a key assumption, which is that the behavior of the driver will remain unchanged.

In an influential economic analysis, Sam Peltzman hypothesized that driver behavior would not remain unaffected. This theoretical hypothesis in turn has led to extensive empirical work

5. This value is in $2,000 and is derived from estimates in Glenn Blomquist, “Value of Life Saving: Implications from Consumption Activity,” *Journal of Political Economy* 87, no. 3 (1979): 540–58. A large literature has followed, yielding values of statistical life from $2 million to $7 million.

by Peltzman and others, including Glenn Blomquist. In the case of safety belts, for example, the safety improvement from the new technology would reduce the potential hazards to the driver of driving fast. As a result, the relative benefits of taking the safety precaution of driving slowly would decline once an individual had buckled up. Faster speeds would become more desirable.

Figure 22.3 illustrates Peltzman’s reasoning diagrammatically. Suppose that initially the driver is at point A, where the line 0A gives the relationship between driving intensity and the driver’s risk of death before regulation. With the use of safety belts, the risk curve drops

Figure 22.3
Relationship of Driving Intensity to the Regulatory Regime

---

to $0BC$. After the introduction of the safety devices, if the driver leaves his degree of precautions unchanged he will be at a death risk at point $B$. However, because the marginal benefits to the driver of taking the precaution have been reduced, he will increase his driving intensity to a point such as $C$, thus muting some of the effect of the safety belts.

The factors driving the movement from point $B$ to point $C$ are based on elementary economic principles. Figure 22.4 indicates the marginal benefits and marginal costs to the driver of driving slowly. For simplicity, suppose that the marginal cost of driving slowly has a constant value per unit time, leading to the flat marginal cost curve. Originally the driver faced a marginal benefit curve of $MB_0$ for driving slowly. However, after the introduction of the safety device—in this case, auto safety belts—the marginal benefit of driving slowly has been reduced, assuming that the belt is worn. As a result, the optimal slowness of the driving speed has been reduced, which is to say that the driver now finds it desirable to drive faster once he is using devices that will decrease his risk of injury or property damage from an accident.
Consumer’s Potential for Muting Safety Device Benefits

The overall economic story is consequently that once individuals buckle up, they will have an incentive to drive faster, thus muting and possibly offsetting the beneficial effects of the safety device. This line of argument has long aroused considerable controversy because of the surprising nature of the results, as well as the extent of the empirical effect that has been claimed for it.

The underlying theory is quite sound and is based on the same kinds of marginal benefit and marginal cost reasoning that is fundamental to all of economics. If one recast the safety belt issue in a somewhat different context, then most individuals would accept the economic mechanisms at work. Suppose that instead of making the car safer by introducing safety belts, we make driving riskier by making the streets icy. Few would question that it is optimal for people to drive with greater care and slower speeds when the roads are icy and slick than when they are dry. Once the ice melts and the streets return to their dry conditions, one would expect people to drive faster. In essence, what safety belts do is take us from a risky regime such as that of icy streets to a safer regime such as dry streets, and the benefits from exercising driver care will decline.

How much of a decline will occur has long been a matter of dispute. In Peltzman’s original paper, he did not claim on theoretical grounds that the effect of safety belts would necessarily be counterproductive, although it could be. However, in his empirical analysis of both time-series and cross-sectional data pertaining to motor vehicle death rates, he was unable to find any statistically significant effect of the introduction of safety belts. He concluded that the behavioral response effect offset the technological gains.

This issue has become an ongoing controversy in the automobile safety literature. The general consensus appears to be the following. First, automobile safety regulations have reduced the risks to drivers and motor vehicle occupants. However, there is also evidence that drivers wearing safety belts do drive faster, inasmuch as there has been an increase in the fatalities of motorcyclists and pedestrians with increased safety belt utilization. Thus the overall mechanism described here has strong support.

There is less general agreement on the extent of the offset from the deaths of motorcyclists and pedestrians. Although there remains some adherence to the view that these effects completely offset the beneficial impacts of automobile safety regulation, the mainstream view is that on balance safety regulation has a risk-reducing effect, although there is a muting of the impact of safety regulations by the decrease in the care exercised by drivers.

This role of individual responses to regulations is not restricted to safety belt issues. In particular, other product safety regulations that rely on changes in the technology similarly

7. These overall themes are articulated in Crandall, et al., *Regulating the Automobile.*
will interact with individual usage of the product to govern the ultimate product safety that will be experienced by the consumer.

**The “Lulling Effect”**

A case that is particularly intriguing is that of safety caps. The government has imposed child-resistant safety cap requirements for two decades for products such as aspirin, prescription drugs, and selected other hazardous products, such as antifreeze. Safety caps not only reduce the benefits to parents of putting medicines in a location for which access by children is difficult, but also may give parents a false sense of security. In a phenomenon that Viscusi terms the *lulling effect*, there may be a misperception on the part of consumers of the efficacy of the safety device, leading to an additional decline in safety precautions.8 Consumer product safety commissioners and the public at large routinely refer to these caps as being childproof, whereas in fact they are not.

Figure 22.5 indicates the impact of these various influences. Suppose that before the advent of safety caps, the expected loss suffered by the consumer from any given level of safety effort is given by the curve $EL_0$. Suppose that the consumer originally was at point $A$, which reflects the optimal amount of expected loss that the consumer is willing to incur, given the costs associated with undertaking safety efforts. After the advent of the safety caps, the expected loss associated with any degree of safety precautions declines to $EL_1$. If the consumer does not change his precautions, the postregulation safety effort will be at point $B$. For the usual economic preferences, one will necessarily decrease the level of safety-related effort, so that the individual will be to the left of point $B$. If this decrease in safety precautions is sufficient, there will be no effect of the safety device on the safety outcome, and we will be at point $C$. For the safety device to be counterproductive, leading to an outcome such as point $F$, one must impose very severe restrictions on the shape of individual preferences. Such a counterproductive effect of regulations is conceivable, but it requires that very special and unusual assumptions be met.

**Effect of Consumer’s Perception of Safety Device Efficacy**

However, matters are quite different if consumers do not accurately perceive the efficacy of the safety device. If they believe the safety mechanism reduces the perceived risk from the curve $EL_1$ to $EL_2$, then we may end up at a counterproductive outcome such as point $D$ much more easily. Consumers believe they are at point $D$, whereas they are actually at point $F$. The danger of safety mechanisms is not simply that of a falloff in the optimal level of consumer behavior but also an inducement to misperceptions regarding the risk, leading to a further drop-off in safety precautions.

In the safety-cap case, there is detailed evidence regarding the character of consumer precautions before and after the caps went into effect. Poisonings from safety-cap bottles are the main source of poisonings from aspirin and analgesic products. In most of these instances, the bottles have been left open by consumers. The rash of open-bottle poisonings is not surprising, inasmuch as there have been widespread complaints regarding the difficulty of grappling with the caps. In an effort to deal with these caps, many consumers have responded by simply leaving the caps off altogether.

The poisoning context also provides some intriguing evidence regarding possible spillover effects for other products. If consumers undertake a common safety precaution for their medicines, then the introduction of safety caps may lead to a decrease in safety precautions

---

9. See Viscusi, 1992, for empirical documentation of the effects discussed here.
overall. In the case of products with safety caps, the net effect of the decrease in precautions may be simply to have no observable effect on safety, which has in fact been the case for aspirin. For products not covered by safety caps, the decrease in the overall level of precaution taking will increase the risk from these products. Such an increase in risk has in fact been observed, as there has been an observed adverse spillover effect on other products. After the introduction of safety caps for aspirin bottles, the poisoning rates for analgesic products such as Tylenol escalated from 1.1 per 1,000 in 1971 to 1.5 per 1,000 in 1980. Taking into account the rise in the sales of Tylenol and related products accounts for only half of this increase. The overall implication of this analysis is that there have been 3,500 additional poisonings annually of children under five that resulted from the decreased safety precautions after the advent of safety caps.

The presence of such behavioral responses does not imply that all government regulations are bad or that these particular regulations are necessarily ill-conceived. What the responses do suggest, however, is that one cannot view safety as simply being a matter of engineering controls. Individual behavior plays a key role. These responses are not restricted to safety caps and seat belts. Empirical evidence suggests that the newly introduced safety mechanism for butane cigarette lighters also reduces parental care. Regulations that attempt to influence safety behavior through hazard-warning programs, safety training efforts, and other efforts should be regarded as a central component of any product safety regulatory strategy.

The Costs of Product Safety Regulation: The Automobile Industry Case

A principal target of product safety regulation has been automobiles. There have been dozens of regulations affecting the safety of autos, as well as their environmental impact from emissions. Much of the impact of these regulations occurred in the 1970s—the decade in which the major wave of auto regulations emerged.

Let us begin with the various regulations affecting auto safety. These include occupant protection requirements, steering column protection, seat belt assemblies, side door strength, bumper requirements, fuel system integrity standards, and a variety of other specific safety standards. There will also be a fuel penalty due to the added weight that will be imposed on the car because of these safety devices. The standard that requires a bumper to withstand low-speed crashes is a chief contributor to this greater weight. The total costs are on the order of over $1,000 per car.

These costs are in addition to the costs imposed by emissions standards. Table 22.3 summarizes these costs associated with compliance with environmental regulations pertaining to automobiles. As in the case of safety standards, the overall price tag by the 1980s had reached over $1,000 per car.
What we see, then, is a situation in which both safety and environmental regulations were imposing substantial and increased costs on the automobile industry in the 1970s. This decade was also a period of dramatic change because of the inroads being made by foreign imports of small cars as a result of the dramatically higher gasoline prices in the late 1970s. U.S. sales of Toyota, Honda, and Nissan (then known as Datsun) soared in the 1970s. Some critics charged that government regulation was undermining the previously superior position of the U.S. automobile industry by imposing a required shift in the technology for automobiles, thus making much of the U.S. production system and U.S. design of automobiles obsolete. With the sunk costs in the earlier designs no longer being useful, U.S. producers sacrificed much of their previous advantage over foreign competitors, thus making it easier for foreign firms to compete in American markets.

Even for political observers who did not blame government regulation as playing a central role in the demise of the automobile industry, the relationship between the health of the industry and government regulation was an important concern. The Carter administration, for example, undertook a substantial financial bailout of the Chrysler corporation in an effort to avoid its bankruptcy. In addition, it initiated some modest efforts designed to target regulatory relief at that industry.

On taking office, the Reagan administration instituted a sweeping program of regulatory relief for the automobile industry. Table 22.4 summarizes the components of this effort. The diversity of the regulations included in the auto reform package is quite impressive. Government regulations pertain to almost every aspect of the design of automobiles, ranging from speedometer standards to emissions requirements. Moreover, the price tags associated with many of these regulations are on the order of hundreds of millions of dollars. What is perhaps most impressive is the broad range of the twenty-nine regulations listed in table 22.4 and the fact that a single industry could account for so much regulatory activity.

### Table 22.3
Estimated Cost per Car of Meeting the Automotive Emissions Standards, Pre-1968–1981

<table>
<thead>
<tr>
<th>Model Year</th>
<th>Current Costs (1981 dollars)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-1968</td>
<td>0</td>
</tr>
<tr>
<td>1968–1969</td>
<td>30</td>
</tr>
<tr>
<td>1970–1971</td>
<td>50</td>
</tr>
<tr>
<td>1972</td>
<td>370</td>
</tr>
<tr>
<td>1973–1974</td>
<td>950</td>
</tr>
<tr>
<td>1975–1976</td>
<td>640</td>
</tr>
<tr>
<td>1977–1979</td>
<td>700</td>
</tr>
<tr>
<td>1980</td>
<td>1,000</td>
</tr>
<tr>
<td>1981</td>
<td>1,400</td>
</tr>
</tbody>
</table>

Table 22.4
The Reagan Administration’s Auto Reform Package

<table>
<thead>
<tr>
<th>Issue</th>
<th>Action (Date of Completion)</th>
<th>Five-Year Savings (millions)</th>
<th>Industry</th>
<th>Public</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gas-tank vapors</td>
<td>Declined to order new controls on cars (April 1981).</td>
<td></td>
<td>$103</td>
<td>$1,300</td>
</tr>
<tr>
<td></td>
<td>Raised allowable “failure rate” for test of light trucks and heavy-duty engines from 10 to 40 percent (Jan. 1983).</td>
<td></td>
<td>19</td>
<td>129</td>
</tr>
<tr>
<td></td>
<td>Reduced spot checks of emissions of vehicles on assembly lines by 42 percent; delayed assembly-line tests of heavy-duty trucks until 1986 (Jan. 1983).</td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>High-altitude autos</td>
<td>Ended assembly-line tests at high altitude, relying instead on industry data (April 1981).</td>
<td>0.2</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Allowed industry to self-certify vehicles as meeting high-altitude emission standards (April 1981).</td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Pollution waivers</td>
<td>Consolidated industry applications for temporary exemptions from tougher emissions standards for nitrogen oxide and carbon monoxide (Sept. 1981).</td>
<td></td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Test vehicles</td>
<td>Cut paperwork required to exempt prototype vehicles from environmental standards (July 1982).</td>
<td></td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Driver vision</td>
<td>Scrapped existing 1981 rule and second proposed rule setting standards for driver’s field of view (June 1982).</td>
<td>160</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Speedometers</td>
<td>Revoked rule setting standards for speedometers and tamper-resistant odometers (Feb. 1982).</td>
<td></td>
<td>—</td>
<td>20</td>
</tr>
<tr>
<td>Tire rims</td>
<td>Scrapped proposal to set safety standards for explosive multipiece tire rims (Feb. 1982).</td>
<td>300</td>
<td>75</td>
<td>—</td>
</tr>
<tr>
<td>Brake tests</td>
<td>Eased from 30 to 20 percent the steepness of grades on which post-1984 truck and bus brakes must hold (Dec. 1981).</td>
<td></td>
<td>—</td>
<td>1.8</td>
</tr>
<tr>
<td>Tire pressure</td>
<td>Scrapped proposal to equip vehicles with low-tire pressure indicators (Aug. 1981).</td>
<td></td>
<td>—</td>
<td>130</td>
</tr>
<tr>
<td>Battery safety</td>
<td>Scrapped proposal to set standards to prevent auto battery explosions (Aug. 1981).</td>
<td></td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Tire safety</td>
<td>Revoked requirement that consumers be told of reserve load capacity of tires; eased tire makers’ reporting requirements (June 1982).</td>
<td></td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Antitheft protection</td>
<td>Eased antitheft and locking steering wheel standards for open-body vehicles (June 1981).</td>
<td></td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>
### Table 22.4 (continued)

<table>
<thead>
<tr>
<th>Issue</th>
<th>Action (Date of Completion)</th>
<th>Five-Year Savings (millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Industry</td>
</tr>
<tr>
<td>Fuel economy</td>
<td>Streamlined semiannual reports of auto makers on their progress in meeting fuel economy goals (Aug. 1982).</td>
<td>—</td>
</tr>
<tr>
<td>Tire ratings</td>
<td>Suspended rule requiring industry to rate tires according to tread wear, traction, and heat resistance (Feb. 1983).</td>
<td>—</td>
</tr>
<tr>
<td>Vehicle IDs</td>
<td>Downgraded from standard to administrative rule the requirement that all vehicles have ID numbers as an aid to police (May 1983).</td>
<td>—</td>
</tr>
<tr>
<td>Seat belt comfort</td>
<td>Scrapped proposal to set standards for seat belt comfort and convenience (June 1983).</td>
<td>—</td>
</tr>
<tr>
<td><strong>Rules with Uncertain Futures</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-altitude emissions</td>
<td>Failed to revise Clean Air Act order ending weaker high-altitude emissions standards in 1984; eased through regulatory changes.</td>
<td>38</td>
</tr>
<tr>
<td>Emissions reductions</td>
<td>Failed to revise Clean Air Act order to cut large trucks’ hydrocarbon and carbon monoxide emissions by 90 percent by 1984; standard was delayed until 1985.</td>
<td>105</td>
</tr>
<tr>
<td></td>
<td>Failed to ease Clean Air Act order reducing nitrogen oxide emissions from light trucks and heavy-duty engines by 75 percent by 1984. Regulatory changes under study.</td>
<td>150</td>
</tr>
<tr>
<td>Particulate pollution</td>
<td>Delayed a proposal to scrap specific particulate standards for some diesels in favor of an average standard for all diesels. Stiffer standards delayed from 1985 to 1987.</td>
<td>40</td>
</tr>
<tr>
<td>Methane standards</td>
<td>Shelved because of “serious” costs; questions a plan to drop methane as a regulated hydrocarbon.</td>
<td>—</td>
</tr>
<tr>
<td>Passive restraints</td>
<td>Delayed and then revoked requirement that post-1982 autos be equipped with passive restraints; revocation overturned by Supreme Court in June 1983.</td>
<td>428</td>
</tr>
<tr>
<td>Bumper damage</td>
<td>Cut from 5 to 2.5 mph the speed at which bumpers must resist damage; change is on appeal.</td>
<td>—</td>
</tr>
</tbody>
</table>

The automobile industry will remain a principal target of product safety regulation. The purpose of these efforts is not to drive a leading American industry out of business but rather to address the main source of product safety problems. Motor vehicles account for over half of all accidental deaths in the United States, and it is inevitable that continued regulation of automobiles and other motorized vehicles will remain a prominent policy concern.

In addition to being highly regulated, automobiles are the target of substantial litigation. Nevertheless, one should not lose sight of the fact that market forces also play a substantial role in providing incentives for automobile companies to provide cars with characteristics in line with consumer preferences. For example, the prices of used cars are higher for cars that are more powerful, have a higher resale value retention rate, have a better maintenance rating, have automatic transmissions, and have similar value characteristics. Moreover, in their used car purchases, consumers pay a higher price for safer cars, as reflected in the accident history for these model lines, with the implicit value per life saved ranging from $3.8 million to $5.4 million, estimates that are not out of line with the labor market estimates discussed in chapter 20.10

Trends in Motor Vehicle and Home Accident Deaths

The purpose of regulation of home and automobile safety is to produce improvements in the accident rate. Focusing on these two classes of injuries can yield substantial dividends. In 2002, motor vehicle accidents accounted for 41 percent of all accidental deaths, and home accidents other than motor vehicle accidents accounted for an additional 33 percent. The contribution of these accident groups to disabling injuries is also quite substantial, although the emphasis is reversed, with home accidents accounting for more disabling injuries. The remainder of the accidents that are not due to either motor vehicle or home accidents stem from accidents that occur at work or in public places, such as falls in public places, deaths from firearms, and crashes involving planes and trains.

The administrators of the regulatory agencies responsible for auto safety and home accidents generally refer to the improvements in the trends for these accidents as evidence of the efficacy of the agency. Annual press releases announcing decreases in accident trends portray these declines as evidence of the agency’s success. As will be noted in chapter 23, this approach has also been used by Occupational Safety and Health Administration (OSHA) administrators in defending the accomplishments of their agency.

Accident Rate Influences

Such a test does little to show that regulation has had a demonstrable effect on safety. Accident rates have declined steadily for the past 70 years for which ••. Most important is that because of the greater affluence of society, consumers demanded greater safety from their products. This wealth effect alone should continue to lead to safety improvements.

These safety improvements are in evidence in auto accident rate trends provided that one defines the risk measure properly. This definitional issue arises most particularly with respect to motor vehicles. Two types of variations are most important. The first pertains to the intensity of usage of motor vehicles. People drive cars more often today than they did fifty years ago, so that we cannot simply compare death rates across the population but must take into account the intensity of the product’s use. One mechanism for doing so is to look at the automobile accident rate on a mileage basis rather than a population basis. Doing so yields a quite different picture of the trend in automobile safety. Automobile death rates have declined from 21.65 per 100 million vehicle miles in 1923 to 1.56 per 100 million motor-vehicle miles in 2002, whereas the accident rate on a population basis has been quite stable, declining from 16.5 deaths per 100,000 population in 1923 to 15.7 deaths per 100,000 population in 2002.

Temporary shifts in the age structure of the population also may influence the accident rate. The rise in the proportion of teenage drivers in particular eras, for example, has also contributed to temporary swings in the motor vehicle accident rate. Changes in the character of highways and the driving speed on these highways also affect motor vehicle accident rates, even though the safety of the car itself may not have changed. In general, one should take all of these various factors into account when analyzing safety trends.

The Decline of Accident Rates

Since the 1930s, motor vehicle accident death rates per 100 million miles driven have declined quite steadily. The decline preceded the advent of government regulation just over three decades ago. While there was some flattening of the drop-off in motor vehicle accident rates in the 1960s, that decade also marked an increase in the total motor vehicle death risk arising from the changing age structure in the population and increased driving due to the growth of the interstate highway system.

The decline in home accident rates has been steady since the 1940s. Once again, the advent of a decline in the accident rate preceded the establishment of the Consumer Product Safety Commission. Statistical studies of the Consumer Product Safety Commission have failed to indicate any statistically significant impact of these efforts on product safety.11 These studies

do not imply that no regulations of that agency have ever been effective, only that their impact has been sufficiently small that their influence is not evident in examination of national accident statistics.

The econometric studies of motor vehicle accidents have yielded somewhat more optimistic results. Almost all of these studies have indicated that there has been an acceleration in the rate of decline in motor vehicle accident death rates in the 1970s and 1980s, and controlling for other factors indicates that much of this decline is due to the impact of safety regulations. Much more controversial has been the assessment of the overall impact of motor vehicle safety regulation taking into account the spillover effects on the deaths of pedestrians and motorcyclists, which offset at least partially the favorable effect of safety regulations on motor vehicle accident rates.

From an economic standpoint, there is no reason why government regulations should not be effective, provided that the offset from the decrease in consumer precautions is not too great. The efficacy of these regulations in terms of promoting safety is not so limited that there should not be some beneficial effect observed. The main surprise from the earliest studies of these regulations, indicating the absence of a demonstrable effect, stems primarily from the imbalance between the initial projected impacts of these agencies and their observed impacts. This disparity suggests that other economic behavior such as users’ precautions should also be taken into account.

The more important question from the standpoint of long-run regulatory policy is ascertaining whether on balance these regulations are in society’s best interests. By the early 1980s, the price tag for safety in emissions regulations had reached over $2,000 per car, but there were observable benefits as well. Econometric estimates indicated that the automobile death risk would be as much as 40 percent greater in the absence of such safety regulations. By some calculations, these regulations also produced benefits in excess of their costs, in large part because the safety regulations did not impose stringent deadlines for the adoption of specific technologies, but rather proceeded on an incremental basis in which there was a gradual development of the cluster of safety standards that reflected an evolving knowledge of the changing safety technologies.12

The Rise of Product Liability

Direct government regulation of product safety is not the only influence on firms’ safety decisions. Increasingly product liability awards by the courts have played an important role in establishing safety incentives.13 The rise in product liability costs is reflected in the trends in

12. See Crandall et al., Regulating the Automobile.
13. The following discussion is based most directly on W. Kip Viscusi, Reforming Products Liability (Cambridge, Mass.: Harvard University Press, 1991). Other treatments of these issues include Robert E. Litan and Clifford...
In the 1960s the total insurance premiums that were paid for general liability insurance such as product liability coverage were only $746 million. In the last forty years, however, there has been a dramatic expansion in the role of product liability. Changes have included shifts in the criteria used to assign liability to corporations. The earlier negligence doctrine has been replaced by a strict liability doctrine that requires companies to bear the costs of product injuries in a greater share of situations. In addition, there has been a tremendous expansion in hazard-warnings cases and in the concept of what constitutes a product design defect.

The net effect of these changes was to increase product liability premiums to $1.7 billion in 1970, and they rose even further to $6.6 billion in 1980. The greatest expansion occurred in the mid-1980s, as liability premiums jumped to $11.5 billion in 1985, reaching $19.1 billion in 1988. Since that time, costs have been fairly stable, as premiums were $19.9 billion in 2000. Even these impressive costs do not capture the full liability cost to firms. Corporations also must pay for the cost of extensive legal staffs. Moreover, there may be costs of liability judgments that are not covered by insurance, such as punitive damage awards and awards in excess of the policy limits. Many industries such as the pharmaceutical industry are unable to receive any liability insurance coverage at reasonable rates from conventional insurers and as a result have established separate insurance mechanisms outside of the standard industry channels. These insurance costs are in addition to the amounts in table 22.5.

The impact of this product liability revolution on businesses has been substantial. Pharmaceutical companies have responded by withdrawing vaccines that have been hard hit by

---

Table 22.5
Trends in General Liability Insurance Premiums

<table>
<thead>
<tr>
<th>Year</th>
<th>Premiums ($ millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1960</td>
<td>746</td>
</tr>
<tr>
<td>1970</td>
<td>1,658</td>
</tr>
<tr>
<td>1980</td>
<td>6,612</td>
</tr>
<tr>
<td>1985</td>
<td>11,544</td>
</tr>
<tr>
<td>1988</td>
<td>19,077</td>
</tr>
<tr>
<td>1990</td>
<td>18,123</td>
</tr>
<tr>
<td>1995</td>
<td>18,582</td>
</tr>
<tr>
<td>2000</td>
<td>19,917</td>
</tr>
</tbody>
</table>


---

liability costs. A National Academy of Sciences panel blamed the lagging research and development of contraceptive devices by United States firms on the rise in liability costs.\textsuperscript{14} The domestic aircraft industry manufacturing planes for private use has all but disappeared. Between 15 and 25 percent of the purchase price of ladders goes to pay for liability costs, and 17 cents of every fare dollar on the Philadelphia mass transit system goes to pay for liability insurance expenses.

One of the seminal legal events that led to this product liability revolution was the emergence of the strict liability doctrine that replaced the earlier negligence criteria. Under a negligence test, a firm is required to provide the efficient level of product safety. Consider the statistics in table 22.6, which gives the benefits and costs of different levels of product safety. The objective is to set the level of product safety so as to minimize total social costs. This optimal point is achieved at the medium degree of safety, which offers a net social payoff of $50,000. Higher and lower levels of safety each provide fewer social rewards.

### The Negligence Standard

Under a negligence regime, a firm is liable if it does not meet the medium degree-of-safety standard. Thus, all firms providing a low level of safety become liable for the $50,000 in accident costs that could have been prevented if they had provided the efficient degree of safety at the medium level of safety. Firms facing this penalty consequently must choose between providing a low level of safety, which will cost them $25,000 from a manufacturing standpoint and $150,000 from a legal liability standpoint, or choosing a higher level of safety such as the medium level of safety that will impose $50,000 of manufacturing costs but no liability costs. A negligence standard such as this will create incentives for firms to choose the medium level of safety, which in this case is the socially efficient level. Once they have done so, firms are free of any possible liability burden because they have met the efficient standard of care.

The Strict Liability Standard

In contrast, under a strict liability standard, a firm is liable for all of the accident costs incurred by the consumer irrespective of whether the firm has met the appropriate level of safety. (Actually, this is a variant of strict liability known as absolute liability, but consideration of this extreme case facilitates the exposition.) In the case of the high degree of safety, the cost to the company is the same as the social cost, which is $140,000. At the medium level of product safety, the social cost that must be borne by the company under strict liability is $100,000, and at the low level of product safety, the social cost that must now be internalized by the company is $175,000. The company will choose the level of product safety that minimizes these social costs—the medium level of product safety.

Strict liability achieves this outcome almost by definition. Since strict liability requires that the company bear all of the product-related costs associated with accidents, in effect what this doctrine does is force the company to internalize all accident costs. The social objective function and the company’s profit function consequently become one and the same.

Tracing Accident Costs and Causes

A danger enters with respect to our inability to distinguish which are the accident costs traceable to the product. In addition, problems of moral hazard also arise. The result may be that entire product markets may disappear. If all ski manufacturers were required to pay for the hospital bills of those injured while using skis, the price of skis would become exorbitant. Similarly, the prices of automobiles would escalate if companies had to pay for all of the accident costs resulting from automobiles, irrespective of the parties at fault and the behavior of the drivers.

If we abstract from such complications, from the standpoint of achieving an economically efficient outcome, both the negligence rule and the strict liability rule are equivalent. Each leads to the medium level of safety. The difference is that under strict liability companies pay a share of the accident costs in a much broader set of instances. As a consequence, the legal system’s movement toward a strict liability doctrine shifted the balance of power in the courts toward the consumers, who are consequently able to collect from companies in a larger share of cases than they were before.

The Ford Pinto Case

An instructive example of how the negligence standard could be applied is with respect to the calculations for the Ford Pinto’s gas tank safety. Recall that Ford’s decision to place the gas tank at the rear of the vehicle saved costs but led to burn injury and deaths upon rear impact. Moving the gas tank would save lives on an expected basis but would raise the price of the vehicle.
Table 22.7 summarizes the calculations done by Ford with respect to this safety risk. The magazine *Mother Jones* won a Pulitzer Prize for its exposé of these calculations, though apparently the estimates were prepared not just for the Pinto but with respect to proposed auto safety regulations more generally. In Ford’s defense, it should also be noted that General Motors has undertaken similar analyses with respect to side-impact fires.

The cost estimates at the bottom of table 22.7 indicate that moving the gas tank to reduce the risk of fires is not inordinately costly, only $11 per vehicle. However, when this cost is imposed on an entire vehicle fleet, the cost total is nontrivial—$137.5 million.

The quite appropriate and sensible economic question to ask is whether the benefits exceed the costs. As indicated by the calculations in the top panel, the benefit levels fall short of what is needed to make the safety improvement desirable. In preparing these estimates, Ford estimated the value of burn deaths at $200,000, which was in fact the average court award for burn deaths in product liability cases at that time. Ford’s conclusion that the benefits of the safety improvement did not exceed the costs was not persuasive to jurors, who levied both compensatory and punitive damages in Ford Pinto cases.

Did the company err in its analysis, or was it the victim of a runaway jury that failed to understand economic reasoning? The key flaw in Ford’s analysis was using court awards to value the prevention of burn injuries and deaths. These values primarily capture income losses and medical costs, and understate the value of preventing such serious outcomes. To value prevention of burn deaths, one should use values of statistical life, not court awards. For that earlier era the estimates of the value of statistical life were in the $3 million range rather than the $7 million value that now prevails. Use of such an appropriate benefit measure boosts the
burn death value to $540 million, which alone is sufficient to justify moving the gas tank to a safer location. Whether the context is government risk policies, corporate decisions, or personal actions, the appropriate method for attaching a benefit value to reduced risks to life is the value of statistical life number. Focusing on the incentives provided by court awards alone will lead to undervaluation of the risk.

**Escalation of Damages**

There has also been a rapid escalation in the role of damages. The penalties levied by regulatory agencies are often quite modest—on the order of $1,000 or less, and only a few million dollars even in the most severe cases. In contrast, million-dollar awards in product liability cases are routine. Newspapers throughout the country gave prominent coverage to the woman who spilled a cup of hot McDonald’s coffee on her lap and suffered burns for which she received a several-million-dollar award (later reduced by a higher court).

In addition to the phenomenon of runaway juries, there also may be a reasonable basis for some large liability awards. For a consumer who loses twenty-five years of his or her work life at a rate of pay of $40,000 per year, the lifetime earnings loss is $1 million. Because the size of the award roughly doubles when one also takes into account the pain and suffering associated with the accident, as well as the loss to the family associated with such injuries, one can see how severe injury awards on the order of $1 million or more can become routine rather than exceptional.

While significant awards of $1 million or more may have merit, the blockbuster awards have generated increasing controversy. The greatest level of stakes is with respect to punitive damages. In the past two decades there have been sixty-four punitive damages awards of at least $100 million. All but three of these awards have been the result of jury trials rather than bench trials, leading some observers to propose that judges be given greater control over the level of punitive awards. Even though many punitive damages awards are reduced on appeal, the high stakes involved mean that the entire company’s survival is often on the line in these cases.

This escalation in awards has led to a variety of product liability reform efforts designed to limit the role of product liability damages. A chief target for these efforts has been the awards for pain and suffering because of the absence of well-defined legal criteria for determining such damages. Some lawyers, such as Melvin Belli, suggest that jurors should ascertain the value of pain and suffering for a small time interval such as a second and then scale it up by the length of time the pain and suffering was endured. In the case of very lengthy injuries this procedure could produce astronomical pain-and-suffering awards. A useful exercise is to consider whether from the standpoint of good economics one should simply multiply the value of pain and suffering for small time intervals by the amount of time the pain and suffering is experienced to generate the total welfare loss.
The lobbying over the economic stakes involved in product liability reflects the patterns of political influences one would expect for rent-seeking behavior. Table 22.8 provides a summary of the groups in favor of a pain-and-suffering damages cap in the left column, and a listing of the groups opposed to such a cap in the right column. This pattern reflects the economic stakes involved. Parties who bear the cost of such pain-and-suffering awards, ranging from business representatives at the U.S. Chamber of Commerce to groups representing the construction industry, favor pain-and-suffering caps. In contrast, labor and consumer groups generally oppose such caps because they limit the awards that the victims of accidents can potentially receive. The debate over the pain-and-suffering cap proposals and the institution of these caps by various states has been almost devoid of compelling economic reasoning. In the case of each party’s arguments, it has been the economic self-interest and the stakes involved that have driven the debate rather than any underlying rationale concerning the appropriateness of particular pain and suffering concepts.

### Table 22.8
Proposal to Impose Limits on Damages Awards for Pain and Suffering

<table>
<thead>
<tr>
<th>For</th>
<th>Against</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alliance of American Insurers</td>
<td>Association of Trial Lawyers of America</td>
</tr>
<tr>
<td>American Consulting Engineers Council</td>
<td>Brown Lung Association</td>
</tr>
<tr>
<td>American Medical Association</td>
<td>Consumer Federation of America</td>
</tr>
<tr>
<td>National Association of Home Builders</td>
<td>Consumers Union</td>
</tr>
<tr>
<td>National Association of Manufacturers</td>
<td>Environmental Action</td>
</tr>
<tr>
<td>National Association of Realtors</td>
<td>National Council of Senior Citizens</td>
</tr>
<tr>
<td>National Association of Towns and Townships</td>
<td>Public Citizen</td>
</tr>
<tr>
<td>National Federation of Independent Business</td>
<td>United Auto Workers</td>
</tr>
<tr>
<td>National School Boards Association</td>
<td>United Steelworkers Union</td>
</tr>
<tr>
<td>U.S. Chamber of Commerce</td>
<td>Women’s Legal Defense Fund</td>
</tr>
</tbody>
</table>


Risk Information and Hazard Warnings

One of the rationales for market failure is that consumers do not have perfect information regarding the safety of the products they purchase. In some cases, consumers may be able to monitor the overall riskiness of products as a group, but not the riskiness of products manufactured by particular companies. Consumers know that chain saws are hazardous, but they may have less ability to discriminate between the differing degrees of riskiness of Echo chain saws and Stihl chain saws.

In situations in which consumers know the average product risk, but not the risk used by the individual product, there will be a phenomenon akin to the classic lemons
Table 22.9
Markets with Imperfect Information: Lemon Markets for Risky Cars

<table>
<thead>
<tr>
<th>Fraction of Cars</th>
<th>Safety</th>
<th>Consumer Value with Perfect Information (dollars)</th>
<th>Group-Based Value (dollars)</th>
<th>Gain or Loss (dollars)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>High</td>
<td>30,000</td>
<td>23,500</td>
<td>+6,500</td>
</tr>
<tr>
<td>0.3</td>
<td>Medium</td>
<td>25,000</td>
<td>23,500</td>
<td>+1,500</td>
</tr>
<tr>
<td>0.5</td>
<td>Low</td>
<td>20,000</td>
<td>23,500</td>
<td>−3,500</td>
</tr>
</tbody>
</table>

problem.\textsuperscript{15} Table 22.9 presents an example for the automobile safety case. Suppose that there are three classes of cars ranging in safety from low to high. If consumers had perfect information regarding the properties of the cars, they would be willing to pay up to $30,000 for the safe car and as little as $20,000 for the average-safety car. Because they cannot distinguish the differing degrees of safety, they will make their judgments based on the average safety across this entire group of cars, which produces an average value to consumers of $23,500. The losers from this group-based value approach are the producers of the high-safety cars and the winners are the producers of the low-safety cars. This kind of redistribution from the high-quality to the low-quality market participants is a standard property of lemons markets. This property holds whether we are dealing with the properties of used cars or the salaries given to graduates of a college in a situation where the individual’s performance cannot be distinguished from the group average. The presence of such group-based pricing provides a disincentive for firms at the high end of the market because their safety efforts will not be rewarded. Rather, they will be sharing the benefits of safety with all of the other firms on the market.

**Self-Certification of Safe Products**

Firms can potentially avoid this pooling problem by identifying themselves as being producers of safer products. The gains and losses that would result from full disclosure of this type appear in the final column of table 22.9. The incentives for such revelation are greatest at the highest end of the quality spectrum and decline as one moves toward the lower end of the spectrum. Firms that produce the low-quality cars in this example would be willing to pay to have information on the differing riskiness of the cars suppressed.

The practical issue from the standpoint of the companies producing the high-quality cars is how to convey credibly to consumers the lower riskiness of their cars. This is the classic

economic signaling issue. Firms cannot simply claim that they produce high-quality products because such claims will have no credibility. All firms could make their claims without cost. The classic signaling problem is how to establish an economic mechanism so that the firm can credibly convey to consumers the safety of its products. Such mechanisms include the provisions of warranties and guarantees, since the costs of offering these product attributes are higher if one is marketing a risky product. Purely informational efforts such as ratings provided by government agencies and by consumer groups also may be of value in enabling consumers to make quality judgments.

Government Determination of Safety

Increasingly, the government has become active in trying to meet the informational needs of consumers. Beginning in 1965, Congress mandated hazard-warning labels on cigarettes. This system evolved over time, and in 1984 cigarettes began to include a system of rotating warnings alerting consumers to a diverse array of potential hazards from cigarettes.

Congress also instituted similar warnings for products containing saccharin. This artificial sweetener is potentially carcinogenic, but the scientific evidence and the extent of the risk have long been a matter of dispute. Indeed, in 1998 a panel of scientists convened by the government concluded that there was not sufficient evidence to designate saccharin a human carcinogen. Moreover, the benefits of the product from the standpoint of reducing obesity and its associated risks have greatly complicated the debate over the appropriate regulation of this product.

Similarly, in 1989, Congress mandated warnings on all alcoholic beverages. The first warning alerts consumers to the risk of birth defects from consumption of alcohol by pregnant women. The second warning notes the presence of health problems linked to alcohol and the effect of alcohol on one’s ability to drive and operate machinery.

These measures do not exhaust all initiatives of this type. Many states have also joined in with these efforts. Chief among these state initiatives is California Proposition 65, which has mandated the labeling of all significant carcinogens in food products. Many companies have avoided the stigma of labeling by reformulating their products, such as Liquid Paper, which formerly contained carcinogens. Although the implementation of this regulation remains a matter of debate, the proliferation of right-to-know measures has represented a major shift in the emphasis of regulation over the past quarter century.

Alternatives to Direct Command and Control Regulation

The rationale for employing an informational regulation rather than a direct command and control regulation is twofold. First, in many situations we do not wish to ban an activity alto-
gether. The regulatory agency may not have sufficient information to proceed with a ban, but would nevertheless like to alert consumers to a potential hazard in the interim so that they can at least exercise caution until the information for taking more stringent action becomes firmer.

Second, regulation through information may sometimes be the most appropriate and most effective response even when the agency is in doubt as to the appropriate course of action. If individuals differ in their tastes and their willingness to bear risks, then information provides consumers with the ability to make these market judgments and to choose the level of risk that is most efficient given their own preferences.

In addition, many decisions must necessarily be made on a decentralized basis. The care consumers take when using household chemicals cannot be monitored by direct regulatory agency, so that the most effective way to promote safety is to give consumers the motivation to undertake the appropriate level of precautions.

More generally, the task of promoting product safety is not simply one of designing an appropriate technology for the product. As we saw in the case of safety belts and aspirin caps, consumer behavior often plays a central role in determining the safety level that will result from a particular product design. The government has at its disposal two mechanisms to influence safety—producer and consumer actions. In general, we will be able to achieve greater gains for safety if we take advantage of both these mechanisms rather than simply rely on a technology-based approach.

The potential impact of such regulations is reflected in the statistics in table 22.10. This table provides information on the efficacy of the Drano label. Table 22.10 reports on the response of consumers to drain opener products for which the warning information had been purged, as opposed to their response to a label patterned after the current Drano label. The addition of risk information increases the frequency with which consumers wear rubber gloves or store the product in a childproof location.

Not all of the consumers would choose to wear rubber gloves even though the label urges that they take these precautions, because this precaution is onerous. Consumers might rationally choose to forgo this precaution if they believe that the benefits to them of taking it did not outweigh the costs associated with the nuisance value of wearing rubber gloves. The

<table>
<thead>
<tr>
<th>Precaution</th>
<th>No Warning (n = 59)</th>
<th>Drano (n = 59)</th>
<th>Incremental Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wear rubber gloves</td>
<td>63</td>
<td>82</td>
<td>19</td>
</tr>
<tr>
<td>Store in childproof location</td>
<td>54</td>
<td>68</td>
<td>14</td>
</tr>
</tbody>
</table>

warning label also has an effect on storage in a childproof location, as 14 percent more of the respondents will store the product in a childproof location after receiving the label. Moreover, for the subsample of the population with children under age five—the high-risk-of-poisoning group—almost all respondents who received the hazard warning would take the appropriate childproofing precaution.

Finally, it is noteworthy that even when the hazard-warning information is purged, over half of all consumers would undertake the precaution. It is important to recognize that consumers are not working in a vacuum. Even for a drain opener purged of warnings, over half of all consumers know enough about the potential hazards to take the precaution. Moreover, the studies of hazard warnings indicate that these warnings are effective only to the extent that they provide new knowledge to consumers. Programs of education that are intended to browbeat consumers into changing their behavior or that are intended to remind consumers about desirable courses of action generally are not successful. The programs that have been shown to be effective are those that provide new information to consumers in a convincing manner rather than those that fail to recognize that consumers are rational decision makers who are sometimes in need of important risk information.

An attraction of informational regulations from the standpoint of economists is that such regulations do not interfere with market operations to the same extent as technological standards. Nor do they impose substantial costs on firms. Rather, by providing information to consumers, the market can work more effectively in generating the incentives for efficient levels of safety.

---

**Regulation Through Litigation**

A phenomenon that emerged with particular force in the late 1990s was the use of litigation to force regulatory changes. The most prominent use of regulation through litigation was the settlement of the state tobacco suits in a form that was not a conventional damages payment. Rather, the agreement led to payments to the states that would be funded by a per-pack levy on cigarettes of about 40 cents, which in effect would function as an excise tax. The companies also agreed to a variety of restrictions that were regulatory in character, such as limitations on sponsorship of sporting events.

The tobacco lawsuit model has led to similar efforts against other risky products. Guns, fast food, health maintenance organizations, and lead paint have also been litigation targets. In some instances, the apparent objective is a financial payoff to the attorneys, while in others, the driving force appears to be to compel “voluntary” regulatory changes.
Breast Implant Litigation and Regulation

The often complex interaction of regulation and litigation is exemplified by the legal and regulatory debate over breast implants. Women had long attempted to enlarge their breasts, but these efforts increased substantially after 1962, when manufacturers introduced silicone-gel-filled breast implants. Silicone gel in a silicone envelope was less likely to cause health problems and adverse physical consequences than direct injections of silicone.

Table 22.11 summarizes the timeline of the regulatory and litigation events that followed the introduction of this medical device. At the time of their introduction, breast implants were not regulated by the FDA. Unlike prescription drugs, medical devices did not face tests of safety and efficacy. In 1976 medical devices such as breast implants were now under the FDA’s jurisdiction, but this product was grandfathered in. Thus, there was no explicit examination of the safety of breast implants by the agency.

While breast implants had cosmetic appeal to many women, especially in the South and the West, the product was not risk-free. Breast implants often leaked, ruptured, and led to a hardening of the surrounding tissue known as capsular contracture. Beginning in 1977, women began to win lawsuits against breast implant manufacturers and, as table 22.11 indicates, the degree of scrutiny by the FDA increased following these successful lawsuits.

Perhaps the watershed regulatory event was a successful lawsuit in December 1991 in which a plaintiff received over $7.3 million in total damages for her claims that ruptured implants caused connective tissue disease. FDA head Dr. David Kessler called for a moratorium on breast implant usage the following month. He viewed the situation as one of scientific uncertainty, but the effect of his action was to create widespread panic that led to the largest class action lawsuit, which eventually included over 400,000 women. The leading manufacturer of breast implants, Dow Corning, subsequently filed for bankruptcy. Preventing the initial marketing of a product posing uncertain risks has quite different consequences than declaring a moratorium on a product already implanted in millions of women. Fearing dire health consequences, a large number of women underwent costly and painful operations to have their breast implants removed. Subsequent scientific evidence indicated that the extreme fears concerning serious consequences such as systemic diseases lacked a sound scientific basis, though there are many adverse morbidity effects. Today, saline breast implants remain on the market and are becoming increasingly popular despite their adverse morbidity effects and leakage problems. In the George W. Bush administration an FDA advisory panel recommended that silicone breast implants be approved for use, but the agency withheld approval.

### Table 22.11
Timeline of Critical Breast Implant Events

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1962</td>
<td>Silicone-gel-filled breast implants first used.</td>
</tr>
<tr>
<td>1965</td>
<td>Silicone injections classified as a drug and not approved for human use.</td>
</tr>
<tr>
<td>1976</td>
<td>Medical Devices Amendments give FDA authority to regulate breast implants. Implants grandfathered in.</td>
</tr>
<tr>
<td>1977</td>
<td><em>Mueller v. Corley</em>. Plaintiff is awarded $170,000 due to rupture.</td>
</tr>
<tr>
<td>1978</td>
<td>FDA General and Plastic Surgery Devices Panel recommends Class II status. FDA concerns in 1978 include gel leakage in intact implants.</td>
</tr>
<tr>
<td>1982</td>
<td>FDA proposes Class III status.</td>
</tr>
<tr>
<td>1984</td>
<td><em>Stern v. Dow Corning</em>. Plaintiff is awarded over $1.7 million for claim that ruptured implants caused connective tissue disease. Internal Dow Corning documents showed Dow had suppressed risk information. These documents were then sealed by court order.</td>
</tr>
<tr>
<td>1988</td>
<td>Silicone implants are classified as Class III requiring manufacturers to submit safety information. FDA concerns in 1988 include capsular contracture, breakage, bleeding outside the shell, migration of silicone to organs, interference with the accuracy of mammogram, calcification of the fibrous capsule, immune disorders, and cancer.</td>
</tr>
<tr>
<td>Nov. 1991</td>
<td>Manufacturers’ safety information deemed inadequate by FDA.</td>
</tr>
<tr>
<td>Dec. 1991</td>
<td><em>Hopkins v. Dow Corning</em>. Plaintiff is awarded $840,000 in compensatory damages and $6.5 million in punitive damages for claim that ruptured implants caused her connective tissue disease.</td>
</tr>
<tr>
<td>Jan. 1992</td>
<td>FDA imposes a moratorium on silicone implants.</td>
</tr>
<tr>
<td>Feb. 1992</td>
<td>First class action filed in wake of FDA moratorium. Eventually 440,000 women join.</td>
</tr>
<tr>
<td>April 1992</td>
<td>Silicone implants withdrawn from market except in limited cases.</td>
</tr>
<tr>
<td>1994</td>
<td>Mayo clinic study shows systemic health risks not likely.</td>
</tr>
<tr>
<td>1995</td>
<td>Federal settlement approved with Dow dropping out.</td>
</tr>
<tr>
<td>1995</td>
<td>Dow Corning files for Chapter 11 bankruptcy reorganization, citing 19,000 individual implant lawsuits and at least 45 putative class actions.</td>
</tr>
<tr>
<td>1996–98</td>
<td>Courts appoint science panels. All panels conclude implants do not cause systemic diseases. Various courts do not allow plaintiffs’ experts to testify under Daubert.</td>
</tr>
<tr>
<td>1999</td>
<td>Institute of Medicine concludes only localized risks of silicone implants, including “overall reoperations, ruptures or deflations, contractures, infections, hematomas, and pain.”</td>
</tr>
</tbody>
</table>

The breast implant chronology reflects the interactions and failures of both regulatory action and the courts. For years the FDA failed to scrutinize the product risks adequately, and it was only when breast implant litigation put the product on the agency’s agenda that the agency acted, and, in the view of some critics, overreacted. Because the adverse health consequences of breast implants will not be apparent to women getting the implants, this product represents an ideal situation in which meaningful hazard warnings could play a constructive role. The role of the courts in some respects was constructive in that the litigation highlighted some actual product defects and called the risks to FDA’s attention. However, there were also court awards for ailments not apparently caused by breast implants. These awards also continued after some early scientific studies rebutting these claims began to appear in the literature.

This case study also illustrates how the informational environment plays a critical role in determining regulatory and judicial policies. These institutions were operating on the basis of highly imprecise information in a situation where the stakes from not acting were considerable. In such situations, it will often be imperative to make decisions before all the evidence is in. Some of these decisions may appear wrong in hindsight once the informational environment has changed, but that does not always mean that the decisions were incorrect at the time.

The Future of Product Safety Policy

The regulator choice process in the product safety area is even more diverse than figure 22.1 suggested. As a society, we have a broad set of choices involving the appropriate role of government regulation and tort liability. However, within these classes of institutional mechanisms there are also important decisions with respect to the particular mechanism for intervention.

Consider first the case of regulation. The alternative mechanisms by which we can intervene through government regulation are quite diverse. The government can specify in advance the technological standards that must be met by a product. A second possibility is to have some premarket approval mechanism whereby the company submits the product to the government for review and approval before the product can be marketed. A third possibility is to have government regulations that operate after the fact, as in the case of the product recall strategies for defective automobiles and consumer products.

One mechanism that has not been used as a regulatory arena is an injury tax approach, whereby the government imposes financial penalties on risky products rather than specify their technical characteristics. One reason why this final strategy has not met with widespread adoption is that in the product case it is particularly difficult to ascertain the contributory role of the product to a particular accident. The Consumer Product Safety Commission, for
example, keeps a tally of all accidents involving the use of particular consumer products, such as ladders. However, what we do not know is whether the ladder itself buckled under the consumer or whether it was simply the case that the consumer fell from the ladder. We know that the use of ladders is potentially risky, but ideally we want to establish a tax approach that will penalize unsafe ladders and not simply raise the price of all ladders irrespective of their safety. The latter approach is similar to the impact of a strict liability standard for ladders. In addition to raising the possibility of potentially substantial costs, this option also may create substantial moral hazard problems as well.

Since the mid-1980s the main policy debate has not focused on regulatory reform but rather on product liability reform. The reason for this emphasis is that the stakes imposed by product liability have escalated considerably and frequently dwarf the stakes involved with direct government regulation. Pharmaceutical companies are routinely faced with liability costs that may be in excess of the value of the sales of a product, particularly for small-market products such as vaccines, and the result is that such penalties have had a chilling effect on product innovation.

The fact that there has been some product innovation effect is not necessarily bad. Ideally, we do want tort liability to discourage the introduction of unduly risky new products. However, at the same time we do not want to deter beneficial innovations because firms are excessively cautious owing to the prospect of potentially enormous liability burdens. A fundamental task that must be addressed in the coming years will be to restructure the tort-liability system to strike a balance between the competing objectives of promoting safety and at the same time fostering the legitimate interests of the businesses affected by the regulation.

A final issue on the policy agenda is the overall coordination of regulatory and liability efforts. These are two different institutional mechanisms that affect similar classes of economic concerns. In some cases, the companies are hit twice by these institutions. They may adopt particular technological devices to come into compliance with formal government regulations, but the presence of such compliance does not provide any guarantees against additional liability costs being imposed on the firm.

The regulation through litigation phenomenon also raises concerns over whether the courts are usurping legitimate functions of the legislature and regulatory agencies. Large stakes litigation may be used to force regulatory changes and damages awards that are actually disguised excise taxes. Is the role of litigation a constructive one that fills the gaps left by other government policies, or is there an inappropriate overlap of these efforts? While governmental regulatory bodies appear to be better suited to the regulatory tasks, what if these efforts fell short? Whether the courts are overstepping their bounds will be a matter of continuing debate for which the answer surely will vary with the particular regulatory context and the degree of vigilance of the regulatory agency.
In addition to this potentially duplicative impact, there is also the issue of the appropriate division of labor between these social institutions. Which classes of hazards are best suited to being addressed by government regulations, and which are better suited to the ex post facto approach of tort liability that addresses specific accident cases identified in the courts? Differences in the temporal structure and character of the accidents no doubt will be important considerations in making these institutional allocations, but another issue may also be the difference in expertise. Many observers are beginning to question the ability of jurors to make the society-wide product safety decisions that are often part of the judgments that must be made to determine whether a particular product design is defective. Resolving such issues, which only recently have begun to be raised, will remain a central component of the future regulatory agenda.

Questions and Problems

1. Ideally, we would like to determine whether there is a product market failure before we intervene. Suppose that we do not have information on consumer risk perceptions, but we do observe the price reductions consumers are willing to accept in return for greater objective levels of safety with their product. For example, economists have estimated the implicit values of life associated with purchases of different automobiles. How would you use these implicit value-of-life estimates to determine whether there is a market failure?

2. The Food and Drug Administration uses a premanufacturing screening program for determining the marketability of new pharmaceutical products, whereas the U.S. Consumer Product Safety Commission relies on recall actions for products found to be defective, such as electric coffee-makers that short out. Can you think of any rationale other than historical accident for the difference in regulatory approaches between the two agencies?

3. The AIDS lobbyists have forced the FDA to accelerate the approval of AIDS-related drugs. In some cases, therefore, drugs that have not undergone the full FDA testing process to determine safety and efficacy will be used on patients. Is there a legitimate rationale for the expedited approval of AIDS-related drugs, or is this result simply due to the political power wielded by the AIDS lobbyists? Which other classes of drugs do you believe merit accelerated approval? More generally, if the FDA were to set up two different approval schedules, one being a thorough approval process and the second being a rapid approval process, what factors would you use in distinguishing whether a particular drug merited the thorough or the rapid approval approach?

4. Although Coase theorem types of influences can potentially lead to market provision of non-smoking areas and similar restrictions, current smoking policies do not reflect market influences alone. Municipalities throughout the country enacted various smoking ordinances, and the federal government is considering measures as well. What do you believe has been the impetus for such policies? Are there any efficiency-oriented reasons why it might be desirable for the government to set uniform standards in this area? More specifically, is there any likely source of market failure? In addition, are there any equity concerns that may be motivating these measures? Which parties will gain from an equity standpoint, and which will lose? Is there any reason to believe that the political outcomes will be efficient, and how would we judge efficiency?
5. Much of the debate over the efficacy of seat belt requirements and the influence of the counter-productive effect of decreased driver precautions noted by Peltzman stems from the crudeness of the empirical information that is available. If we can look only at accident rate totals by year or by state, then much key information will be lost. If you had unlimited resources and could commission your state police to develop an empirical database for you, what factors would you ask them to assess so that you could test the Peltzman effect conclusively?

6. After the advent of an increased role of tort liability, some products have disappeared. Let us take the case of diving boards at motels. What factors would you want to examine to determine whether this change in product availability is efficient?

7. It has often been noted that Melvin Belli’s procedure for determining pain-and-suffering damages is biased. Recall that his technique is to ask jurors to assess the pain-and-suffering value for a small unit of time, and then to extrapolate the value to the total time period over which the victim experienced the pain. Defense lawyers argue that the approach overstates the value of pain and suffering. What is the structure of the utility function for pain that must hold for the Belli approach to lead to an overstatement? Under what circumstances will it lead to an understatement?
Workplace health and safety levels are governed largely by three sets of influences: the market, direct regulation of risk levels by the Occupational Safety and Health Administration (OSHA), and the safety incentives created through workers’ compensation. In each case safety is promoted by creating financial payoffs for firms to invest in workplace characteristics that will improve worker safety. These incentives arise because improved safety leads to reduced wage premiums for risk, lower regulatory penalties for noncompliance, and reduced workers’ compensation premiums.

The labor market incentives provide the backdrop for the analysis of job safety. As discussed in chapter 20, these wage incentives are often quite substantial. The principal direct regulatory mechanism consists of OSHA’s health and safety standards. OSHA has long been a target of criticism. Critics have not questioned the agency’s fundamental objective. Promoting worker health and safety is a laudable and widely shared objective. Rather, OSHA is generally regarded as not fulfilling its mission of promoting this objective. Some observers claim that the agency imposes needless costs and restrictions on American business, whereas others claim that the agency’s efforts are not vigorous enough.

This branch of the U.S. Department of Labor began operation in 1971, after the Occupational Safety and Health Act of 1970 created it so as “to assure so far as possible every working man and woman in the nation safe and healthful working conditions.” Because ensuring a no-risk society is clearly an unattainable goal, the initial OSHA mandate established the infeasible as the agency’s mission. Nevertheless, a regulatory agency focusing on worker safety issues could serve a constructive function.

The early operations of OSHA did not, however, even begin to fulfill the agency’s initial promise. OSHA was the object of widespread ridicule for standards that prescribed acceptable toilet seat shapes, the placement of exit signs, the width of handrails, and the proper dimensions of OSHA-approved ladders. Many of the more frivolous standards were never among the most prominent concerns in the agency’s enforcement effort. Nevertheless, they did epitomize the degree to which the federal government was attempting to influence the design and operation of the workplace—matters that previously had been left to managerial discretion.

In recent years the stories of OSHA’s misguided regulatory efforts have been less prominent. One no longer reads amusing anecdotes like that concerning the OSHA inspector who penalized a firm for allowing its employees to work on a bridge without the required orange life vests, even though the riverbed was dry. The tone has shifted. Strident criticism in the 1970s gave way to comparative inattention. This inattention did not necessarily imply that the agency had been given a clean bill of health. There has been no widely publicized reform of the agency. Moreover, unlike transportation, natural gas, oil, and airlines, there have been

no legislative changes or major administrative reforms. The decrease in coverage of controversial OSHA policies occurred because a continuation of past policies, however ill-conceived, was simply no longer newsworthy. Moreover, firms had complied in many instances, so that the decisions regarding the standards were behind them.

In the 1990s the tone of public debate concerning OSHA shifted. After decreasing its enforcement effort in the early 1980s, OSHA became increasingly criticized for not doing enough. Instead of media coverage of apparently frivolous regulations, attention shifted to the continuing death toll in the American workplace. The status of job safety regulation began to be epitomized by the meatpacking worker who had become disabled because of lax regulatory enforcement and the thousands of asbestos workers who will die from job-related cancers. The late 1980s and the 1990s marked a substantial increase in activity in the job safety regulation area. Job health and safety remains an area where society is still striving to devise workable and effective regulatory mechanisms.

This chapter focuses on a general assessment of the effort to promote worker health: why we have such policies, how the initial effort failed, whether there has been any improvement in this effort, and how these policies can be reformed. Although only just over three decades old, OSHA has been the subject of a variety of proposed reform efforts. That OSHA has already become a chief target of proposed regulatory reforms suggests the kinds of fundamental changes needed in the agency’s initial orientation.

Several presidential administrations have promised an overhaul of OSHA policies. The Carter administration sought to provide this risk regulation effort with greater legitimacy by eliminating some of the more frivolous standards and by enforcing the sounder portions of OSHA regulations more vigorously. Under the Reagan and Bush administrations the attention shifted to decreasing OSHA’s confrontational character so as to foster a cooperative business-government approach to promoting workplace safety. The Clinton administration initiated a long-overdue increase in the scale of regulatory penalties, coupled with a substantial expansion in the scale of regulation. In 1994, for example, OSHA proposed indoor air-quality standards that would affect matters such as workplace smoking and, by OSHA’s estimates, would cost $8 billion per year. In part because unions objected to these smoking restrictions, the

regulation was never issued. Under the George W. Bush administration, the focus was on a less confrontational approach emphasizing outreach and cooperative programs.

Although these efforts rectified many of the more extreme deficiencies of OSHA’s initial strategy, calls for reform continue. Regulation of workplace conditions is a legitimate role for the government, but, as with other regulatory policies, there is a continuing need to maintain a balance between competing objectives. In this case the principal trade-off is between the costs imposed by the regulation and the health and safety benefits they provide. There is also a need to enforce the regulations that are promulgated in a manner that will foster effective incentives for compliance.

The Potential for Inefficiencies

There also may be more fundamental shortcomings whereby the government policy is failing to achieve as much safety improvement as is possible for the costs imposed. In more technical terms, the difficulty may be that we are not on the frontier of efficient policies (that is, those policies that provide the greatest safety for any given cost), as opposed to simply making the wrong trade-off along such a frontier.

Consider the set of feasible risk-cost combinations in figure 23.1. All points on the frontier ABC or to the right of it are potentially achievable. Ideally, the policy debate should be where along the policy frontier ABC our policy choice should be. Some finite rate of trade-off is required, inasmuch as complete safety is prohibitively costly in the case that is drawn. Unfortunately, a danger with ill-conceived policies is not that we are setting the trade-off rate incorrectly, but rather that we are wasting resources. If OSHA policies are now at point D, we could have greater safety at the same cost as point C or the same safety at less cost at point B.

Often the debate over ill-conceived regulations is miscast as one of values—where along the frontier should we be?—whereas the more fundamental problem is that a better policy could be designed regardless of one’s disposition toward regulation. Many of the most widely publicized standards initially promulgated by OSHA fall in the category of policies dominated by less costly and more effective alternatives. As in the case of other regulatory reforms, proper application of fundamental economic principles will illuminate the nature of the policy changes required.

How Markets Can Promote Safety

Before instituting a government regulation, it is instructive to assess how the market functions. Basically, one should inquire whether there is any inadequacy in the way in which market forces operate. Although individual life and health are clearly valuable attributes, there are many other market outcomes that are valued by consumers and workers but are not
regulated by government. Because markets that operate well will allocate resources efficiently, there should be some perceived inadequacy in the way these forces function before one interferes with their operation.

To ensure that market outcomes will be efficient, a number of stringent conditions must be met. For example, the outcome of any employment decision must affect the worker and employer only, not society at large, because these broader concerns will not be reflected in the job choice. A particularly pertinent requirement is that the job choice must be the outcome of a fully rational decision. Individuals must be cognizant of the risks they face and be able to make sound decisions under uncertainty. As we will discuss, these assumptions are especially likely to be violated for many important classes of risks.

Even if there is a consensus that market outcomes are not optimal, it is essential to ascertain the extent of the market failure. It is important to understand whether the operation of the market is fundamentally flawed or whether there is a narrower market failure, such as an informational shortcoming that can be remedied through an information transfer effort rather than direct control of workplace conditions.

**Figure 23.1**
The Policy Frontier for Job Safety
Finally, the market mechanisms will be pertinent insofar as they establish the context in which the government regulation operates. Regulations do not dictate health and safety outcomes, inasmuch as it is impossible for regulators to monitor and influence the health and safety attributes of all firms. Instead, these policies simply create incentives for firms and workers to take particular actions, such as installing new ventilation equipment. Whether regulations have any impact will hinge on the strength of the incentives created by the policy and the safety incentives the market generates for firms.

**Compensating Wage Differential Theory**

The fundamental economic approach to worker safety was sketched by Adam Smith more than two centuries ago. Smith observed that workers will demand a compensating wage differential for jobs that are perceived as being risky or otherwise unpleasant. This theory was the basis for the labor market estimates of the value of life discussed in chapter 20. The two critical assumptions are that workers must be aware of the risk (which may not always be the case) and that they would rather be healthy than not (which is not a controversial assumption). Attitudes toward health will, however, differ. Hersch and Viscusi, for example, have shown that smokers and those who do not wear seat belts are much more willing to work on hazardous jobs for less pay per unit risk than their safety-loving contemporaries, such as non-smoking seat-belt users. These differentials in turn will establish an incentive for firms to promote safety, since doing so will lower their wage bill. In particular, these wage costs are augmented by reduced turnover costs and workers’ compensation premium levels, both of which also provide incentives for safety improvements by the firm. In effect, it is primarily the risk-dollar trade-offs of the workers themselves that will determine the safety decision by the firm.

Figure 23.2 illustrates how these forces will influence the level of safety provided. Suppose that the health outcome involved is reduction of job-related accidents and that improvements in safety have diminishing incremental value to workers, just as additional units of other types of “economic goods” have diminishing importance. The marginal value of the safety curve in figure 23.2 consequently is a downward-sloping curve, for the initial increments in safety have the greatest value. Workers’ marginal value of safety is transmitted to firms through the wage rate the firm must pay to attract and retain workers. The firm can provide greater levels of safety, but doing so entails additional marginal (or incremental) costs that increase as the level of safety becomes increasingly great. Some initial safety improvements can be achieved

---


inexpensively through, for example, modification of existing machines or work practices. The addition of exhaust fans is one such measure for airborne risks. More extensive improvements could require an overhaul of the firm’s technology, which would be more expensive. This marginal cost curve consequently is increasing rather than staying flat, because safety equipment differs in its relative efficacy, and the firm will choose to install the most effective equipment per unit cost first.

The price of safety set by worker preferences will determine where along this marginal cost curve the firm will stop. The optimal level of safety from the standpoint of the market will be \( s^* \). The shaded area under the marginal cost curve will be the total safety-related expenditure by the firm. This level is short of the no-risk level of safety. At the level of safety provided, workers would have been willing to pay \( \nu \) per expected accident to avoid such accidents. Additional safety beyond this point is not provided because the cost to the firm for each extra accident avoided exceeds workers’ valuation of the improvement.

**Figure 23.2**
Determination of Market Levels of Safety
The level of health and safety selected will not be a no-risk level, because promoting safety is costly. Almost all our daily activities pose some risk because of the costs involved in reducing the hazards. Consumers, for example, routinely sacrifice greater crashworthiness whenever they select more compact automobiles in an effort to obtain greater fuel efficiency, for the typical small car is less crashworthy than the average full-sized car. Moreover, the order of magnitude of risks we regulate is not too dissimilar from the order of magnitude of risks that we encounter in other activities. As the data in table 19.3 indicated, the accident risk posed by one day of work in a coal mine (a relatively hazardous pursuit) is comparable in size to the risk of smoking 3.7 cigarettes, or riding 27 miles by bicycle, or eating 108 tablespoons of peanut butter, or traveling 405 miles by car. Individuals trade off these and other risks against other valued attributes, such as the recreational value of cycling.

### Risk Information

The first link in the compensating differential analysis is that workers must be aware of the risks they face. For example, if there is no perception of the risks, workers will demand no additional compensation to work on a hazardous job. The available evidence suggests that there is some general awareness of many of the risks workers face. Given data from the University of Michigan Survey of Working Conditions, there is a strong correlation between the risk level in the industry and whether workers perceive their jobs as being dangerous in some respect. This evidence is by no means conclusive, however, inasmuch as the risk assessment question ascertained only whether workers were aware of the presence of some risk, not the degree of risk posed by the job.

A more refined test can be provided using data based on a survey of workers at four chemical plants. In that study workers were asked to assess the risks of jobs using a continuous scale that could be compared with published accident measures. Overall, workers believed that their jobs were almost twice as hazardous as the published accident statistics for the chemical industry suggest, which is expected in view of the degree to which health hazards, such as cancer, are not reflected in the accident data. Particularly noteworthy was that after the health hazards were excluded from consideration, the risk assessments equaled the accident rate for the chemical industry.

5. These calculations were made by the authors using data from Richard Wilson, “Analyzing the Daily Risks of Life,” *Technology Review* 81, no. 4 (1979): 40–6.


8. The health risks were in effect excluded by informing one subsample of the workers that the chemicals with which they worked would be replaced by sodium bicarbonate (household baking soda).
These studies should be regarded as evidence of some reasonable perception of job risks by workers. They do not, however, imply that workers are perfectly informed. It is unlikely that workers have completely accurate perceptions of the risks posed by their jobs. These risks are not fully known even by occupational health and safety experts.

The degree to which there will be errors in the risk assessment will not, however, be uniform across all classes of risk. As a rough generalization, one would expect safety risks (external hazards such as inadequate machine guards) to be better understood than health risks (internal risks such as excessive exposure to radiation). Safety hazards tend to be more readily visible and familiar risks, such as the chance of a worker in a sawmill losing a finger. In contrast, health hazards usually are less well understood. These risks often involve low-probability events about which nothing is known whatsoever. Such risks may affect the individual decades after the exposure, so that learning by observation and experience is infeasible. These difficulties are enhanced in some instances by the absence of any clear-cut signals that a health risk is present. The odor and color of gases emitted in the workplace, for example, are not a reliable index of their potential carcinogenicity.

In situations where workers are aware of the hazard, the riskier jobs should be expected to command a wage premium. The value-of-life estimates in chapter 20 indicated that these compensation levels are quite substantial, on the order of $7 million for each workplace fatality. Similar compensating differential values exist for the implicit value of nonfatal injuries and illnesses, and these estimates are in the general vicinity of $50,000 for each lost workday illness or injury. Total wage premiums for job risk in the U.S. private sector are $34 billion for the 4,900 fatalities each year and $77 billion for the 1,537,567 injuries and illnesses, for a total wage premium of $111 billion. This amount, which is in addition to costs of workers’ compensation, represents the potential wage savings firms could achieve by making their workplaces safer.

These figures represent what workers’ risk-dollar trade-offs are, given their current information about the risk, not what they would be if they had full information about the risk. In addition, the calculations assume rational decision making, whereas in practice workers may overreact to risks or may neglect to take them into consideration. Although market behavior may not be ideal, the substantial magnitude of compensation per unit risk does suggest that there is substantial awareness of risks and their implications.

The value-of-life results are bolstered by analogous findings for nonfatal job injuries. These studies suggest that there is substantial compensation for job risks, when viewed both in terms of the total wage bill (6 percent of manufacturing workers’ wages) and the rate of compensation per unit risk.

The level of compensation may vary by industry. Unions, with a strong interest in health and safety issues, are particularly interested in securing workers’ hazard pay. Some unions, such as that for petroleum and chemical workers, often have specialized expertise in the health and safety area and have the ability to bargain with greater expertise than workers
could individually. Table 23.1 summarizes the overall breakdown in risk premiums by industry. These premium levels range from 3 to 5 percent of total earnings for industries such as apparel to as high as 12 to 15 percent for the lumber industry.

### Table 23.1
Risk Premiums as a Percentage of Total Earnings in Manufacturing Industries*

<table>
<thead>
<tr>
<th>Risk premiums of 3%–5%</th>
<th>Chemicals and allied products</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Petroleum refining and related industries</td>
</tr>
<tr>
<td></td>
<td>Electrical machinery, equipment, and supplies</td>
</tr>
<tr>
<td></td>
<td>Transportation equipment</td>
</tr>
<tr>
<td></td>
<td>Instruments and related products</td>
</tr>
<tr>
<td></td>
<td>Printing, publishing, and allied services</td>
</tr>
<tr>
<td></td>
<td>Tobacco manufacturers</td>
</tr>
<tr>
<td></td>
<td>Apparel and related products</td>
</tr>
<tr>
<td></td>
<td>Nonelectrical machinery</td>
</tr>
<tr>
<td>Risk premiums of 6%–9%</td>
<td>Textiles</td>
</tr>
<tr>
<td></td>
<td>Paper and allied products</td>
</tr>
<tr>
<td></td>
<td>Primary metals</td>
</tr>
<tr>
<td></td>
<td>Rubber and plastics</td>
</tr>
<tr>
<td></td>
<td>Fabricated metal products</td>
</tr>
<tr>
<td></td>
<td>Leather and leather products</td>
</tr>
<tr>
<td></td>
<td>Stone, clay, and glass products</td>
</tr>
<tr>
<td>Risk premiums of 12%–15%</td>
<td>Food and allied products</td>
</tr>
<tr>
<td></td>
<td>Furniture and fixtures</td>
</tr>
<tr>
<td></td>
<td>Lumber and wood products</td>
</tr>
</tbody>
</table>

* These premiums are derived from earnings equations that are estimates of the relationship between injury rates and workers’ earnings.


On-the-Job Experience and Worker Quit Rates

The presence of possibly inadequate worker knowledge concerning the risks remains a potential impediment to the full operation of the compensating differential mechanism. The result will not be that market mechanisms will work less effectively, although some decreased efficacy will undoubtedly occur. Rather, there will also be new market forces that may be influential.9

Consider a situation in which a worker starts a job without full knowledge of the potential risks. After being assigned to the position, he or she will be able to observe the nature

of the job operations, the surrounding physical conditions, and the actions of co-workers. Similarly, during a period of work on the job the worker learns about some particular difficulties in carrying out the job tasks, and even more directly, he or she observes whether co-workers are (or have been) injured. The worker can then use these experiences to evaluate the risk potential of the job.

If the worker’s risk perceptions become sufficiently unfavorable, given the wage paid, he or she can quit and move to another firm. Overall, job risks account for one-third of all manufacturing quit rates. Similarly, the periods of time that workers spend at hazardous firms before leaving are shorter than for safe firms. As a consequence, there will always tend to be more inexperienced workers in high-risk jobs, because the high turnover rates from these positions lead to frequent replacements.

The standard observation that younger and more inexperienced workers are more likely to be involved in accidents is not entirely attributable to greater riskiness of this demographic group. Rather, the causality may be in the opposite direction, inasmuch as new hires are more likely to be placed in the high-risk, high-turnover jobs. The firm will also have a strong incentive to avoid placing its most experienced workers in these positions, because it will lose the training investment if the worker is injured or quits.

All of the labor market responses by workers are simply variations of the compensating differential theme. If the job appears to be risky initially, the worker will require extra compensation to begin work on it. Similarly, once he or she acquires information about the risks that are present, the worker will reassess the job’s attractiveness and remain with it only if the compensating differential is sufficient.

### Inadequacies in the Market

If market operations were fully efficient, there would be no need for government regulation of health and safety. The decentralized operation of the market would be sufficient to ensure appropriate levels of the risk. Two broad classes of shortcomings limit the efficacy of market outcomes: (1) informational inadequacies and problems with individual decisions under uncertainty and (2) externalities.

### Informational Problems and Irrationalities

For the compensating differential model to be fully applicable, workers must be cognizant of the risks they face and be able to make sound decisions based on this knowledge. The available evidence suggests that in many contexts, workers have risk perceptions that appear plausible, but these studies in no way imply that all workers are fully informed. There is a general consensus that many health risks in particular are not well understood, and indeed, workers may be completely ignorant of some of the risks they face.
With on-the-job experience, workers undoubtedly will revise their perceptions of many risks. Once again, safety hazards are more likely to be treated in a reliable manner because they tend to be readily visible and to occur with much greater frequency than many health risks, which often are low-probability events. Thus the worker has fewer observable incidents of adverse health outcomes to use in forming his risk assessment. The long time lags involved in many health risks further impede efforts to learn about the implications of these risks through experience. A worker may get cancer two decades after job exposure to a carcinogen, but tracing the cause to the job usually is not feasible. As a rough generalization, there is probably reasonable but not perfectly accurate perception of many safety risks and much less reliable assessment of the pertinent health risks.

Even with accurate perceptions of the risk, however, one cannot be confident that the decisions ultimately made by the workers will be ideal. Decisions under uncertainty are known to pose considerably more difficulties than decisions made in cases where the outcomes of alternative actions are known in advance. These difficulties are likely to be particularly great in situations involving very low-probability events that have severe outcomes after a substantial lag. The low probabilities and substantial lags make these decisions difficult to conceptualize. How averse, for example, is a worker to taking a one in 20,000 risk of cancer twenty-five years from now? Because of the high stakes—possibly including the worker’s life—the cost of mistaken choices will be high. Once again, it is likely that health hazards pose relatively greater demands on individual rationality than safety risks.

The final class of shortcomings in individual behavior relates to the degree workers can choose from a variety of alternative risk-wage combinations. For the relatively mobile, modern U.S. economy there seems to be substantial range of job options for almost all workers. Certainly, the classic textbook discussions of the one-company town no longer seem relevant and, even if true, would not have as great an impact in an era of interstate highways and substantial worker mobility. This mobility may be restricted during cyclical downturns, when job opportunities are less plentiful, but because accidents move procyclically, the net influence of adverse economic conditions is not clear-cut.

Perhaps the most important constraint on individual mobility is related to the character of the employment relationship. Once on the job, individuals acquire skills specific to the particular firm, as well as seniority rights and pension benefits that typically are not fully transferable. If workers had full knowledge of the risk before accepting the position, these impediments to mobility would not be consequential. The basic difficulty, however, is that workers may not have been fully cognizant of the implications of the position and will subsequently become trapped in an unattractive job situation. Available evidence

for chemical workers suggests that the extent of serious job mismatches of this type is not high.

Externalities

An additional class of market inadequacies arises even if individual decisions are fully rational and ideal in all respects. Parties outside of the market transaction for the job may have a stake in the risky job insofar as there is a broader altruistic concern with individual health. This type of health-related altruism is probably of greater consequence than redistributional concerns in this context. Life and health are clearly quite special, as society has undertaken a variety of health-enhancing efforts, such as Medicare, to promote individual well-being.

The overall importance of these altruistic interests has not yet been ascertained, however. The evidence summarized in chapter 20 was exploratory in nature. In contrast, individuals’ values of life and health are considerable, and it is not obvious that the external interests of society would boost these values substantially. Whether society’s broader altruistic concerns are of great consequence in this area is an open empirical issue that merits further attention.

Moreover, there is the ethical issue of whether there is a legitimate altruistic concern or simply an attempt by more affluent citizens to impose their own risk-dollar trade-offs on others. High-income, white-collar workers may view most blue-collar jobs as unattractive, but this opinion does not mean that social welfare will be enhanced by preventing anyone from working on this class of jobs. Until these questions can be resolved, the primary impetus for regulation of occupational hazards probably should be the shortcomings of worker decisions.

OSHA’s Regulatory Approach

The general approach OSHA has taken to regulating job safety is dictated at least in part by the Occupational Safety and Health Act of 1970. This legislation authorizes OSHA to set standards and to do so in a manner that will ensure worker health and safety. OSHA’s enabling legislation did not, however, specify what these standards should be, what general character they should take, or how stringent they should be.

In addition, the legislation did not specify the nature of the enforcement of the standards. For example, OSHA could couple standards with a penalty for firms out of compliance, where the penalty is set at a level that could give firms some discretion as to whether compliance is desirable. For example, the penalty could be related to the health impacts on workers, and the firm could comply with the standard only if the health benefits exceeded the costs to firms. (The frequency of OSHA inspections could also influence the penalty.) In actuality, OSHA imposes an ever-escalating series of penalties on firms out of compliance; thus the standards
can be viewed as rigid guidelines. Because of this binding character, the level and nature of the standards are of major consequence to firms regulated by OSHA.

**Setting OSHA Standard Levels**

One could characterize OSHA's general approach as that of adopting technology-based standards whose stringency is limited only by their affordability. Cost considerations enter only insofar as OSHA is concerned with shutting down affected firms. To see how OSHA's strategy differs from a standard benefit-cost approach, consider figure 23.3. For simplicity, suppose that the marginal safety-benefit curve is flat, so that there is a constant unit

![Figure 23.3](image-url)

**Figure 23.3**

OSHA Standard Setting versus Efficient Standard Setting
benefit value. The marginal cost of providing safety is rising, as it becomes increasingly more expensive to promote safety.

The strategy of OSHA is to look for the kink in the marginal cost curve—at what point does added safety become prohibitively expensive? For figure 23.3, that point is at \( s_2 \), whereas the efficient level of safety is at \( s_1 \). The strategy advocated by most economists is that the agency should pursue a more balanced approach that recognizes the necessity of taking into account both the costs and risk-reduction benefits in a comprehensive manner. What matters is the relationship between marginal benefits and marginal costs, not whether costs happen to jump at a particular point. Costs should always be a matter of concern, not simply when a firm may go out of business as a result of OSHA policies. Such a shift in emphasis need not always lead to more lenient regulations. Some very hazardous firms probably should go out of business if provision for efficient levels of safety and health will not permit them to earn a profit.

Much of the policy-oriented debate over the safety standards has concerned their stringency. Those advocating a more balanced approach note that the Occupational Safety and Health Act does not require a risk-free workplace, only one that promotes safety “as far as possible.”\(^\text{11}^-\) This and other qualifiers in the act suggest that OSHA might have some leeway in being able to take costs into consideration. This view was bolstered somewhat by the U.S. Supreme Court’s decision in the 1980 benzene case, in which it overturned the standard because OSHA had not shown that the reduction in risks would be “significant.”\(^\text{12}^-\) This “significant risk” criterion imposes a threshold benefit level, but it does not impose a requirement that OSHA balance benefits and costs.

Indeed, such benefit-cost tests were explicitly ruled out in the 1981 U.S. Supreme Court decision regarding the OSHA cotton dust standard.\(^\text{13}^-\) The court upheld the OSHA cotton dust standard and interpreted the feasibility provisions of the Occupational Safety and Health Act as meaning “capable of being done.” It is the technical possibility of compliance rather than benefit-cost trade-offs that should guide OSHA decisions.

In fact, however, in this instance OSHA had based its cotton dust standards on cost-effectiveness concerns, not simply affordability. Specifically, the standard is varied across different stages of processing because of difference in the severity of the risk in these areas and differences in the cost of reducing the risk. Further reductions in the risk were clearly “capable of being done,” and in fact many firms have already achieved cotton dust levels well below those specified in the standard.\(^\text{14}^-\)

Clearly, technological feasibility cannot be divorced from cost considerations, since almost any risk can be reduced at sufficiently large costs. Drivers, for example, would face a lower risk of injury in an auto accident if everyone drove full-sized cars at speeds under 35 miles per hour. Such measures have not been undertaken because the safety benefits do not justify the increased travel time and loss in fuel efficiency. Likewise, OSHA varied the cotton dust standard because the severity of cotton dust exposures differs according to the stage of processing (because different types of fibers and dust are airborne at different stages) and because compliance costs differ.

Indeed, under the Reagan administration, OSHA began to routinely calculate the costs and benefits of its proposed regulations. The agency does not, however, explicitly compare these magnitudes when discussing the reasons for its policy recommendations. Inevitably, some comparisons of this type are made by OSHA, the Office of Management and Budget, and other players in the regulatory process. There would be greater likelihood of balanced policies if the Supreme Court reversed its narrow and unrealistic interpretation of OSHA’s mandate or if Congress amended OSHA’s legislation. In the absence of such a change, primary emphasis will continue to be placed on the level of risk reduction rather than the associated costs. Regulations sometimes may impose costs that appear to be well out of line with any reasonable values, such as almost $70 million per expected life saved by the OSHA arsenic standards.

The Nature of OSHA Standards

The structure of OSHA’s regulatory approach also has been overly restrictive, as the agency has adopted a narrow technology-based approach to safety regulation. Ideally, OSHA should permit firms to achieve any given level of safety in the least expensive manner possible, consistent with having well-defined regulations that are enforceable. Instead, OSHA has typically adopted uniform standards that attempt to prescribe the design of the workplace.

This orientation derives in part from the pattern set in OSHA’s initial standard-setting activity. Shortly after beginning operations, OSHA issued more than 4,000 general industry standards for health and safety, the preponderance of which were safety related. These standards, which continue to constitute most of OSHA’s safety policies, were derived from the national consensus standards of the American National Standards Institute, the National Fire Protection Association, and some existing federal standards for maritime safety. In this process, OSHA converted a set of discretionary guidelines into a mandatory prescription for workplace design.

The upshot of this effort was to establish OSHA as a leading object of ridicule for its portable toilets for cowboys and other seemingly trivial standards. Perhaps more significant than these well-publicized OSHA horror stories was the specification character of the regulations. The OSHA handrail regulation specifies their required height (30 to 34 inches), spacing of posts (not to exceed 8 feet), thickness (at least 2 inches for hardwood and 1\(\frac{1}{2}\) inches
for metal pipe), and clearance with respect to the wall or any other object (at least 3 inches). Likewise, in its requirements for band guards for abrasive wheels, OSHA specifies the required thickness, the minimum diameter of rivets, and the maximum distance between the centers of rivets.

In each case the specification standard approach may have imposed greater costs than equally effective alternatives. To provide guidelines for how such flexibility could be achieved, President Ford’s Task Force on OSHA, headed by economist Paul MacAvoy, designed a model standard for machinery and machine guarding that indicated, for example, several ways to guard a punch press. This flexibility also may enhance the safety that could be achieved through a performance-oriented approach. A performance-oriented approach would stress the need for firms to achieve a particular health and safety level through whatever means they chose rather than be required to install a particular type of technology. The present OSHA specification standards are so narrowly defined that they pertain to only 15 percent of all machines. This model standard has not yet been adopted, but it provides an operational example of how OSHA could achieve greater flexibility in its regulatory approach without jeopardizing worker safety.

It is also noteworthy that the primary orientation of the standards remains in the safety area. Externally visible aspects of the workplace, such as handrail width, are given comprehensive and meticulous treatment. In contrast, only a small fraction of the carcinogens in the workplace have been addressed by OSHA standards. There are some health standards, such as those for radiation exposure, but for the most part the standards have been dominated by safety concerns.

In view of the earlier discussion of market inadequacies, this emphasis seems misplaced. Health risks rather than safety risks are handled least effectively by the market. The greatest potential gains from OSHA regulation are likely to come from addressing the dimly understood health risks that pose the most severe difficulties for worker decision making.

Moreover, the structure of the health standards is also more likely to be conducive to more effective promotion of worker health. The health standards typically limit worker exposure rather than specifying particular technologies. For example, the cotton dust standard specifies permissible exposure limits to airborne concentrations of respirable cotton dust in different stages of processing, and it indicates the circumstances under which protective equipment must be worn. Respirators are needed during cleaning operations because of unusually high levels of cotton dust in that period. The standard does not specify how the

18. Ibid., preface.
lower levels of cotton dust are to be achieved, whether through use of exhaust fans, new machines for drawing and carding the cotton, or some other approach.

While health risks may present greater opportunities for productive regulatory interventions, health regulations in some instances have been less cost-effective in saving lives. The following comparison is not atypical. A 1997 OSHA standard for methylene chloride exposures had a cost of $13 million per life saved, while the 1996 OSHA safety standards for scaffolds had a cost per life saved of $200,000. While health hazards present potential productive opportunities for regulation, if the standards are set at a too-stringent level, the potential net gains will not be realized.

The Reform of OSHA Standards

Proposals for reforming OSHA standards have focused on three dimensions. The first recommendation is that there should be a shift in emphasis from safety to health. Second, there should be greater opportunities for firms to find less expensive techniques for promoting safety. Standards should consequently be more performance oriented when that approach is feasible. Finally, the level of the standards should be set in a more balanced fashion that attempts to recognize the health benefits to workers and the costs to firms.

Regulatory Reform Initiatives

Compared with its initial activity, OSHA’s standard setting has been relatively modest in the past decade. During the Carter administration, much new regulation was stymied by the uncertainties caused by the court challenges of OSHA’s legislative mandate in the cotton dust and benzene cases. The Reagan administration’s emphasis was on slowing the pace of new regulation rather than changing its character, so that OSHA was less active than in its earlier years. Neither the Clinton administration nor the George W. Bush administration returned OSHA to a vigorous pace of regulatory initiatives. Nevertheless, OSHA has not been completely dormant in the standards area.

Changes in OSHA Standards

The chief legacy of the Carter administration in the area of regulatory reform was its overhaul of the safety standards. The primary emphasis was not on a general restructuring of the standards approach but on eliminating those portions of the standards that were most extraneous and ill-conceived. This emphasis was quite appropriate, in view of the importance of establishing the agency’s credibility. Assistant Secretary of Labor for Occupational Safety and Health Eula Bingham eliminated or modified 928 OSHA regulations in all in October 1978. In many cases these changes were only editorial and had no major substantive impact.

Nevertheless, the net effect of the elimination of the “nit-picking” features of OSHA regulation was to reduce some of the harsher criticisms of the agency’s regulatory approach. Because of the magnitude of OSHA’s initial credibility problem, the importance of even cosmetic changes in the standards should not be underestimated.

**Chemical Labeling**

The most important structural change in regulatory policy was OSHA’s chemical labeling regulation, which was proposed at the end of the Carter administration and finalized by President Reagan. By providing workers with information, this regulation represented an effort to use market forces to promote safety. The chief forms of information provision required were labels on the chemicals, material safety data sheets on the nature of chemicals used in the workplace, and a program for training workers in the handling of chemicals. This regulation addresses the primary source of market failure directly and, as a consequence, preserves the constructive aspects of the health-related decisions by firms and workers. In addition, the focus of the regulation is strongly oriented toward health hazards rather than safety risks.

Indeed, much of the impetus for this regulation came from the inability of direct regulatory controls to address the entire range of chemical hazards. Setting standards for all of the thousands of carcinogens in the workplace was viewed as unfeasible.

In addition to addressing long-term health impacts and acute health effects (for example, skin rashes from chemical exposures), the regulation also affects accidents from fires and explosions. These safety hazards also are likely to merit greater attention than more visible workplace characteristics, since the safety-related properties of chemicals will not be well understood in the absence of some information about risk.

**The Economic Role of Hazard Warnings**

The attractiveness of hazard warnings from an economic perspective is that they work in conjunction with market forces by eliminating the informational market failure directly. Moreover, in many instances it may be that altered worker actions are a more efficient means of promoting safety than technological changes in workplace conditions.

The manner in which hazard warnings exert their influence is reflected in the data in table 23.2, which are based on reactions of workers in four major chemical plants to different hazard warnings. Each worker was shown a hazard warning for a particular chemical and was told that the chemical would replace the chemical with which the worker currently worked. In each case, the worker was shown a single warning, where the four different chemical labels used were for sodium bicarbonate (household baking soda), chloroacetophenone

(an industrial chemical that is an eye irritant), TNT (a well-known explosive), and asbestos (a leading occupational carcinogen). Workers were then asked a series of questions regarding their attitudes toward the job after it had been transformed in this manner.

The first row of statistics in table 23.2 gives the change in the fraction of workers in the sample who viewed their jobs as being above average in risk after being given the hazard warning information. In the case of sodium bicarbonate, the fraction of the workers who viewed their job as above average in riskiness dropped by 35 percent, so that overall for the sample the workers all viewed their jobs as relatively safe. In contrast, for the remaining three chemicals there was a substantial increase in the fraction who believed that their jobs were risky, particularly in the case of the most severe hazards posed, asbestos and TNT. If the market operates efficiently, these risk perceptions in turn should lead to additional wage compensation for the jobs. Workers who were shown the sodium bicarbonate label did not require any additional wage compensation to work on the job, whereas workers shown the other three chemicals required amounts ranging from $1,900 to $5,200 per year in order to remain on the job. The final market mechanism discussed earlier is that if workers are not compensated sufficiently, they will quit. For this sample, if there were no change in the wage rate after the introduction of the hazard warning, quit rates would decline by 23 percent for sodium bicarbonate and would rise by up to 63 percent for workers who are exposed to asbestos. In terms of creating incentives for safety, hazard warnings serve to augment market forces by informing workers of the risks that they face. In addition, other studies of individual precautions indicate that hazard warnings are also likely to lead to increased precautions as individuals become better informed of the risks they face, as well as the precautions needed to reduce these risks.

### Effective Hazard Warnings

It should be emphasized that for these hazard warnings to be effective they must provide new information. Thus the source of the market failure is an information gap. Education efforts
that are primarily efforts of persuasion and that attempt to browbeat individuals into changing their behavior have met with far less success.

**Innovations in OSHA Regulation**

The chief new safety standard under the Reagan administration is also noteworthy because it also marked a change in the character of OSHA regulation. That regulation consisted of a set of extensive rules intended to decrease the risks associated with grain handling. These hazards are often well publicized, for explosions in grain-handling facilities may lead to the deaths of dozens of workers. Perhaps in part because of this publicity and the safety incentives created by the market and workers’ compensation, there were no deaths from explosions in 1983.

The 1984 OSHA regulation was intended to reduce this risk further by decreasing the dust levels in grain elevators, which in turn will reduce the risk of explosions. What is noteworthy about this standard is that firms are given several alternative options to decrease the dust: (1) to clean up the dust whenever it exceeds $\frac{1}{8}$ inch, (2) to clean up the dust at least once per shift, or (3) to use pneumatic dust-control equipment. This flexibility represented a major innovation in the design of OSHA safety standards. The regulation provides an opportunity for firms to select the most cost-effective option and will lead to lower compliance costs than would a uniform specification standard. OSHA’s effort to use the advantage of a performance-oriented approach represents a significant, constructive contribution to OSHA policy development. Such efforts are likely to put us on the frontier of efficient regulatory policies.

Overall, there has not been a dramatic change in the structure of OSHA safety standards since OSHA’s initial standard-setting efforts. Some of the extraneous and more frivolous standards have been pruned, other standards have been updated to take technological changes into account, and a few new standards have been added.

Further reform in standards that have already been promulgated is expected to be minimal, inasmuch as there is not a strong constituency for such changes. To the extent that more firms comply with the revisions of the OSHA standards, any impetus for relaxations or modifications of existing regulations will be diminished.

Some progress may be made regarding future standards in the form of greater recognition of the costs of the regulations and the introduction of innovative approaches to regulation. Two OSHA efforts of the 1980s, the chemical labeling standard and the grain-handling standard, represent significant advances in OSHA’s regulatory approach. On balance, however,

OSHA’s level of activity in the standards area has not been great since its inception, as it has retained most of its original approach.

OSHA’s Enforcement Strategy

To design and enforce its standards, OSHA now has 2,303 employees, ranking second behind the EPA among social regulation agencies. Almost half of these employees, 1,123, are inspectors who monitor the compliance of workplace with OSHA standards. This staff, in conjunction with the inspectors from states that choose to enforce OSHA regulations with state inspectors, come to the workplace, ascertain whether there are any violations, and penalize violators. The inspectors may return for a follow-up inspection, continuing to assess penalties until compliance is ensured.

Firms will choose to comply with OSHA standards if OSHA establishes effective financial incentives for doing so. The firm must consequently find it more attractive financially to make the safety improvements than to risk an adverse OSHA inspection. The penalties that result include fines levied by OSHA, as well as possible adverse effects on the firm’s reputation, which may in turn affect worker turnover or wages. To assess whether these safety incentives are strong, we will consider each link in the OSHA enforcement process.

Before OSHA can affect a firm’s policies, it either must inspect the firm or create an effective threat of possible enforcement. OSHA undertakes four types of inspections: (1) inspections of imminent dangers, (2) inspections of fatalities and catastrophes, (3) investigations of worker complaints and referrals, and (4) programmed inspections.23 This priority ranking has remained virtually unchanged since OSHA’s inception. Somewhat surprisingly, complaint inspections produce few violations per inspection, a finding which suggests that disgruntled workers may be using the OSHA inspection threat as a means of harassing the employer.24 This pattern is unfortunate, since the role of workers and unions in promoting safety could potentially have been instrumental.

There have been five different eras of OSHA enforcement. The Nixon and Ford administrations established the general inspection approach, and there was little change in emphasis except for a gradual expansion in the enforcement effort. Under the Carter administration there was an attempt to eliminate some of the less productive aspects of the enforcement policy. The number of inspections and less important violations declined, and penalties for violations increased. The first term of the Reagan administration marked the start of what was termed a less confrontational approach. In effect, the inspection effort was scaled back. The Reagan administration also introduced more conscious inspection targeting. The biggest change was that the level of penalties assessed for OSHA violations plummeted. In the second

Reagan term and in the Bush administration, the enforcement effort was not bolstered, with penalties at an all-time low. In the Clinton and George W. Bush administrations, the enforcement staff has remained sparse, but penalties have increased by several orders of magnitude.

**Inspection Policies**

The total number of inspections rose steadily through fiscal year 1976, after which it dropped by one-third as a result of the Carter administration’s attempt to reduce the less productive inspections. The present total of state and federal inspections of 96,000 annually may seem substantial, but it covers very few workplaces. At this rate of inspection an enterprise would be inspected about once every century.

Because many firms are small businesses with few employees, a more accurate index of coverage is the inspection rate per worker. There are 111 million workers at firms under OSHA’s jurisdiction. The total number of all state and federal inspections per worker is one per 1,156 workers. Many of these are follow-up inspections, so that these statistics overstate the coverage of OSHA inspections. Some OSHA critics have correctly suggested that the chance of seeing an OSHA inspector is less than the chance of seeing Halley’s Comet. In contrast, the EPA inspects all major water polluters roughly once per year. Moreover, there has been a substantial drop in the rate of coverage of employees.

Two aspects of inspections that reflect desirable changes in emphasis pertain to the emphasis on health rather than safety and the emphasis on serious violations. Health violations merit relatively more attention, for there are greater inadequacies in the way these risks are treated. Safety risks are often well known to workers and generate compensating wage differentials, higher quit rates, and larger workers’ compensation premiums, all of which establish incentives for firms to promote safety. In contrast, health hazards are less well understood and, because of difficulties in monitoring causality, are not covered as effectively by workers’ compensation.

The role of health inspections doubled under the Carter administration, in part because the decline in overall inspections in fiscal year 1977 primarily represented a drop in safety inspections. The pattern through fiscal year 1981 was one of a gradual rise in the absolute number of health inspections. This increase was reversed under Reagan not so much because of a conscious decision to abandon the health area but because of the shift toward construction inspections, which are primarily safety-related.

Ideally, inspections also should identify serious violations rather than less consequential threats to worker safety. This emphasis on serious violations escalated considerably under the Carter administration, as almost one-third of all inspections began to generate serious violations. The frequency of serious violations under the Reagan administration was roughly the same as under the Carter administration.
Trivial Violations

On entering the workplace, the OSHA inspector attempts to identify violations of OSHA standards for which he or she will assess penalties. In determining whether a firm is in compliance an OSHA inspector cannot consider costs of meeting the standard, only technical feasibility.

In fiscal year 1977, when OSHA eliminated less important inspections and citations for trivial violations, there was a dramatic drop in the number of OSHA violations. Thereafter there has been a gradual and steady decline in the number of violations, with an additional small downward shift under Reagan. At present, each federal inspection results in just two violations of OSHA standards. An important change has been the emphasis on violations for serious threats to worker health, which now constitute two-thirds of all violations cited by inspectors.

OSHA Penalties

The ultimate determinant of the financial impact of an OSHA inspection is the amount of the penalties that are assessed for noncompliance. Notwithstanding the widespread notoriety of the enforcement effort, these penalty levels have always been inconsequential. Annual penalties had typically been well below $26 million—roughly the same financial incentive created by the wage response to up to four additional deaths per year.

One change in the penalty structure occurred in the reforms of fiscal year 1977 when, at the insistence of Congress, OSHA eliminated penalties for firms with fewer than ten nonserious violations. The overall level of penalties, however, increased under the Carter administration to more than double its earlier level.

Under President Reagan, OSHA adopted a less confrontational approach in which penalties were well below their earlier levels. A particularly noteworthy change was that firms could obtain reductions in the assessed penalties by up to 30 percent if they made a serious effort to comply with the standards.

The resulting financial incentives for safety were not great. Penalties have long averaged about $50 per violation, an amount that will provide little incentive for safety. The Clinton administration was most vigorous in boosting the penalty levels so as to create more meaningful financial incentives for safety. The penalty level per violation under the George W. Bush administration remained much higher than the historical average, as it was $1,005 for federal inspections and $527 for state inspections. Even the most egregious violations do not receive penalties that are inordinately large. Federal penalties for willful violations are just over $28,000. Total OSHA penalties in fiscal year 2002 were $149 million.

While the penalty level has ratcheted up from the level in the early days of OSHA, these financial incentives are dwarfed by other economic forces at work. Market incentives for safety through wage premiums for risk total $111 billion, and workers’ compensation
premiums were $26 billion in 2001. Since each of these costs will drop if firms have a better safety record, they provide a powerful incentive for safety.

Firms that are found in violation can reduce the financial costs of being out of compliance by making the mandated improvements. Because of the reduction in penalties for firms that remedy OSHA violations, there is little threat from a random OSHA inspection. A firm need do little to promote safety, but simply await the OSHA inspector. The firm will avoid correcting safety problems that the inspector may not identify, since the inspector will typically find only a couple of violations, and the firm will face few penalties if it makes the suggested changes. The elimination of the expected losses from inspections suggests that OSHA will have little impact on the great majority of firms that are not inspected, for inspections now have little deterrence value.

**Enforcement Targeting**

In addition to changes in the level of OSHA enforcement, there have also been shifts in the focus of the enforcement effort. Perhaps the most controversial recent change in OSHA enforcement policies was the introduction of records-check inspections in October 1981. In these programmed safety inspections the OSHA inspector first examined the firm’s lost-workday accident rate for the past two years (three years for very small firms). If this rate was below the most recently available national manufacturing lost-workday rate, the firm was not formally inspected. For example, a firm inspected in 1985 would have available its 1983 and 1984 lost-workday accident rates for comparison with the 1983 manufacturing rate, because there is a two-year lag in publishing the Bureau of Labor Statistics data.

Ideally, OSHA should target riskier firms. Indeed, every administration has introduced some targeting policy. Inspecting these outliers provides greater opportunities for safety gains. Once the risk information has been acquired, it is clearly desirable to use the data to target OSHA inspections. The OSHA procedure is not as sophisticated as it could be, however. From an economic standpoint, one would like to identify the risky outliers based on what is achievable within a particular context, which will depend on the costs of compliance for that industry. A procedure of targeting firms based on whether their record is better than the national manufacturing average does not incorporate this heterogeneity in the costs of promoting safety. A sawmill with an accident rate above the national manufacturing average may have a very safe technology for that industry, whereas a garment manufacturer with an injury rate just below the manufacturing average may be a high-risk outlier for that industry.

The changing character of the OSHA enforcement effort is exemplified as well by the change in the mix of violations cited by OSHA inspectors. Although the OSHA standards have not changed dramatically, the role of different violation categories has undergone many significant modifications. In OSHA’s initial years violations for walking and working surfaces (for example, misplaced exit signs) constituted about one-fifth of all violations. Many of these violations were for less important risks, some of which were readily visible
to workers as well. The substantial drop in this category suggests that OSHA’s resources have been redirected from a less profitable area.

The two categories that displayed the greatest relative increases are health-related. The role of health and environmental control (for example, noise, ventilation, and radiation) has risen, as have violations for toxic and hazardous substances (for example, asbestos and coke oven emissions). OSHA enforcement policies remain primarily safety related, but health hazards no longer constitute a trivial portion of the enforcement effort.

The Impact of OSHA Enforcement on Worker Safety

Firms will choose to make the necessary investments in health and safety if the OSHA enforcement policy in conjunction with market incentives for safety makes it in the firm’s financial self-interest to do so. More specifically, a firm will comply with an OSHA regulation if

\[
\text{Expected cost of compliance} < \text{Probability of inspection} \times \text{Expected no. of violations per inspection} \times \text{Average penalty per violation}
\]

This framework is of general value in assessing the desirability of regulatory compliance.

As discussed, the three links in establishing these incentives—inspections, violations, and penalties—are all relatively weak. A firm has about one chance in 100 of being inspected in any given year. If inspected, it expects to be found guilty of two violations of the standards, and for each violation the average penalty is under $700. Overall, the financial cost per worker is only $1.34. A useful comparison is that market forces, through compensating differentials, in combination with workers’ compensation premiums, imposed costs in excess of $1,234 per worker. Quite simply, OSHA’s enforcement effort is too modest to create truly effective financial incentives for safety.

The manner in which the safety incentives created by OSHA influence the decisions of firms can be seen by examining the payoffs from safety investments for a firm on a compliance–no compliance margin, which is illustrated in figure 23.4. The curve ABC gives the payoffs to the firm from different levels of safety investment in the absence of OSHA policy. If there were no government regulation, the firm would choose the optimal level of safety, which is that at \( s_0 \), because that point yields the highest payoff on the curve ABC. Suppose now that OSHA standards require a minimal safety level given by \( s^* \), which is above the level at \( s_0 \) shown on the diagram. For safety levels below \( s^* \), the firm will face an expected penalty level that will shift its curve downward, so that over what formerly was the payoff curve range \( AB \), the firm’s payoff levels now are given by \( DEF \). If the firm invests enough in safety to achieve a safety level equal to or in excess of \( s^* \), its payoff function will be given.
by BC as before. The real issue from the standpoint of compliance is whether the OSHA enforcement effort is sufficiently stringent so that the expected cost of noncompliance shifts the firm’s payoffs downward sufficiently to make compliance worthwhile. In this example, the highest payoff the firm can get from complying with the standard will be at point \( B \), whereas the highest payoff that the firm can get from noncompliance will be at point \( E \). For the case shown, the expected costs of noncompliance are sufficient to induce the firm to choose to invest in greater safety.

**OSHA Regulations in Different Situations**

The preceding analysis need not always be the case. In particular, as the three situations in figure 23.5 indicate, there are three broad classes of situations that can arise. In the first case, a firm initially has a payoff curve \( AA \), which is shifted downward to \( A'A' \) once OSHA penal-
ties are introduced for safety levels below $s^*$. That firm will continue to choose safety level $s_0$, because in this diagram the firm’s safety level is so far below that needed to achieve compliance that it would not be feasible or desirable for the firm to comply with the standard. In the intermediate case, the firm’s initial payoff curve is given by $BB$, which shifts downward to $B'B'$. For that firm, the standard will be sufficient to induce compliance, because the downward shift in payoffs introduced by the additional expected penalties has given the firm enough of a financial incentive to make additional investments in safety up to the standard worthwhile. The final situation, given by the payoff curve $CC$, represents a situation of a firm that already is in compliance with the standard, so that OSHA regulation will be irrelevant to its conduct.

What figure 23.5 illustrates is that in general, there will be three classes of firms. For two of these classes, firms that are already in compliance with the standard and firms that are substantially below the safety level required, one would expect little effect from the regulation. Effective enforcement in terms of raising the expected penalties for noncompliance is essential, inasmuch as that will drive the extent to which there is a substantial population of firms that will choose to comply with the regulation. Moreover, the level of stringency of the regulation is also of consequence, because if a regulation is very tight many firms will choose not to comply at all, so that a very tight regulation that is ignored may actually produce less of a beneficial safety effect than a more modest regulation for which compliance is feasible.\(^{25}\)

In terms of its regulatory strategy, OSHA in effect may have adopted a strategy that was doomed to fail—stringent regulations coupled with weak enforcement.

**OSHA and Other Factors Affecting Injuries**

Because of these limitations and the weakness of the OSHA enforcement effort, it is not surprising that OSHA has no dramatic effect on workplace safety. Not all workplace injuries are due to factors under OSHA’s influence. Many accidents stem from aspects of the work process other than the specific technological characteristics regulated by OSHA. That most workplace risks have not been readily amenable to the influence of OSHA regulations is in stark contrast to the optimistic projections of the framers of OSHA’s legislative mandate, who anticipated a 50 percent drop in workplace risks.\(^{26}\)

The chief contributing factor relates to worker actions. Although the estimates of the role of the worker in causing accidents vary, in part because of the difficulty in assigning accidents caused jointly by worker actions and technological deficiencies, it is clear that worker actions play a substantial role. OSHA found that over half of all fatal accidents on oil/gas well drilling rigs were caused by poor operating procedures, and worker actions also have been found to be a major contributor to 63 percent of the National Safety Council’s accident measure, 45 percent of Wisconsin workers’ compensation cases, and the majority of accidents among deep-sea divers in the North Sea.

Recent studies reinforce the view that at best OSHA regulations could have a significant but not dramatic effect on workplace safety. One statistical analysis estimated that if there were full compliance with OSHA standards, workplace accidents would drop by just under 10 percent.\(^{27}\) A detailed analysis of workplace accidents in California presented somewhat more optimistic conclusions. At most, 50 percent of all fatal accidents were contributed to by violations of OSHA standards that potentially could have been detected by an OSHA inspector visiting the day before the accident.\(^{28}\)

Because even a fully effective set of OSHA regulations would not revolutionize workplace safety, it is appropriate to take a more cautious view of the prospective effects of OSHA regulation than the original framers of the Occupational Safety and Health Act did. The critical economic issue is whether OSHA regulation has had any beneficial effect on safety. Agency officials at OSHA as well as at other safety-related agencies frequently point to improvements in accident rate trends as evidence of the efficacy of their agency. As recently as 2004, the official OSHA pronouncements linked safety improvements to the agency’s efforts: “Since

---

the agency was established in 1971, workplace fatalities have been cut in half and occupational injury and illness rates have declined 40 percent.”

There are two difficulties with linking these declines to agency efforts. First, there may be year-to-year changes in risk levels for reasons wholly unrelated to changes in safety standards or their enforcement, such as cyclical fluctuations. Second, there was a long-run trend toward safety improvements throughout the twentieth century as a result of the increased wealth of American society and the increased demand for safety that we placed on our social institutions. Thus, even in the absence of any government regulation, one would have expected a safety improvement as a result of society’s increased affluence.

The extent to which there has been such a safety trend is evidenced in the death rate statistics sketched in figure 23.6. The death risk for American workers dropped from 15.8

---


30. Note that the pronounced drop in fatality rates in 1992 was due to a shift in the database used to calculate the death rate and did not reflect an improvement in safety.
per 100,000 workers in 1928 to 6.8 per 100,000 workers in 1970, more than a 50 percent decline, and these improvements were achieved before the existence of OSHA. In the post-OSHA era these improvements have continued, as the risk level in 2002 was 1.7 deaths per 100,000 workers. The appropriate test of the agency’s effectiveness is whether OSHA has shifted this trend in any demonstrable fashion, controlling for determinants of accident rates other than occupational safety and health regulation.

The methodology for approaching this issue is illustrated in figure 23.7, which presents a stylized view of the statistical tests. The curve $AB$ represents the injury trend before the establishment of OSHA, and the curve $BC$ represents the predicted trend in injuries after OSHA, had there been no safety regulations in place. Similarly, the curve $BD$ represents the actual trend that injuries have had. If there is no statistically significant difference between the actual and predicted injury trend, then the agency has not had the intended effect. It is the vertical spread between $BC$ and $BD$ that represents the incremental effect of the agency on the injury rate, not the extent to which the injury level at point $D$ lies below point $B$ at the establishment of the agency, because much or all of the injury decline may have occurred in the absence of the agency.
Determining OSHA’s Impact on Safety

Two approaches can be used to ascertain the efficacy of the agency. Under the first, one estimates the equation to characterize the injury rate performance during a pre-OSHA era, which is the injury rate trend given by $AB$. One such model that could be used to estimate this relationship would be

$$
\text{Risk}_t = \alpha + \beta_1 \text{Risk}_{t-1} + \beta_2 \text{Cyclical effects},
+ \beta_3 \text{Industry characteristics},
+ \beta_4 \text{Worker characteristics}, + \varepsilon.
$$

The dependent variable in the analysis is the risk level in some year $t$, which will be determined by the series of variables on the right side of the equation. The risk level in the previous year is influential, inasmuch as it is a proxy for the character of the technology of the industry, and typically this lagged risk level has a positive effect. Cyclical effects are also pertinent because accident rates generally move procyclically as additional shifts of workers are added, new hires are added to the workforce, and the pace of work is increased. Industry characteristics are also consequential because the mix of industries in the economy and factors such as the presence of unionization may be influential. Principal worker characteristics include the experience mix of the workforce. Once an equation such as (23.1) has been estimated, one can use the fitted value of this equation to project out the accident rate trend $BC$ in figure 23.7, and this predicted trend can be compared with the actual observed risk levels along $BD$ to determine whether the agency has, in effect, shifted the injury rate downward.

The main prerequisite to using this postregulation simulation approach is that one must have a substantial period of preregulation data to do so. Although this is the case for overall accidental death statistics presented in figure 23.6, all data series gathered by the U.S. Bureau of Labor Statistics on an industry-specific basis changed after the advent of OSHA, so that no preregulation and postregulation comparison is possible. In this situation, one can adopt an alternative approach in which one estimates an equation using only data from the post-regulatory period. Equation 23.2 summarizes such a model:

$$
\text{Risk}_t = \alpha + \beta_1 \text{Risk}_{t-1} + \beta_2 \text{Cyclical effects},
+ \beta_3 \text{Industry characteristics},
+ \beta_4 \text{Worker characteristics}, + \beta_5 \sum_{i=0}^{n} \text{OSHA}_{t-i} + \varepsilon.
$$

This equation is the same as the preregulation simulation, equation 23.1, except for two differences. First, it will be estimated using the postregulation data rather than the preregulation data. Second, it includes variables that capture measures of the effect of the regulation, where the principal variables that have been used in the literature pertain to the rate of OSHA
inspections for the expected penalty level. The equation indicated that a distributed lag on
the OSHA variable has been included to recognize the fact that it may take some time for an
OSHA inspection to have an effect. Health and safety investments that require capital invest-
ments on the part of the firm take some delay before they can be made, so that it would be
unrealistic to assume that OSHA regulations will always have a contemporaneous effect. In
practice, most studies have indicated that to the extent that OSHA has an effect, it is with a
one-year lag.

Mixed Opinions Regarding OSHA’s Impact

The general consensus of the econometric studies is that there is no evidence of a substan-
tial impact of OSHA. Viscusi analyzed the 1972–1975 period and failed to find any signifi-
cant OSHA impact. Smith found a drop in the lost-workday rate at firms inspected in 1973,
but not for firms inspected in 1974. A replication of Smith’s analysis by McCaffrey for the
1976–1978 period failed to yield any significant effects on manufacturing firms. Similarly,
Bartel and Thomas’s analysis of the 1974–1978 experience did not reveal any significant
OSHA impacts. Mendeloff’s analysis of the California workers’ compensation records from
1947 to 1974 likewise produced mixed results, as some risk levels rose and others declined.
Perhaps the strongest published evidence of OSHA’s efficacy is by Cooke and Gautschi, who
found a significant drop in lost workdays because of accidents in Maine manufacturing firms
from 1970 to 1976. A more recent study by Viscusi of the 1973–1983 period suggested that
there may have been a modest decline in the rate of accidents as a result of OSHA enforce-
ment, although this effect was not great. The greatest effect appears to be in the most severe
category of injuries, for OSHA regulations have reduced the total number of lost workdays
due to injuries by 5 to 6 percent. The OSHA policy variable that seemed most instrumental
was the rate of OSHA inspections rather than the penalty level. The degree of OSHA’s pres-
ence in the workplace and the threat of penalties for continued noncompliance appear to be
the fundamental determinants of OSHA’s impact.

Resources 14 (1979): 145–70.
34. Bartel and Thomas, “Direct and Indirect Effects of Regulation.”
35. Mendeloff, Regulating Safety.
Ruser and Smith similarly found that the recent OSHA efforts had a beneficial effect. They estimate that the records-check inspections of the early 1980s decreased injuries by 5 to 14 percent.38 Evidence reported by Scholz and Gray for 1979–1985 also indicated significant OSHA effects.39

The possibility of a favorable impact of OSHA on workplace conditions is also borne out in more refined studies of workplace standards. One case study is that of the OSHA cotton dust standard, which was the subject of a Supreme Court decision. That standard was directed at controlling cotton dust exposures in the workplace, because these exposures lead to potentially disabling lung diseases. The promulgation of such a regulation was viewed by the business world as a dramatic event, as there were severe stock market repercussions of the various regulatory events that were involved in the issuance of the regulation. Overall, an event study analysis indicates that the market value of the cotton firm fell by 23 percent in response to the cotton dust standard.40 If it was expected that firms would be able to completely ignore the regulation, then no market effect would have been observed. A major reason for the expected compliance with the regulation was that the controversy surrounding it as well as the vigorous action by the union involved in this particular instance ensured that the cotton dust standard would be a prominent target of OSHA enforcement.

Although compliance with the cotton dust standard was not required until 1984, by the end of 1982 the majority of the exposed workers were in work situations in compliance with OSHA standards. Firms’ investments in cotton dust controls from 1978 to 1982 will lead to an annual reduction of about 6,000 cases of byssinosis (a lung disease) annually. The standard remains controversial, however, because it is a costly means for promoting worker health. For example, the cost per case-year of total disability prevented has been estimated at $1.2 million.41

In addition, there remain a number of advocates of the greater use of more performance-oriented alternatives to control cotton. One possible policy alternative is to require the use of lightweight dust masks for low to moderate cotton dust levels, which would produce the same benefits as engineering controls at negligible cost. Because byssinosis is a progressive disease that moves through a series of grades and is reversible in its early stages, disposable masks could be coupled with a worker rotation policy. Only for severe cotton dust

levels would respirators or engineering controls be required. To date, protective equipment alternatives have not been treated as a viable policy option because of union opposition to such efforts.

The available empirical results for the overall OSHA impact and in the cotton dust case suggest that OSHA enforcement efforts may be beginning to enhance workplace safety. An improvement over the early OSHA experience should be expected, as the standards have been refined and there is more systematic targeting of the inspection effort.

The Role of Workers’ Compensation

There is a tendency to think of OSHA as establishing the incentives for safety and to think of workers’ compensation as simply a social insurance mechanism that reduces the litigation costs that would arise if workers could sue their employers for job-related injuries. However, the financing mechanisms for workers’ compensation create very powerful incentives for safety. These incentives will be particularly great for larger firms, for whom their injury experience will have an instrumental role in affecting their insurance rates. Moreover, even small firms are subject to inspections by insurance underwriters, who have a strong financial incentive to price the insurance appropriately.

The role of workers’ compensation soared in the latter part of the twentieth century. Workers’ compensation premiums were just over $15 billion annually in 1984, but more than doubled to $31.3 billion by 1991. As a result of both legal reforms and increased efforts to promote workplace safety, workers’ compensation premiums had fallen to $26.0 billion annually by 2001. The linkage of these insurance premium costs to firms’ accident performance can potentially create extremely powerful incentives for safety—far greater than those generated by the comparatively negligible OSHA sanctions.

Empirical evidence on the incentive effects of workers’ compensation has always been contaminated by the presence of moral hazard effects. Even if higher workers’ compensation premiums foster a safer workplace, the increased generosity of workers’ compensation benefits will increase the incentives of workers to report accidents. Indeed, some researchers have speculated that workers may misrepresent accidents, such as lower back injuries, in order to collect such benefits. The net result is that most studies of the link between workers’ compensation and nonfatal accidents suggest that there is a positive correlation, which is the opposite of what one would expect if there were safety incentive effects.

Matters are, however, quite different in the case of fatalities. One cannot misrepresent one’s death, and the usual moral hazard arguments do not apply. As a result, the empirical evidence presented by Moore and Viscusi indicates that had it not been for the existence of workers’ compensation, job fatality rates in the United States would be as much as one-third greater
than they are. This estimate of the safety consequences of workers’ compensation is roughly an order of magnitude greater than the consensus estimates of the effect of OSHA safety regulations.

The importance of financial incentives for safety is not surprising, but recognition of the role of such incentives could transform the approach that policymakers take toward fostering greater safety. For example, an injury tax levied by OSHA based on the accident level at the firm could serve a similar function. Cornell University economist Robert S. Smith has advocated this approach since OSHA’s inception. Moreover, it would greatly increase the scope of OSHA’s regulatory efforts, enabling it to focus its efforts on the less readily monitorable hazards, rather than on errors that will be captured in the accident statistics and can be handled through the injury tax mechanism.

Economic analysis is also instructive in guiding policymakers in setting the level of social insurance benefits. Workers will value workers’ compensation and, as a consequence, will be willing to accept lower wages in return for insurance coverage of their job accidents. Indeed, workers’ compensation now pays for itself through such wage offsets. By comparing this wage offset with the cost of providing the insurance one can ascertain whether the level of earnings replacement under workers’ compensation is optimal from an insurance standpoint. Evidence on these wage-workers’ compensation offset levels suggests that current benefit levels are reasonable. In recent years there has been considerable concern about the growth in workers’ compensation. While employers often complain about rising workers’ compensation costs, in most instances it is the workers themselves who are paying for those costs through reduced wages. However, there is no evidence that the level of insurance coverage is overly generous. Rather, the main problems seem to stem from an extension of coverage of workers’ compensation to new classes of injuries, such as carpal tunnel syndrome.

**Agenda for Policy Reform Efforts**

Even with a reform of its policies, OSHA will not be the dominant force influencing worker safety. The role of the market in determining safety will continue to be instrumental. OSHA can augment the existing forces for safety, but even full compliance with all current OSHA regulations or those likely to be promulgated will not markedly reduce workplace risks. The no-risk society that some might envision as OSHA’s ultimate goal is simply unattainable.


Nevertheless, constructive reform of OSHA could enable this agency better to foster the interests of workers and at the same time diminish the associated burden on society. A number of specific reforms have been advocated in the literature. Rather than review each of these proposals, the following focuses on changes for which there is likely to be a broad consensus about the nature of OSHA’s inadequacy or the proposed remedy.

The first area of proposed reform concerns the area of emphasis. In more than two decades of regulation, OSHA policies have exhibited a slight shift toward health but have remained largely safety oriented. The emphasis of both the structure of new regulations and OSHA enforcement has continued to be predominantly in the safety area. This emphasis is misplaced, for market forces are better equipped to address safety risks through compensating differentials and related mechanisms. In addition, the incentives created by workers’ compensation premiums already augment to some extent the market incentives for safety. Health hazards are handled less adequately by both the market and workers’ compensation. Moreover, the coupling of substantial uncertainties with low-probability events involving potentially catastrophic outcomes makes health risks a promising target for governmental regulation.

An increased emphasis on health risks is the cornerstone of a policy reform proposal by economists Thomas J. Kniesner and John Leeth. They propose that more resources be given to the National Institute of Occupational Safety and Health to research and publicize health risks. Their proposal similarly would increase financial incentives for safety, such as those provided by workers’ compensation, and decrease the emphasis on inspections, particularly for firms with strong safety records.

A second class of reforms is to ensure that we are “on the frontier” of efficient policies; that is, that we are achieving as much health and safety improvement as possible for the costs imposed. Much of the adverse reaction to OSHA’s initial wave of regulations of toilet-seat shapes and the like stemmed largely from the belief that the regulatory mechanisms had not been well chosen. Strong enforcement of regulations will only be warranted for regulations that are sensible.

Some of the most extraneous features of OSHA policy have been pruned, but there is continued need in all regulatory contexts to find ways to promote safety at less cost. The use of performance standards rather than narrowly defined specification standards could, for example, enable firms to select the cheapest means of achieving the health and safety objective. Such flexibility would reduce compliance costs and increase the incentive of firms to develop innovative technologies to foster health and safety. Moreover, if structured appropriately, as in the grain dust standard, a performance standard need not greatly increase firms’ uncertainty whether they are in compliance.

The final reform target is to strike a more explicit balance between the health improvements and the costs imposed on society. Labor market estimates of the value of statistical life are now being used to provide guidance in terms of the appropriate trade-off. If policymakers viewed regulatory alternatives in light of the cost per health benefit achieved, they would at least confront explicitly the nature of the trade-offs and ideally would pursue only those policies that they judged to be in society’s best interest.

Although reforming OSHA’s regulatory strategy remains a major item on any agenda of important regulatory reforms, it would be an oversimplification to say that OSHA has not improved its efforts. The agency has introduced several promising new regulations, has eliminated some of the worst initial regulations, and has better targeted enforcement efforts than they once did.

The future of OSHA policies no doubt will continue to exhibit the need for reflecting the three classes of reform elements that we have suggested, for they are at the heart of any regulatory strategy for workplace health and safety. As a result, complete regulatory reform will never be achieved with the same finality as economic regulation, where, for example, deregulation has transformed the airline industry into what some observers consider to be a more competitive situation. The need in the health and safety area is for better regulation, not deregulation, and opportunities for improvement will always remain.

Questions and Problems

1. What are the rationales for occupational safety and health regulation? Does the existence of compensating differentials for risk imply that there is no rationale for regulation? What if we also knew that workers had perfect information regarding the risks they faced and markets worked competitively? Could you think of any other possible rationale for intervention?

2. OSHA inspectors could guarantee compliance by imposing infinite penalty amounts on firms that did not comply with their regulations. If arbitrarily large penalties were permitted by OSHA’s legislation, would it be desirable to adopt such penalties? What are the factors that you would want to consider in establishing the penalty level?

3. Suppose that a technological innovation has made it easier for firms to provide a safe work environment. How would you illustrate this effect using the diagram in figure 23.2? Can we tell whether safety expenditures will rise or fall after such a shift?

4. OSHA and other regulatory agencies have typically followed a specification standard approach rather than a performance-oriented approach. What are the considerations that make a technology orientation attractive to government officials even though it has not found great favor among economists?

5. When setting the optimal penalty level for noncompliance for the regulation, should the regulatory agency vary the penalty with firm size? Should the profitability of the company be a concern?

6. Suppose that there are different types of firms in the industry, old firms and new firms. Suppose that old firms have existing technologies for which it is more costly to adopt risk-reducing
innovations, whereas new firms can incorporate these innovations in their new plant investments. Use a variant of figure 23.2 to illustrate how the optimal safety level will differ in these two different situations. Should there be heterogeneity in the standards set by regulatory agencies?

7. Economists frequently advocate the use of personal protective equipment, such as gas masks and ear muffs, as less costly solutions for promoting worker safety. These cost considerations typically focus only on the purchase cost of the equipment. What other cost components are associated with personal protective equipment that might make engineering controls a more attractive alternative?

8. New information becomes available pertaining to potential safety innovations on a very regular basis. There are always new potential engineering controls that could be adopted by OSHA. Yet the agency tends to vary its standards very little over time. Can you think of any economic rationales for having a relatively stable regulatory regime even in the presence of technological changes that might enhance safety?
Patents and Pharmaceuticals

In chapter 4 we discussed the importance of technical progress (or dynamic efficiency) in comparison to static efficiency. It was observed that an economy may have to tolerate market power that stimulates or results from technical change if a rapidly rising standard of living is desired. In this chapter, we focus on the economics of patents and their role in one of the most technologically progressive U.S. industries: pharmaceuticals. It is generally believed that patents are more important in pharmaceuticals and chemicals than in any other industries. Pharmaceuticals is also an industry in which there is recurring controversy about this very trade-off. Some argue that its profits and prices are excessive, and others warn that policies to curb prices will likely harm the innovativeness of one of America’s most progressive industries.1

There are two main sections in this chapter. In the first section, the economics of patents in providing incentives for innovation will be discussed. Both positive and normative lines of analysis will be pursued. It should not be surprising that a definitive set of results does not exist. For example, there are some market factors that make for too little investment in inventive activities, and others that make for too much—from a welfare point of view. Nevertheless, economic analysis can be quite instructive for policy purposes in certain cases.

The second half of this chapter is a case study of the U.S. pharmaceutical industry with a particular emphasis on the role of patents. A principal topic is an examination of a 1984 law that increased the life of patents in the pharmaceutical industry while simultaneously easing entry conditions faced by imitators. Also, the safety of pharmaceuticals is regulated by the Food and Drug Administration, as discussed in chapter 22.

Economics of Invention and Patents

In an important article in 1962, Nobel laureate Kenneth Arrow explored the problems created for the market by inventive activity. Arrow observed that the product of inventive activity is new knowledge, or information. For example, after spending $200 million on research, a pharmaceutical firm might come up with the chemical structure of an important new drug. In principle, the information is extremely valuable and may be worth more than $200 million in terms of future revenues. However, the point here is that the product of the research and development (R&D) is simply knowledge of the chemical structure. This knowledge, or information, can often be described completely in a short document of five to ten pages!

As Arrow has explained,

Information is a commodity with peculiar attributes, particularly embarrassing for the achievement of optimal allocation. In the first place, any information obtained, say a new method of production, should, from the welfare point of view, be available free of charge (apart from the cost of transmitting information). This insures optimal utilization of the information but of course provides no incentive for investment in research. . . . In a free enterprise economy, inventive activity is supported by using the invention to create property rights; precisely to the extent that it is successful, there is an underutilization of the information.\(^2\)

One difficulty is the problem that the inventor has in appropriating the economic value of the invention. Because what is possessed is information, the problem is how to sell the information. Suppose the inventor discovers an important drug, Panacea. The inventor could keep the chemical structure secret and try selling the drug as a cure for certain diseases. But a rival could easily buy a few pills, hire a chemist to figure out the structure, and begin selling exact copies at a lower price. In such a case, the inventor would not appropriate all of the economic benefits of the invention; rivals would share in the rewards although having invested very little. From a social point of view, this situation would result in too little investment in inventive activity.

Of course, if the inventor could obtain a legal right to exclusive use of Panacea—a patent—then the problem is partially resolved. The patent gives the inventor property rights to the invention for a fixed period of time. But, as Arrow has described, patents cannot solve all problems of appropriability:

Suppose, as a result of elaborate tests, some metal is discovered to have a desirable property, say resistance to high heat. Then of course every use of the metal for which this property is relevant would also use this information, and the user would be made to pay for it. But, even more, if another inventor is stimulated to examine chemically related metals for heat resistance, he is using the information already discovered and should pay for it in some measure; and any beneficiary of his discoveries should also pay.\(^3\)

If the inventor cannot expect to appropriate all of the economic value of the invention, there will be underinvestment in inventive activity. The real quandary, though, is raised when we consider the optimal use of the information. Now the argument is that the price of the information should be zero, implying that the inventor would appropriate none of the economic value of the invention!

The argument that the price of the information should be zero follows from the fact that its marginal cost is zero. Once the information exists, any number of people can “consume” it without any cost to anyone else. Anyone who would derive any benefit whatsoever from the information should be permitted to use it freely; if not, the information would not be used efficiently.


\(^3\) Ibid., p. 617.
Clearly, firms would not invest in inventions without expectations of rewards. Arrow suggested one way out of the dilemma: “In an ideal socialistic economy, the reward for invention would be completely separated from any charge to the users of the information.”\(^4\) That is, “prizes” could be given by the government to successful inventors, with the understanding that the inventions would be freely available to all. In response to this suggestion, Harold Demsetz observed that there would be serious problems in implementing a system of prizes. “How would such a system produce information on the desired direction of investment and on the quantities of resources that should be committed to invention?”\(^5\) Similar problems would exist if the government implemented a system of subsidizing or contracting with firms to undertake research deemed worthy by the government.

Patents, then, can be regarded as one way of achieving a balance between appropriability and use. On the one hand, a long patent life (by giving the inventor a monopoly of long duration) favors appropriability at the expense of use. That is, optimal use is not achieved during the patent life because of pricing above marginal cost. On the other hand, a short patent life favors use at the expense of appropriability—with the result being levels of investment that are too low. Hence, a “second-best” patent life lies somewhere in between. Later in this section we will consider a model of the optimal patent life in detail.

**Background on Patents**

A patent is an exclusive right to one’s invention. In the United States, the right lasts for twenty years. Basically, either products or processes can be patented; an idea itself cannot be patented unless it is applied. To obtain a patent, the inventor must make an application to the U.S. Patent and Trademark Office. The Patent Office must be satisfied that the invention is new, useful, and nonobvious. The twenty-year life begins when the patent application is made. In most other countries the patent life is twenty years also.

The idea of patents is quite old; the first patent law was adopted by the Republic of Venice in 1474, and the first U.S. patent statute was enacted in 1790. The usual rationale for the patent includes the belief that the inventor is entitled to his or her discovery, that the patent is a device for promoting invention, and that the patent system encourages inventors to disclose their inventions to others. The granting of a patent in itself does not ensure the inventor exclusive rights. Rather, the inventor must bring suit against anyone who infringes the patent, and the courts then make the final determination of the validity of the patent.

Although patents are technically issued only to individuals, many large corporations engaged in R&D require their employees to assign the right to any invention that they make.

---

4. Ibid., p. 623.
to the company. Less than a quarter of the patents issued today are assigned to individual inventors.

The holder of a patent may either make sole use of the discovery or license others to use the invention at a mutually agreed-on royalty rate. For example, General Electric once licensed Westinghouse to produce electric lamps at a royalty rate of 2 percent of Westinghouse’s sales revenues. The rate jumped up to 30 percent once a certain level of revenues was reached.

Incentives to Invent: Monopoly versus Competition

In this section we consider the following limited question. Assume that an industry can be organized either competitively or as a monopoly. In either case, assume that a single inventor is considering investing in R&D in order to achieve a cost-reducing invention of a particular size. The inventor is not concerned about competition from other inventors, and complete protection from imitation is assumed. In the competitive case, the inventor has an infinitely lived patent, and in the monopoly case, the inventor is the monopolist and entry is completely barred.

Minor Invention Case

Figure 24.1 shows both the competitive industry and the monopoly for the case of a minor invention. That is, the original equilibria for both cases are based on a constant cost of production $C_0$ and the demand $DD'$. Hence, the competitive industry equilibrium before the invention is at price $P_0$ and quantity $Q_0$, where demand and the constant cost supply curve intersect. The original monopoly equilibrium is determined by the intersection of marginal revenue, $DJ$, and marginal cost (constant at $C_0$), or at the quantity $M_0$, yielding price $P_m$.

First, we focus on the incentive to the inventor in the competitive industry. What royalty rate (expressed in dollars per unit output) would maximize the inventor’s total royalty if the invention lowers cost from $C_0$ to $C_i$? Analytically, we can proceed in either of two equivalent ways. The derived demand for the patent could be determined, or the inventor could be

6. This section is based on the analysis in Arrow, “Economic Welfare.”
7. There are two types of inventions, cost-reducing processes and new products. Here we focus on inventions that lower costs. Although a bit strained, the same analysis can be used for new products. Imagine that the demand exists but is everywhere below the current cost of production, so quantity purchased is zero. Now a cost-reducing invention lowers cost such that positive quantities are purchased—and we have a new product!
8. A minor invention reduces cost by a relatively small amount. The exact difference between a minor and a major invention will become clear later in this section.
9. The inventor would then equate the marginal cost of the patent (zero) with the marginal revenue, thereby determining the optimal royalty rate to charge users of the lower cost process.
assumed to monopolize the industry (because the inventor alone has access to the lower cost process). Because the second method is analytically a bit simpler, we will use it.

The monopolist-inventor in the competitive industry would face a kinked demand curve, $P_0AD'$. The price could not be set above $C_0$ or the existing firms would find it profitable to compete. Thus the maximum price for output levels up to $Q_0$ would be $P_0$ (or just a bit below), and above that output level the market demand curve would be the relevant demand. The marginal revenue curve would be $P_0AHJ$, with the usual vertical discontinuity at the kink. Hence, the monopolist-inventor would choose price $P_0$ and output $Q_0$, because marginal revenue intersects marginal cost ($C_1$) at this output.

We pause for a moment to distinguish exactly the difference between a minor invention and a major invention. Notice in figure 24.1 that the marginal cost $C_1$ lies within the “gap” $AH$ of the marginal revenue curve. This ensures that the quantity $Q_0$ remains unchanged after
the invention. However, if the marginal cost $C_1$ should be so low as to intersect the $HJ$ segment of the marginal revenue curve, then the monopolist-inventor’s quantity would be larger than $Q_0$, leading to a price decrease as a result of the invention. Large cost reductions of this sort that induce price reductions are termed major inventions. For minor inventions, market price is unaffected. (A major invention can also be defined as one that makes the inventor’s monopoly price below the original marginal cost.)

Given the inventor-monopolist’s equilibrium as explained, the inventor’s profit in the competitive industry case is therefore the rectangle equal to the cost saving per unit ($C_0 - C_1$) multiplied by the output level $Q_0$. Or, the incentive in the competitive case is the sum of the two shaded areas I and II in figure 24.1.

We now consider the case in which the industry is organized initially as a monopoly. Originally, the monopoly would charge a price of $P_m$, as noted earlier. At this price the monopoly profit is the triangular area $DEP_0$. It equals the area under marginal revenue (or total revenue) less the area under marginal cost (or total cost).\(^\text{10}\) The monopolist’s incentive to invest in a cost-reducing invention is simply its increment to profit due to the lower cost process. It is easy to show that this is the trapezoid $P_0EFG$ in figure 24.1, or area I. The reason is that profit with the lower cost process increases from $DEP_0$ to $DFG$, and the difference is $P_0EFG$.

The key conclusion is that the incentive to invent in the competitive industry case is the sum of areas I and II, while in the monopoly case it is only area I. Hence, for the minor invention case, the incentive is greater if the industry is organized competitively.

Before examining the same question in the case of a major invention, it is useful to consider the “first-best” social benefit of the cost-reducing invention in figure 24.1. If the lower cost process were made available to firms at the efficient price of zero, the competitive equilibrium would change to a price equal to $C_1$ and a quantity of $Q_1$. The social benefit would then be equal to the sum of areas I, II, and III, the increase in consumer surplus due to the price decrease. The ranking is therefore that the social benefit ($I + II + III$) exceeds the incentive in competition ($I + II$), which in turn exceeds the incentive in monopoly ($I$).

**Major Invention Case**

Figure 24.2 shows the case of a major cost-reducing invention. As we explained earlier, the major invention leads to a price decrease after the invention, unlike the minor invention case.

Before invention, the two equilibria are exactly as in the minor invention case. Competition has price $P_0$ and quantity $Q_0$, while monopoly has price $P_m$ and quantity $M_0$. After invention, both the inventor-monopolist in the competitive industry and the monopolist

\(^{10}\) For a more detailed explanation of this geometrical method for measuring profit, see the section in chapter 8 on successive monopolies.
choose price $P_m'$ and quantity $M_1$. The inventor-monopolist in the competitive industry therefore obtains a profit incentive equal to the large shaded rectangle, $P_m'SVW$.

To find the profit increase for the monopolist, and therefore its incentive, simply subtract the preinvention profit, which equals the small shaded rectangle $P_m'RTP_0$, from the large shaded rectangle. The comparison of incentives is therefore clear. It is again the case that the incentive is greater in the competitive industry than in the monopoly—the monopolist has a preincentive profit that must be subtracted from the large rectangle, whereas the inventor in the competitive industry does not need to subtract anything.

Tirole has described this lesser incentive in monopoly as the *replacement effect*. “The monopolist gains less from innovating than does a competitive firm because the monopolist ‘replaces himself’ when he innovates whereas the competitive firm becomes a monopoly.”

Welfare Analysis of Patents

Earlier we noted that patent life could be too long or too short from a social viewpoint. Longer life increases the inventor’s appropriability at the cost of a longer period of monopoly pricing. A shorter life reduces appropriability but brings about efficient pricing sooner. The implication is that an optimal patent life lies somewhere in between the extremes. In this section, we describe an optimal patent life model developed by W. D. Nordhaus. This section also considers some complications regarding the Nordhaus model—competitive patent races and new product inventions when close substitutes exist.

Optimal Patent Life Model

Nordhaus assumed the case of a single inventor in a competitive industry who makes a minor cost-reducing invention. Hence, the model builds directly on the development of that case earlier in this chapter.

The Inventor’s Equilibrium

In the analysis relating to figure 24.1, the magnitude of the cost savings per unit \((C_0 - C_1)\) was taken as given. Here, we shall examine what determines the magnitude of the cost savings, or what can be termed the size of the invention. The magnitude of the cost savings will be referred to as \(B\).

Although one of the major characteristics of inventive activities is the uncertainty of the outcome, the model rules out uncertainty. Hence, it is assumed that the inventor has a total cost of R&D function, \(TC\), that gives \(TC\) as a function of the cost savings, \(B\). In Nordhaus’s book, he justified this assumption by observing that “if there is no relationship between pecuniary rewards and inventive inputs on the one side and inventive output on the other, the optimal life is zero.”

Hence, assume that the inventor’s total cost of R&D, \(TC\), is a quadratic function of \(B\). This assumption means that there are diminishing returns to R&D as the size of the invention increases. It is

\[
\text{be argued that the incumbent monopolist has a greater incentive to invent than a rival inventor who is a potential entrant into the industry. That is, if the rival inventor is successful and develops the low-cost process } C_1 \text{ first, a Cournot duopoly would result between the incumbent with cost } C_0 \text{ and the inventor. Hence, the incumbent monopolist’s incentive is the difference between winning the race and obtaining monopoly profit with } C_1 \text{ and its duopoly profit with } C_0 \text{. This is larger than the entrant’s duopoly profit. For a specific example, see problem 4 at the end of this chapter. Also, see R. Gilbert and D. Newbery, “Pre-emptive Patenting and the Persistence of Monopoly,” American Economic Review (June 1982).}
\]

13. The use of this cost function is based on exercise 10.4 in Tirole, *The Theory of Industrial Organization*.
where $TC$ is the total cost of R&D, $B$ is cost savings ($C_0 - C_1$), and $\alpha$ is a positive constant.

The problem is now easily formulated. The inventor must choose the amount of cost savings, $B$, that will maximize the difference between the present value of the stream of royalties, $PV$, and the R&D cost, $TC$. Recall that the royalty (or profit), as determined earlier for the minor invention case, is the rectangle equal to areas I and II in figure 24.1. The rectangle is also simply $BQ_0$. An important parameter is $T$, the patent life, because the stream of royalties will cease after year $T$. It is easy to show that the present value of the stream of royalties from the present to year $T$ is:

$$PV = BQ_0 (1 - e^{-rt}) / r$$

(24.2)

where $PV$ is the present value of royalties, $B$ is cost savings, $r$ is the inventor’s interest rate, and $T$ is patent life.

Figure 24.3 shows both $PV$ and $TC$ as functions of $B$. Notice that $PV$ is just a straight line from the origin with its slope dependent on the value of the patent life, $T$. Two patent lives are assumed in the figure: $T = 10$ years and $T = 20$ years. The inventor will choose the value of $B$ corresponding to the largest vertical distance between $PV$ and $TC$, or where the slopes of the two functions are equal. Hence, if $T = 10$ years, the inventor will choose to have a cost-saving invention of magnitude $B_{10}$; if a longer life of $T = 20$ years is in effect, the inventor will choose the larger invention $B_{20}$. The result is clear: the longer the patent life, the larger the cost-saving invention.16

Determination of Optimal Patent Life

Because the size of the invention, $B$, is positively related to $T$, the patent life, why shouldn’t $T$ be set at an infinitely large number? The reason it should not be is the need to balance off a larger invention against the inefficiency of monopoly pricing. As discussed earlier, the price of new knowledge should be set at zero if it is to be used efficiently—but, of course, this practice would provide no incentive to the inventor.

In order to understand the trade-off as $T$ varies, consider figure 24.4. Figure 24.4(a) shows the benefits of the invention to be the rectangle, area I, for the period of the patent life. Area

---

15. Evaluate the integral $\int BQ_0 e^{-rt} dt$ between $t = 0$ and $t = T$. The integral simply adds up the flows of royalties over the period, properly discounted.

16. In the case here of a single inventor, this result is straightforward. However, other models have been developed in which there is competition among firms to lower costs. In these models, information “spillovers” among the firms increase the efficiency of the industry’s cost-reducing efforts. Larger spillovers, of course, mean that appropriability is lower. Hence, in these models innovative output may decrease with higher appropriability (lower spillovers). See, for example, M. Spence, “Cost Reduction, Competition, and Industrial Performance,” *Econometrica* Vol. 52(1) (January 1984): 101–122.
I represents the total cost savings in real resources, which is captured in the form of a royalty by the inventor. After the patent expires, price falls to $P_1$ and output increases to $Q_1$, as shown in figure 24.4(b). The benefits are now shown as two areas: area I is the same as in figure 24.4(a), though it is now part of an enlarged consumers’ surplus, and area II is the gain of the former deadweight loss triangle due to the removal of monopoly.

Now as $T$ increases, the size of the invention increases and $C_1$ shifts downward in both panels of figure 24.4. This shift increases area I benefits, but area II is put off further into the future, thereby reducing the present value of the stream of area II benefits.

The actual derivation of the optimal patent life requires one to choose the value of $T$ that maximizes the present value of areas I and II, less the total R&D cost, $TC$. Because the mathematical analysis becomes rather complex, we shall merely indicate the result as the intersection of a marginal benefit curve with a marginal cost curve, as shown in figure 24.5. The optimal patent life is therefore $T^\ast$.

The marginal benefit curve is a function of the patent life, and it declines as the life increases. The marginal benefit from an additional year of patent life is the gain in social

\[ $\begin{align*}
$ $\end{align*}$\]
welfare generated by the ensuing larger size of invention. For convenience of exposition, it is net of the additional R&D cost. The decline is primarily because of diminishing returns to R&D investment.

The marginal cost in figure 24.5 is really a marginal opportunity cost. It is the loss of area II in figure 24.4 for an additional year of patent life. For example, as the patent life increases from $T'$ to $T' + 1$, consumers forgo the deadweight loss triangle that they would have attained in $T' + 1$. It also declines with patent life, but not as steeply as the marginal benefit curve. Marginal cost declines because the lost triangle is discounted more heavily as the life increases.

One simple result that the model predicts is that the optimal patent life should vary with the industry’s demand elasticity. For example, higher elasticities of demand imply that area II in figure 24.4 will be larger. This implication, in turn, suggests that the primary effect will be to shift the marginal cost curve in figure 24.5 upward (as indicated by the dashed curve). The result is an optimal patent life that is smaller.¹⁷

Complications

Of course, there are a number of strong assumptions underlying this optimal patent life model. We will consider the relaxation of two important ones here. The first is the restriction of the

analysis to a single inventor un Concerned about rival inventors. Recent work has developed the concept of competition among inventors for the patent—so-called patent races. The second assumption is that the invention is cost reducing. We will examine the welfare economics of a new product invention that has a close substitute that already exists.\textsuperscript{18}

\textbf{Patent Races}

Unlike the model of a single inventor, there are now numerous potential inventor/firms all seeking a particular invention. The winner gets the patent and the others get nothing. For simplicity, it is assumed that there is a social benefit of amount $B$ that also equals the private benefit to the inventor. Problems of monopoly pricing are therefore not considered here.\textsuperscript{19}

It is useful to conceive of a “discovery function” that gives the probability of the invention being made and the patent awarded. Assume that all inventors are equal in size—each must commit an R&D investment of amount $R$ up front if it decides to join in the race. Hence,

\textsuperscript{18} Of course, this issue has been examined to some extent earlier in the book in chapter 6 under the heading Preemption and Brand Proliferation.

we indicate the discovery function as \( P(n) \), where \( P \) is the probability of discovery and \( n \) is the number of inventor/firms. As the number of firms increases, \( P \) rises but at a decreasing rate, approaching \( P = 1 \).

There are two problems to be solved. First, what is the optimum number of firms from society’s viewpoint? Second, what is the number of firms that will engage in the race in a competitive (or free-entry) equilibrium? Looking ahead, it will be shown that there are too many firms in a competitive equilibrium as compared to the social optimum. That is, the model as structured here leads to overinvestment in R&D. We shall explain the rationale for overinvestment as simply a variation of the well-known common pool problem in economics. Next, we consider whether the model’s assumptions are likely to describe the real world.

The social welfare problem is to find the number of firms, \( n \), that maximizes the expected value of social benefit less social cost. In a statistical sense, the expected value of benefit is the “average” benefit if the race were repeated over and over; it is just the probability of discovery multiplied by the benefit amount, or \( P(n)B \). The social objective function is then \( P(n)B - nR \), where \( nR \) is social cost, or just the number of firms times the cost per firm.

Solving this maximization problem by differentiating the social objective function with respect to \( n \) gives the marginal condition that marginal social benefit, \( MSB \), should equal marginal cost, \( MC \). Or,

\[
P'(n)B = MSB = MC = R.
\] (24.3)

In words, \( P'(n)B \) is the marginal increase in probability of discovery brought about by adding another firm, times \( B \). It is the expected increase in social benefit due to one more firm joining the race—and, at the social optimum, it should equal the cost of that firm, \( R \).

The socially optimum number of firms, \( n^* \), is shown in figure 24.6 as determined by the intersection of the \( MSB \) curve and \( MC \). The fact that \( MSB \) declines as \( n \) increases is inherent in the assumption about the shape of the discovery function made earlier. The probability of discovery is assumed to increase with \( n \) but the increases \([P'(n)]\) become smaller and smaller.

Next, consider the competitive equilibrium. A firm must decide whether to enter the race by comparing its expected profit, \( EP \), with its cost \( R \). As long as \( EP \) is greater than \( R \), firms will continue to join the patent race. Hence, the number of firms is determined by the condition of zero profit at the margin, or by \( EP = R \). The expected profit can also be written as \( P(n)B/n \). That is, each firm would believe that it is equally as likely to win the expected prize of \( P(n)B \) as any other firm—so its probability is just \( 1/n \). Hence, its expected profit is \( 1/n \) times \( P(n)B \). Or,

\[
P(n)B/n = EP = MC = R.
\] (24.4)

The \( EP \) equal to \( R \) condition is shown in figure 24.6. It is also possible to interpret \( EP \) as the average social product, \( ASP \). That is, \( ASP \) is the expected social product divided by \( n \).
This interpretation makes it easy to understand why it lies above $MSB$ in the figure: the two curves stand in the usual relationship of average and marginal curves. As the average social product declines, the marginal must lie below it. The competitive equilibrium leads to $n'$ firms, a number that exceeds the optimum number, $n^*$. Just as competitive fishing in a lake can result in overfishing, or competitive drilling of oil from a single pool can lead to too much drilling, our assumptions here lead to too much R&D. The explanation for this result is that there is one “production function” that links the firms in these cases—here it is the discovery function. Each firm affects the others directly through this function but ignores the effect that it has on others in its private decision. This is known as an externality.

One prescription for solving the externality problem is to “internalize” it by placing a single decision maker in charge. In the oil industry, there are cases where the owners of land above an oil pool combine to place the oil drilling decision under a single management, known as unitization.\(^{20}\) The implication here is that the patent race should be placed under the command of a single decision maker.

It is likely that placing R&D decisions under a single decision maker, private or public, would lead to harmful effects that are not captured by this model. That is, the value of inde-

---

20. A more detailed discussion is given in chapter 18 in the discussion under Solutions to the Common Pool Problem.
pendent inventors trying new approaches that may not coincide with the majority view is often important. A single management committee coordinating multiple R&D projects could easily overlook promising approaches.

Another concern about the patent race described here is the assumption that there is one and only one product. In fact, it is often not the case that there is only one winner. In pharmaceuticals, for example, many firms might be seeking a cure for high blood pressure or cancer, and the outcome is a variety of different drugs with different properties—sometimes a drug for a disease not even being considered may be discovered.

Furthermore, patent monopolies are often temporary and can be displaced quite early in their product lives by newer, better products. Hence, it is probably reasonable to say that before R&D overinvestment becomes a widely accepted policy concern, patent race models will need to be refined to incorporate more realistic assumptions.21

A New Product Invention When Substitutes Exist

The optimal patent life model assumed a cost-reducing invention. Here we consider a new product invention that is a close substitute for an existing product. What is the social benefit of a new product \( N \) that is a close substitute for an existing product \( E \)? As we shall see, the introduction of substitute products leads to significant changes in the way social benefits are measured. It is also true that it becomes possible for private incentives to invent to become either too large or too small from a social welfare perspective.

In figure 24.7(b) the demand and cost of the new product \( N \) are shown. Demand is \( dd \) with a constant average cost of \( c' \). The demand \( dd \) takes as given the price \( P_0 \) of a substitute product \( E \) (in panel a of figure 24.7). The introduction of \( N \) at price \( p' \) leads to consumers purchasing quantity \( q' \).

After \( N \) is introduced, the demand for \( E \), which was originally \( DD \), shifts leftward to \( D'D' \). The constant average cost of \( N \) is assumed to be \( C_0 \). This is shown in figure 24.7(a). For simplicity, we shall assume that the supplier of \( E \) keeps the price of \( E \) fixed even after the introduction of \( N \).22 As shown, the quantity of \( E \) purchased decreases from \( Q_0 \) to \( Q_1 \), leading to profits that fall from \( P_0BFC_0 \) to \( P_0GHC_0 \), or by the amount \( GBFH \), or \( D'PSe \).

Now, consider the question of how to measure the social benefit of the introduction of product \( N \). Clearly, the total economic surplus of \( N \) (\( dbp' \) of consumer surplus, or \( CS_n \), and

---

21. In a study of pharmaceutical R&D that attempted to test statistically for the existence of patent races, it was concluded that there was no support for that hypothesis. R. Henderson and I. Cockburn, “Racing to Invest? The Dynamics of Competition in Ethical Drug Discovery,” *Journal of Economics and Management Strategy* Vol. 3(3) (Fall 1994): 481–519.

22. If the price of \( E \) is reduced after entry, a further demand shift for \( N \) occurs and the computation of social benefit becomes a bit more complex. An excellent exposition of this case can be found in a comment by R. W. Hansen in R. B. Helms, ed., *Drugs and Health* (Washington, D.C.: American Enterprise Institute, 1981), p. 295.
\( p'bc' \) of producer surplus, or \( PS_n \)) is a major component of the social benefit of \( N \). However, unlike the case we considered earlier in the optimal patent life section, we must make an adjustment for the effect of \( N \) on the substitute product \( E \).

The adjustment is that the loss in profit on \( E \), \( \Delta PS_e \) in figure 24.7, must be subtracted. This is the “business stealing” effect. The idea is that the firm introducing the new product does not internalize the loss of profit suffered by its rivals. By itself, the effect suggests a tendency for too much innovation.

The social benefit, \( SB \), of the new product \( N \) is

\[
SB = CS_n + PS_n - \Delta PS_e. 
\]  

An often troublesome theoretical point for some is why the shift leftward of demand for product \( E \) does not necessitate a subtraction of consumer surplus for product \( E \) of, say, area \( DBGD' \). The answer is that the demand for product \( N \) assumes the availability of product \( E \) at price \( P_0 \), and it therefore gives the increment to consumers’ willingness to pay due to the introduction of \( N \). This result is exactly what is desired.23

---

23. Another way to explain the point is that total consumer surplus for products \( N \) and \( E \) is \( DBP_0 \) plus \( dbp' \). It is not the sum of the two marginal consumer surplus areas \( D'GP_0 \) and \( dbp' \). Each marginal surplus assumes the availability of the other product at its equilibrium price; hence, the consumers’ value of \( E \) without \( N \) available is not captured by the latter sum.
The social benefit, given by equation 24.5, leads to another possibility for private investment in R&D being socially too large or too small. Assuming that the two products are supplied by different firms, then the private incentive to N’s inventor is simply $PS_n$. Comparing $PS_n$ with $SB$, it is clear that the private incentive can be either larger than the social benefit (if $\Delta PS_e > CS_n$) or smaller than the social benefit (if $\Delta PS_e < CS_n$). Interestingly, the portion of the social benefit not appropriated by the inventor ($CS_n$) acts to offset the business-stealing effect ($\Delta PS_e$) in bringing the social and private benefits closer in value.

It is obviously difficult to know how these magnitudes might compare in particular real-world cases. It would be useful to at least know whether too much or too little inventive activity might be stimulated by this factor. Based on admittedly rough estimates, Scherer and Ross argue that the ready-to-eat cereal industry of the 1960s and the soft drink industry of the 1980s probably had excessive new product innovation.24 The basic idea is that $CS_n$ was probably low because consumers did not perceive great differences among products, while $\Delta PS_e$ was probably large because of the high profit margins in those industries.

Pharmaceuticals and the Role of Patents

We turn now from the theoretical analysis of patents to consider the actual role of patents in the pharmaceutical industry. In particular, the 1984 Drug Price Competition and Patent Term Restoration Act will be examined. Although there was no pretense by policymakers that they were adjusting the patent life in pharmaceuticals to the socially optimal life, it nevertheless represents a rare case in which the legal patent life in an industry was actually changed. We will discuss the reasons for and economic effects of that change.

First, we provide a brief overview of the structure of the U.S. pharmaceutical industry as background.25 The following section will describe the 1984 act and its economic effects.

Industry Structure

The pharmaceutical industry is sometimes referred to as the ethical drug industry. It can be thought of as the industry that discovers, manufactures, and sells drugs that require a doctor’s prescription. It is a leading high-tech industry and is consistently at the top of American


industries in terms of R&D spending per dollar of sales. The industry has contributed greatly to improved health by virtually eliminating certain diseases (e.g., diphtheria, smallpox, and polio) and by reducing deaths from others (e.g., tuberculosis and heart disease). Expenditures on pharmaceuticals account for about 9 percent of total health care expenditures in the United States.

It is a relatively young industry—its beginning as a research-oriented industry dates back to the mid-1930s, when the first important group of anti-infective drugs were introduced. After World War II, pharmaceutical research broadened to cover many different therapeutic areas. Drugs were introduced to deal with cardiovascular, respiratory, neurological, and other disease categories.\(^{26}\)

**Government Regulation**

The development of the industry into a research-based industry competing in terms of new drug innovation was accompanied by the development of extensive government regulations of new drugs.\(^{27}\) Government regulation of this industry dates back to the Pure Food and Drug Act of 1906. This regulation was directed primarily at the adulteration and mislabeling of food and drugs sold in interstate commerce. In 1938, following a drug disaster that killed over a hundred children, the Food, Drug and Cosmetic Act of 1938 was passed by Congress. This law required new drugs to be approved as safe by the Food and Drug Administration before they could be introduced into interstate commerce. And, in 1962, Congress passed the important Kefauver-Harris Amendments to the Food, Drug and Cosmetic Act.

As described in chapter 22, the 1962 amendments required that a drug’s efficacy as well as safety be demonstrated on the basis of well-controlled scientific tests prior to marketing approval by the FDA. Chapter 22 also explains the market failure rationale for product quality regulation in pharmaceuticals as a case of imperfect information.

The 1962 amendments added considerably to the length and cost of the R&D process. Figure 24.8 is an illustration of the average length of the process. It begins with a highly uncertain period of discovery involving laboratory and animal studies. This is shown in the figure as lasting about six years. After a new chemical compound is discovered that has the potential for being an effective drug product, the first of three phases of clinical testing is initiated (FDA approval is required before testing in humans is permitted). On average, the clinical testing adds six more years to the process.\(^{28}\)

---


Following the clinical tests, and assuming the drug is still regarded as a promising new therapy, a New Drug Application (NDA) is submitted to the FDA. These applications cover, on average, clinical trials of over 3,000 patients and contain 90,000 pages. After another two years the FDA gives its decision, which, if positive, permits the firm to begin selling the drug. Of every four drugs that begin clinical trials, one will be eventually approved by the FDA.

The industry typically introduces between twenty and thirty new chemical entities (NCEs) each year. Figure 24.9 shows the trend in total industry R&D expenditures and the number of NCEs introduced between 1980 and 1992. It indicates that total R&D expenditures have been growing rapidly in real terms since the 1980s. The figure suggests the average R&D costs per NCE are also rising.

**Demand**

The concentration of buyers in the retail market is low if one thinks of patients taking their prescriptions to a local pharmacy. Of course, the key decision maker is the doctor who writes the prescription, and who has typically been the main target of pharmaceutical

---

29. An NCE is a new therapeutic molecular compound that has never before been used in humans.

30. Because of the lengthy gestation period for an NCE, and the fact that the period is increasing, the average cost per NCE is not simply the yearly expenditure divided by the number of NCEs in that year. Nevertheless, the cost per NCE has been found in the DiMasi et al. study to have increased dramatically. The reasons include higher clinical trial costs, the adoption of expensive new technologies, and the fact that “firms are focusing development more on treatments for chronic and degenerative diseases, which typically require longer and more expensive testing.” See DiMasi et al., “Cost of Innovation,” p. 133.
marketing. It is often noted that the doctor is more concerned with the quality of the drug than its price, and, in fact, may not even know the price. This characteristic of the demand side makes for relatively inelastic demand. This inelasticity is reinforced by the facts that often drugs are seen as vital for one’s health and many consumers’ drug purchases are covered by insurance.

31. It is possible for pharmacists to substitute a generic drug for a brand name prescribed by a doctor if the doctor does not specify that the brand name is required. Of the new 1989 prescriptions written for drugs that were multi-source, only 19 percent were written generically. Of the remaining brand-written multisource prescriptions, 29 percent were dispensed generically.

Because it is generally more profitable for pharmacists to dispense a generic than a brand-name drug, it is interesting to consider why generic substitution is so low. One reason is that doctors frequently prohibit substitution. For example, in states where doctors can prohibit substitution by simply signing their name on the appropriate line, substitution was prohibited 41 percent of the time. In states where doctors must write “Dispense as Written,” or something equivalent, substitution was prohibited only 11 percent of the time. J. K. Hellerstein, “The Demand for Post-Patent Prescription Pharmaceuticals” NBER working paper 4981 January 1994. Available online at www.nber.org/papers/w4981.pdf.

32. Because over a thousand new drugs have been introduced into therapeutic practice in the past forty years or so, it is very difficult for doctors to be fully informed on such large numbers of drugs.
It should be noted, however, that the preceding characterization of demand is rapidly changing as “managed care” organizations and other large group buying institutions are becoming increasingly important. For example, health maintenance organizations (HMOs) are “firms” of health care providers that are quite cost-conscious because they sell care to patients for a fixed price per year. They often bargain with drug manufacturers over prices and receive large quantity discounts. Their leverage is based on a “formulary,” or list of approved drugs for the HMO doctors. The HMO pharmacy committee’s decision to include or exclude a drug is therefore very important to the manufacturer. Government agencies such as Medicaid also employ various tactics to obtain lower prices.

In addition, firms known as prescription benefits managers (or PBMs) have become quite important. These firms are hired by large employers, insurance companies, HMOs, and other health care providers to lower drug costs. Because they represent many customers, they can negotiate big discounts with manufacturers. In 1993 and 1994 the three largest benefits management firms were bought by manufacturers. Eli Lilly bought the largest manager, PCS Health Systems, which covers prescriptions for 50 million people, in July 1994. In brief, the demand for drugs has become much more price sensitive in recent years.

**Seller Concentration**

Although the pharmaceutical industry is multinational in character, we focus here on the U.S. market. The first point is that there are a large number of research-intensive firms. Considering pharmaceutical sales by manufacturers to drug stores and hospitals in the United States as a single market (about $71 billion in aggregate), the top four sellers in 1994 with their respective market shares (percentages) were

<table>
<thead>
<tr>
<th>Company</th>
<th>Market Share</th>
</tr>
</thead>
<tbody>
<tr>
<td>Merck</td>
<td>7.3</td>
</tr>
<tr>
<td>American Home Products</td>
<td>7.3</td>
</tr>
<tr>
<td>Bristol Meyers–Squibb</td>
<td>6.4</td>
</tr>
<tr>
<td>Glaxo</td>
<td>6.4</td>
</tr>
</tbody>
</table>

Total: 27.4

Hence, the four-firm concentration ratio is relatively low at 27.4. However, if one considers the more meaningful markets to be “therapeutic categories” rather than pharmaceuticals in aggregate, then the concentration ratios are higher. One study found that the average therapeutic market concentration ratio was 70. The rationale for the more narrowly defined

---


markets is that on the demand side, categories like antibiotics and cardiovasculars, for example, are poor substitutes.

In addition to the research-intensive firms listed, there are many “generic” manufacturers. Geneva, Mylan, and Zenith are examples of generic firms that do little research; rather, they specialize in copying brand-name products after the brand product’s patent expires. For example, when the popular brand-name tranquilizer Valium went off patent in the mid-1980s, some fifteen to twenty generic suppliers began selling the generic version of Valium, known as diazepam, at large discounts. In 1993 the average generic price of diazepam was only 2 percent of Valium’s price. Even so, Valium with its much higher price still had a market share of 25 percent of total diazepam pills sold.

**Barriers to Entry**

Three sources of entry barriers can be identified: patents, brand loyalty, and scale advantages in R&D (which includes winning FDA approval).

**Patents**

As noted earlier, patents are generally thought to be more important to foster innovation in the pharmaceutical industry than in most other industries. The reason is that once a new chemical structure is marketed, the cost of imitation is usually quite low.

An important ruling by the Patent Office in 1948 concerning the antibiotic streptomycin opened the door to the patenting of new drugs. That is, new drugs would not be patentable if they were simply natural substances. In the case of streptomycin, the Patent Office ruled that the natural materials found by Waksman, streptomycin’s discoverer, were not in suitable form for medical use. Because chemical modifications to streptomycin had been made so that it could be purified, the Patent Office ruled that a “new composition of matter” had been created.

A patent barrier can be overcome by the development of chemically distinct substitutes for the original product. One strategy for inventing around an existing firm’s patent is termed “molecular modification.” This refers to the development of a similar compound so as to retain a rival product’s main therapeutic effects, but at the same time to have a chemically distinct structure so that it can be patented. This is a controversial practice, and some see it as wasteful; however, it has sometimes led to the creation of superior products that do not have the harmful side effects of the original product.

The antiulcer drug Zantac is an example of such molecular modification. In 1977, SmithKline introduced its newly discovered antiulcer drug, Tagamet, on the U.S. market. It was a drug that was “designed” to fit a receptor site in the body to “turn off” the release of

35. Many generic manufacturers are owned by research-intensive pharmaceutical firms.
histamine in the stomach. Histamine, in turn, could no longer stimulate the secretion of gastric acid in the stomach. This approach proved to be far superior to the use of antacids or surgery in the treatment of ulcers. By 1980, Tagamet became the largest-selling drug in the world, ranking, according to a *Fortune* article, “as one of the most stunningly successful products in the history of American business.”

In the mid-1970s, Glaxo began a search for a similar molecule that would also fit the histamine receptor site. It was granted a patent for its molecule, Zantac, in 1978, and by 1987, Zantac had overtaken Tagamet and had become itself the best-selling drug in the world. Its advantage over Tagamet was claimed to be fewer side effects. Zantac’s daily dosage also was only twice daily, compared to Tagamet’s four times daily (fewer dosages per day are generally preferred by doctors because it increases patient compliance).

By 1992, Zantac’s sales amounted to $3.2 billion worldwide—over $1 billion more than the next-best-selling drug. As might be expected, other companies have also come up with similar antiulcer drugs, but they have not been as commercially successful as Zantac. In 1994, however, the patent on Tagamet expired, and it can be expected that vigorous price cutting by generic suppliers will significantly impact the sales of Zantac.

Glaxo owned two patents on Zantac, one expiring in 1995 and the second in 2002. Glaxo has brought a number of patent infringement suits to defend its belief that the 2002 patent is the valid one. The 1995 patent covered the basic molecular structure of Zantac and one crystalline form of the drug. The 2002 patent covered a second crystalline form, the form actually in use. In the litigation over the patents the courts held that the 1995 patent was legitimate and that the generic companies were not infringing on the 2002 patent for a polymorph. However, due to the Uruguay Trade Agreement that harmonized patent laws throughout the world, Glaxo got an extra 19 months on its 1995 patent, pushing its expiration date to 1997.

**Brand Loyalty**

Brand loyalty has already been illustrated by the case of Valium and the fact that it maintained 25 percent of the total sales of diazepam despite generic prices being only 2 percent of the price of Valium. The point, of course, is that generic versions of the tranquilizer are necessarily bioequivalent to Valium. This equivalence is required by the FDA for generics to be marketed. Hence, quality differences are quite small and must be due largely to consumers’ perceptions that the brand name is superior. Of course, consumers often have poor information about the availability, price, and quality of generic substitutes.

37. To be bioequivalent, the rates at which the active ingredient of two drugs are absorbed by the body must not differ significantly.
In a 1996 study, major drug products whose patents expired in the 1984–1991 period were analyzed with regard to the pattern of generic penetration. The results, updated by the authors to 1993, are reported in table 24.1. As can be seen, the average generic price was introduced at about two-thirds of the brand price, falling to about one-third after two years. Even so, the total market share in units that generics won after two years was only about half for the two earlier cohorts. (We will consider the two more recent cohorts later.) Note also the fact that the brand-name prices rose on average in the face of entry! The table indicates an increase of just over 10 percent.\(^{39}\)

One interpretation of the results in table 24.1 is that the market is segmented into two groups, a price-sensitive group and a group that has strong loyalties to the branded products. If a brand-name product can keep half the market at the unchanged price as opposed to

---

**Table 24.1**


<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average brand-name price index</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At date of entry</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>One year after entry</td>
<td>1.06</td>
<td>1.08</td>
<td>1.06</td>
<td>1.05</td>
</tr>
<tr>
<td>Two years after entry</td>
<td>1.11</td>
<td>1.12</td>
<td>1.10</td>
<td>1.09</td>
</tr>
<tr>
<td><strong>Average generic price index</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At date of entry</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>One year after entry</td>
<td>0.77</td>
<td>0.79</td>
<td>0.86</td>
<td>0.67</td>
</tr>
<tr>
<td>Two years after entry</td>
<td>0.65</td>
<td>0.67</td>
<td>0.63</td>
<td>0.54</td>
</tr>
<tr>
<td><strong>Average ratio of generic price to brand-name price</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At date of entry</td>
<td>0.63</td>
<td>0.59</td>
<td>0.61</td>
<td>0.68</td>
</tr>
<tr>
<td>One year after entry</td>
<td>0.47</td>
<td>0.44</td>
<td>0.49</td>
<td>0.43</td>
</tr>
<tr>
<td>Two years after entry</td>
<td>0.38</td>
<td>0.36</td>
<td>0.35</td>
<td>0.33</td>
</tr>
<tr>
<td><strong>Average generic market share in physical units (proportion of total market)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At date of entry</td>
<td>0.07</td>
<td>0.11</td>
<td>0.13</td>
<td>0.20</td>
</tr>
<tr>
<td>One year after entry</td>
<td>0.32</td>
<td>0.38</td>
<td>0.41</td>
<td>0.64</td>
</tr>
<tr>
<td>Two years after entry</td>
<td>0.45</td>
<td>0.54</td>
<td>0.59</td>
<td>0.73</td>
</tr>
</tbody>
</table>

**Note:** The drugs include ten major 1984–1985 drugs, eight major 1986–1987 drugs, seven major 1989–1991 drugs, and ten major 1992–1993 drugs. Each value is an unweighted average of the values for all drugs in each category.


39. It should be noted that an 11 to 12 percent increase over two years during the period in question is probably less than the average pharmaceutical price increase for all products. The main point is that the brand-name products did not cut their prices. A similar study, but one using a different sample of drugs, found broadly equivalent results. R. Caves, M. Whinston, and M. Hurwitz, “Patent Expiration, Entry, and Competition in the U.S. Pharmaceutical Industry,” *Brookings Papers on Economic Activity: Microeconomics* (1991) 1–48.
keeping the whole market by lowering price to one-third the original level, simple calculations indicate that the unchanged-price strategy is more profitable.\(^{40}\)

It should be noted that managed care organizations and other cost-containment market forces are becoming increasingly important. Hence, brand loyalty should become less and less important over time. In terms of the segmented market story that we have presented, the price-sensitive group is likely to become larger and larger.\(^{41}\) This trend can be seen by examining the two more recent cohorts in table 24.1. The generic market share after two years has increased substantially from about half for the 1984–1985 and 1986–1987 cohorts to 73 percent for the 1992–1993 cohort.

**R&D Scale Economies**

The final type of entry barrier is that of economies of scale in R&D and the need to obtain FDA approval for newly discovered drugs. Several factors can be mentioned here. First, a firm must maintain a portfolio of R&D projects because only one in four that enter clinical testing are ever marketed. Furthermore, it turns out that only three out of ten drugs that are marketed cover their total costs, including their share of failures.\(^{42}\) Therefore, a firm must maintain a large enough R&D budget to ensure that it will have at least the minimal number of successes necessary to maintain financial viability.\(^{43}\)

The R&D process presents several opportunities for traditional economies of scale. For example, the discovery process is characterized by significant fixed costs. Multidisciplinary teams of biologists, chemists, and other scientists are engaged in research to develop concepts and hypotheses concerning new compounds. The clinical development process also requires significant regulatory and legal expertise, which is also characterized by fixed costs and specialization. In principle, larger firms can spread these fixed costs over more R&D projects.

The R&D cost per NCE marketed is also quite high in absolute dollars, and has been increasing in real terms quite rapidly. This point is clear from figure 24.9. A 1993 study by

---

40. Using an estimate of marginal cost of 25 percent of price and assuming that the market is perfectly inelastic with respect to price, profits will fall by 89 percent with price cutting and by only 50 percent with the price unchanged.


41. In 1984, generic drugs accounted for 19 percent (in units) of all pharmaceuticals sold in the United States. By 1995, the share had grown to about 43 percent.


the U.S. Office of Technology Assessment concluded that the average after-tax R&D cash outlay for each new drug that reached the market in the 1980s was about $65 million in 1990 dollars. It went on to say, “The R&D process took 12 years on average. The full after-tax cost of these outlays, compounded to their value on the day of market approval, was roughly $194 million.”

In an 1991 article in the *Wall Street Journal* about a number of recent mergers in the industry, several executives observed that the mergers were occurring for R&D needs. According to the head of SmithKline Beecham—which resulted from the 1989 merger of the U.S. firm SmithKline and the British firm Beecham—“few drug companies on their own can afford truly innovative research these days. More consolidation is inevitable.” The chief executive of Glaxo noted that “to be a big player a company must spend somewhere north of $500 million a year [on R&D] and grow it by more than 10% or 15% a year. Those who can’t spend that will be left behind; they’ll be good candidates for a merger.”

In fact, in the mid-1990s, a relatively large number of mergers took place. Glaxo bought Burroughs Wellcome, Ciba-Geigy and Sandoz merged to form Novartis, American Home bought Lederle, and Roche bought Syntex. In late 1999, American Home and Pfizer made offers to acquire Warner-Lambert. That company’s acceptance of the Pfizer offer made Pfizer the largest pharmaceutical firm in the world, surpassing Glaxo Wellcome. Subsequently, Glaxo Wellcome merged with Smith Klein Beecham, but Pfizer also expanded by merging with Pharmacia, retaining its ranking as the largest pharmaceutical firm.

There have also been a number of statistical studies of R&D productivity in pharmaceuticals. The general findings seem to be that there are economies of scale or advantages to larger firms that at least partly result from the increased regulatory stringency of the FDA after 1962.

This completes our brief overview of the structure of the pharmaceutical industry. It is an industry in which patents are essential to provide incentives for R&D investment and a major form of competition is through the introduction of new drug products. However, recent developments are leading to an increasing role for price competition. We turn now to the analysis of an important 1984 law that simultaneously increased patent lives and lowered barriers to generic competition.

The 1984 Drug Price Competition and Patent Restoration Act

In 1984, President Reagan signed into law the Drug Price Competition and Patent Restoration Act. The law facilitated the entry of generic competitors after patent expiration, thereby leading to price competition. It also restored part of the patent life lost during the premarket regulatory process for new pharmaceuticals. 47

Provisions of the Act

Figure 24.10 can be used to make these two main provisions clear. The figure is a representation of the cash flows of the average new drug innovation. First, there is a twelve-year period of R&D costs and FDA review time from the origin to point a. Point a represents the beginning of marketing, with the pre-1984 profile of net revenues being shown by the path abcd. The post-1984 profile is abefd.

Considering the pre-1984 period, the legal patent life begins during the R&D period and extends, on average, for eight years into the marketing period. Hence, as shown in figure 24.10, there was an “effective patent life” of eight years. The point, of course, is that part of the legal patent life is consumed in premarket R&D and regulatory activities, leaving only eight years for exclusive marketing.

Chapter 24

The act provides for an extension in effective patent life equal to the sum of the FDA review time plus one-half the clinical testing time, subject to certain constraints. For example, there is a maximum extension of five years and no extension beyond fourteen years of effective patent life. In figure 24.10, the amount of patent restoration is represented by the distance $be$.

Figure 24.10 also indicates the second provision of the act as the relatively large drop in net revenues upon patent expiration in the post-1984 case, the distance $ef$, as compared with the smaller drop in the pre-1984 case, or $bc$. The loss is relatively small in pre-1984 because of the significant barriers to entry that confronted generic competitors. (The size of the losses is the product of two factors: the probability of generic entry and the losses of revenues to generics given that entry occurs. It should be recalled that the cash profile is for the average drug—and in the pre-1984 case, generic entry simply did not occur in many cases.)

The entry barriers in the pre-1984 period were due to the fact that generic firms frequently could not rely on the safety and efficacy evidence submitted by the brand-name firms (the innovators). Unless the relevant data were publically available in the scientific literature, an imitator had to duplicate many of the innovator’s tests to gain FDA approval. Under the 1984 law, a generic firm need only submit an Abbreviated New Drug Application (ANDA). This requires only that bioequivalency be demonstrated, a relatively low-cost test. Also, clinical testing by generic firms in the pre-1984 period was subject to litigation if done before patent expiration. Hence, contrary to figure 24.10, entry usually occurred some time after patent expiration.

Reasons for and Welfare Effects of the Act

With regard to the brand-name drug’s revenues, one effect—patent restoration—leads to a gain, whereas the second effect—easier generic entry—leads to a loss. The gain and loss areas are shown in figure 24.10. Of course, the losses to the brand-name firm are gains to generic firms. Hence, as might be expected, the act was passed after a long period of compromises among various interest groups. (In addition to the pharmaceutical firms and their associations, groups representing consumers favored the lower prices and sided with the generic firms.) Obviously, the optimal patent-life model discussed earlier in which the policymaker was assumed to choose the life that maximizes total surplus was not applied!

Senator Hatch, one of the two sponsors of the act, described the bill to the Senate as “carefully balanced . . . in ways that only lawyers could have devised.” More generally, he said:

This is a groundbreaking compromise in the public interest. It reconciles the opposing, competitive interests of two segments of the pharmaceutical industry which have often stymied each other’s attempts to improve the law. The research-based drug industry obtains an extension of patents for new drug
discoveries to compensate them for the time spent off-market in FDA review. The generic drug industry gets to bring generic copies of off-patent drugs to market as soon as the patent expires, without the needless reduplication of studies and tests already in FDA’s files.

The public receives the best of both worlds—cheaper drugs today and better drugs tomorrow.\(^{48}\)

From the perspective of economic welfare, the act is the source of large potential positive gains of two types. First, it eliminated costly scientific testing that served no valid purpose. Second, the act lowered prices to consumers with some elimination of deadweight losses and large transfers from producers to consumers.

At the same time, the act had the potential of lowering the expected returns from drug innovation—it depends on the relative magnitudes of the gains and losses in figure 24.10. One might interpret Senator Hatch’s reference to “better drugs tomorrow” as his belief that the act would increase incentives to invest in R&D. Of course, as we discussed in the earlier part of this chapter, it is possible from society’s viewpoint to have too much appropriability as well as too little. However, if it was the intention of Congress to at least maintain the incentive to introduce new drugs, one can attempt to examine this question empirically.

**The Net Effect on R&D Incentives**

In this section we give the results of a 1998 Congressional Budget Office (CBO) study that estimated the effect of the 1984 act on R&D incentives.\(^{49}\) The study basically followed the strategy suggested by figure 24.10 of comparing the present value of net revenues with and without the act in effect.

The primary finding was that the net revenue stream for a typical new drug in the pre-1984 period had a present value of $27 million (in 1990 dollars) more than in the post-1984 period. That is, the act was estimated to have lowered incentives to invest in R&D. The CBO then compared this figure to other studies’ estimates of the total discounted returns from selling a brand-name drug introduced in the early 1980s. The estimates ranged from $210 million to $230 million. (Those figures represent the present value of the stream of net revenues discounted to the date of marketing, deducting all costs other than R&D.) Hence, as a percentage, the effect of the 1984 act was estimated to have reduced returns by roughly 12 percent.

The study concluded with an interesting argument about proposals to modify the 1984 act:

Some representatives of the pharmaceutical industry would like to modify the Hatch-Waxman Act in various ways to increase the average effective patent term for pharmaceutical products. Although

---


lengthening patents would increase profits today for drugs whose patents are expiring, it would not have a large impact on the incentive to invest in R&D. . . . Extending the average effective patent term by one year would increase the present discounted value of those returns by about $12 million.

In contrast, accelerating the FDA review period by one year would have a much greater effect on the present discounted value of the returns from marketing a new drug—a net benefit of about $22 million, on average.  

Other Policies That Affect R&D Incentives

Of course, patent policy is only one of various ways that the government affects the incentives to invest in R&D. Here we shall discuss several briefly. These include favorable tax treatment, market exclusivity for so-called “orphan” drugs, and various types of price controls.

Tax Subsidies and Orphan Drugs

The tax code favors R&D relative to investment in plants and equipment because R&D can be “expensed” rather than depreciated and written off over time. That is, a drug company can deduct its R&D expenses from income in the year incurred, thereby recovering the expenses much quicker than it can recover its investment in plant and equipment. Also, in 1981, Congress passed a law that permitted a tax credit for increases in R&D. The tax credit was reduced to 20 percent in 1986.

The 1983 Orphan Drug Act, which was passed to encourage firms to develop new treatments for diseases that affect small numbers of people, provides for a tax credit equal to 50 percent of R&D expenses for clinical trials. The trials must be for drugs that have been given orphan drug status—primarily drugs that treat diseases or conditions affecting less than 200,000 people in the United States. The idea, of course, is that drugs for such small markets would not be profitable for firms in the absence of subsidies.

The first firm to receive FDA approval for an orphan drug also may market it exclusively for a seven-year period beginning on the date of approval. Any patent protection covering the drug runs contemporaneously with the market exclusivity. Through September 1992, the FDA had granted orphan status to 494 drugs. There has been much controversy about certain orphan drugs that are perceived by critics as being too profitable, sometimes because of very high prices. Legislation to limit excessive profits has been proposed but has not yet been successful.

50. Ibid., p. 49.
51. The increases equal the difference between R&D expenses in the current year and the average amount spent during the previous three years.
52. OTA, Pharmaceutical R&D, p. 226.
Price Controls and Profits

Price controls for pharmaceuticals is a recurring issue. Although most foreign countries have some type of price control system for drugs, the United States generally does not. However, the possibility of including expanded coverage for prescription drugs under health care reform proposals, along with the perceived need for accompanying cost controls, often makes price controls an issue. In addition, critics of the industry have long been convinced that the industry is too profitable and charges prices that are too high.

To illustrate the two points of view, we quote some of the arguments from a 1992 debate on the Senate floor. The legislation being discussed was a proposal by Senator Pryor to penalize drug companies that increase their prices faster than the rate of inflation.

Senator Pryor:

From 1982 . . . to 1992, 10 years, while the general inflation rate was just 46 percent in that decade, prescription drug prices increased 142 percent . . . Fortune magazine, July 29, 1991, said the manufacture of pharmaceuticals is America’s most profitable business . . . .

I would only say that today those profits are being made at the expense of the most vulnerable members of our society . . . . In the United States, we spend $270 for every man, woman, and child a year for prescription drugs and most of this is not covered by insurance, it is not covered by Medicare, it is coming out of the pockets of our citizens who are least able to pay . . . .

In 1990 . . . the average rate of profit for the Fortune 500 companies was 4.6 percent . . . . What about the pharmaceutical companies? . . . . Let us see how they are getting along—15.5 percent, that was their average profit in the year 1990 . . . . Now how do they make these enormous profits? . . . . By outright price gouging of our American citizens who can least afford the medications—the elderly, the poor, and other vulnerable parts of the American population.

Of course, Senator Pryor’s arguments are more detailed than indicated by his statements. However, his major point is that he thinks that the industry has too much monopoly power. It is interesting that his solution is to use price controls to directly attack what he perceives to be the problem. Given that much of the market power presumably stems from patents, one might think that shortening the patent life would be considered.

Senator Bradley:

So, Mr. President, what I believe is a major concern about Senator Pryor’s amendment is its effect on investment, research, and innovation in this country. Senator Pryor has singled out one sector in the health care economy that is the most heavily research oriented and funds a significant amount of all research on health care . . . .

And although it is not easy to predict the reactions in the marketplace to Government intervention, this one is simple: Price controls, as envisioned in this amendment, will significantly reduce incentives

54. An exception is the Omnibus Budget Reconciliation Act of 1990, which requires drug manufacturers to give Medicaid programs rebates for drugs based on the lowest prices available to any purchaser.


for investment. . . Reduction in research will lead to fewer innovations, fewer cures, and fewer hopes for many Americans who are counting on medical breakthroughs to lengthen their lives.

Certainly, lower prices will help consumers to be able to afford prescription drugs. But the question is, what are they going to be able to buy? 56

As is perhaps natural in a debate, the senators tend to focus on the extremes—one on the monopoly pricing of existing drugs and the other on the benefits of invention, or future drugs. However, there is a background issue that deserves further analysis, and that is the question of the level of profits in this industry. 57 It is a complex issue, and we cannot devote the necessary space to it here. Hence, we will simply make a few key points here, and refer the reader to several relevant articles. 58

Perhaps the most important point is that profit rates taken from the annual financial reports of firms can be very misleading. The profit rates referred to by Senator Pryor from Fortune are of this type. We briefly commented on some of these problems in chapter 9. With regard to pharmaceutical accounting profit rates, the most serious problem is that accountants expense R&D rather than capitalize and depreciate it as they do other plant and equipment. R&D is the major form of investment for pharmaceutical firms, and it clearly has economic effects that last for years—just as plant and equipment do. However, accountants do not attempt to depreciate it over time—probably because of the difficulty of determining its “useful life.”

A simple example may be helpful:

\[
\text{Accounting profit rate} = \frac{(R - VC - r & d - d_k K)}{K}
\]

\[
\text{Economic profit rate} = \frac{(R - VC - d_{rd} RD - d_k K)}{(RD + K)}
\]

where \( R \) is revenues, \( VC \) is variable costs, \( r & d \) is current expenditures on R&D, \( d_k \) and \( d_{rd} \) are the depreciation rates of \( K \) and \( RD \) capital, \( K \) is plant and equipment capital stock, and \( RD \) is R&D capital stock.

As stated earlier, the accounting profit rate expenses R&D. The current expenditures on R&D appear in the numerator. Economic profit rates depreciate the R&D capital stock just as physical capital is depreciated. Assume for simplicity that the firm is in a “steady state”


57. In addition to the level of profits, there is a different point about the rate of increase in pharmaceutical prices being excessive. We have reason to believe that the government pharmaceutical price index is flawed and overstates the true price increase. For an analysis of this issue, see E. R. Berndt et al., “Auditing the Producer Price Index: Micro Evidence from Prescription Pharmaceutical Preparations,” Journal of Business and Economic Statistics Vol. 11(3) (1993): 251–264.

in which its R&D expenditures each year exactly equal the amount needed to offset the deprecation of the R&D capital stock. For this particular case, notice that the two profit rates have equal numerators, but the economic profit rate has a larger denominator. That is, the economic profit rate has both $RD$ and $K$ in its denominator, not just $K$. Clearly, for this particular case, the accounting profit rate overstates the economic profit rate.

Although this particular case—a steady state—is not necessarily applicable to pharmaceuticals, it does indicate the possibility that accounting profit rates may be too high. In fact, based on a review of six sophisticated studies of pharmaceutical industry profits, the Office of Technology Assessment concluded that “correcting pharmaceutical industry profit rates for investment in intangible capital reduces rates of return by roughly 20 to 25 percent.”

The reference to “intangible” capital stock indicates that the argument about R&D also applies to advertising—which is also expensed rather than capitalized. In 1993, the largest drug firm, Merck, invested about $1.2 billion in R&D, $1 billion in physical capital, and about one quarter of a billion dollars in advertising. Hence, the problem of intangible capital is particularly important to pharmaceuticals because of unusually high levels of both R&D and advertising.

An alternative to firm profitability is the profitability of new product introductions from an industry-wide perspective. In a study of the sixty-seven new drug introductions in the United States in the 1980–1984 period, the average rate of return on pharmaceutical R&D was estimated to be 11.1 percent. This return was compared to the industry’s cost of capital of about 10.5 percent (based on a study by S. C. Myers and L. Shyam-Sunder). This finding of the average return being above the cost of capital, but by a relatively small amount, is broadly consistent with a similar study by the Office of Technology Assessment.

It is instructive to note that underlying the average return is a highly skewed distribution of returns. Figure 24.11 illustrates this point. Using a 10.5 percent cost of capital, and grouping the sixty-seven drugs into deciles by sales revenues, the figure compares the average net present value of net revenues to the average capitalized value of R&D cost—all on an

---


60. Most of the marketing expenses of pharmaceuticals have traditionally been promotional visits to doctors, known as detailing. Advertising expenses constitute only about one quarter of total marketing expenses, with detailing accounting for most of the rest.


62. The study was prepared for the OTA report and is partially described in that report.

63. OTA, *Pharmaceutical R&D*. 
after-tax basis. The top decile has an estimated present value of cash flows after launch that is more than five times average R&D costs. In addition, only the top three deciles have present values that exceed average R&D costs.

This extreme skewness of returns to R&D has an important implication for the type of price controls described in President Clinton’s proposed Health Security Act of 1993. The proposal was to create an Advisory Council on Breakthrough Drugs that would focus on the “reasonableness” of the prices of such drugs. Because breakthrough drugs are likely to correspond to the top decile, or “blockbusters,” restricting their prices to a break-even level would significantly reduce the attractiveness of investing in R&D. If one regards R&D investment as somewhat like a lottery—with low probabilities of achieving huge returns—top-decile regulation completely changes the nature of the game. Winning the lottery would now provide only a reasonable or break-even return, with other outcomes worse!

Opinions certainly differ, but according to one expert,

If profits were held to “reasonable” levels on blockbuster drugs, aggregate profits would almost surely be insufficient to sustain a high rate of technological progress. Should a trade-off be required

---

64. The Congressional Budget Office notes, however, that if breakthrough drugs are to correspond to the FDA’s past rating of “significant therapeutic advances,” there may not be a close correspondence to commercial importance. U.S. Congressional Budget Office, How Health Care Reform Affects Pharmaceutical Research and Development, June 1994, p. 35.
between modestly excessive prices and profits versus retarded technical progress, it would be better to err on the side of excessive profits.65

Summary

The chapter has two main parts. The first part dealt with the economics of inventions and patents in general and from a theoretical perspective. We compared the incentives to invest in inventive activity for an inventor in a competitive industry with the incentive that a monopolist would have. Under the relatively special assumptions made, it was found that the incentive was greater for the inventor in a competitive industry.

The welfare economics of patents was next considered, and a simple model of the optimal patent life was presented. The model highlighted the trade-off between greater incentives to invest by giving a patent monopoly, and the cost of that monopoly in terms of inefficient pricing. Two complications were then discussed: so-called patent races and the special problem of measuring social benefits when the invention takes the form of a new product that is a close substitute for an existing product.

The second part of the chapter turned to the role of patents in the pharmaceutical industry. First, some background material on the structure of the industry was presented: the nature of government regulation, as well as the traditional discussion of demand, concentration, and barriers to entry. The 1984 Drug Price Competition and Patent Restoration Act was described and analyzed as to the net effect that it has had on the incentives to invest in R&D. Finally, other policies that affect R&D incentives were discussed, with particular attention given to the possibility of the imposition of price controls because of perceived excess profitability of the pharmaceutical industry.

Questions and Problems

1. If patents were no longer available under the law, would technical progress cease? Explain.

2. Assume that the market demand for shoes is \( Q = 100 - P \) and that the constant average cost of production is $60. Consider two alternatives. In case C the industry is initially organized competitively, and in case M the industry is organized as a monopoly. In each case an invention leads to a lower, constant average cost of production of $50. The R&D cost of the invention is not relevant for this problem, inasmuch as the issue is the magnitude of the incentive to invent in the two cases. Finally, there is no rivalry to make the invention: in case C the inventor may be assumed to have a patent of infinite life, and in case M the inventor is the existing monopolist and entry is barred.

a. Find the initial price and quantity equilibria in the two cases.
b. What is the return that the inventor in case C could expect from its lower cost production process? Assume that the inventor monopolizes the shoe industry.
c. What is the return to the monopolist in case M that results from inventing the lower cost process?
d. Interpret your results regarding the incentive to invent under monopoly and competition.
e. In case C, if society could require that the new process be used efficiently—by making it freely available to all—what would be the increase in total economic surplus? How does this magnitude compare to the returns found in parts b and c?

3. Refer to problem 2 and answer the same questions where the only change is that the invention lowers average cost to $10 (rather than to $50). This is a so-called major invention as compared to the minor invention of problem 2. Does this change affect the comparative incentives under monopoly and competition?

4. In this problem rivalry to make the invention is introduced. Assume that the demand for shoes is $Q = 100 - P$ and the constant average cost of production is $60$. The incumbent monopolist, M, faces competition from a single potential entrant, E, in being first to invent and patent a new, lower cost process for producing shoes. Let the new, lower cost process be one with a constant average cost of $50$. If E wins the race for the patent, a Cournot equilibrium will result with M having an average cost of $60$ and E having an average cost of $50$. If M wins the race, it will remain a monopolist, but with an average cost of $50$.

a. What is M’s incentive to win the race?
b. What is E’s incentive to win the race?
c. Explain the intuition underlying this result. It can be argued that this so-called efficiency effect leads to the “persistence of monopoly.” Why?

5. A breakfast cereal product, Cheers, has demand $Q = 5 - P/2$ and constant average cost of $1$. Under existing conditions, the supplier of Cheers charges $5.50$. Now, a new product, Kips, is introduced by a rival supplier. After equilibrium is reached, Cheers’ price is unchanged at $5.50$ and Kips’ price is also $5.50$. Cheers’ demand has shifted leftward and can be described as $Q = 4.25 - P/2$. Kips’ demand can be described as $q = 4.25 - p/2$, and its constant average cost is $1$.

a. What is the social benefit resulting from the introduction of Kips? (Ignore R&D cost for now.)
b. What is the private benefit of Kips’ introduction?
c. If the R&D cost of introducing Kips is $6$, what is the net social benefit? That is, subtract $6$ from your answer to part a. Compare this result with the net private benefit.
d. Discuss these findings from the perspective of welfare economics. In this instance, Kips would be introduced even though it is not socially beneficial. Is this finding true in general? Explain.

6. This problem illustrates the possibility that excessive resources can be expended in search of a new product, say, Panacea. For an excellent treatment—on which this problem is based—see chapter 17 of D. W. Carlton and J. M. Perloff, *Modern Industrial Organization*, 2nd ed. (New York: HarperCollins, 1994).
Assume that there are an unlimited number of firms that can each undertake one research project at a constant marginal cost of $1. The probability $P$ of discovery of Panacea by one of the $n$ firms searching for it is an increasing function of the number of research projects (firms). In particular,

$$P = 1 - e^{-0.5085n}$$

If Panacea is discovered, it will be priced competitively and the present value of total surplus will be $25. Research all takes place this year; if Panacea is not discovered this year, no research can take place in the future.

a. From society’s viewpoint, find the optimal number of research projects (firms). Hint: Define the expected social benefit as the probability of discovery multiplied by $25. Using calculus, find the marginal social benefit as the derivative of expected social benefit with respect to $n$ and equate the result to the marginal cost of $1$.

b. Assume now that a competitive R&D industry exists and that $n$ is determined by the zero profit condition. That is, firms will join the search for Panacea as long as the expected payoff to a firm is greater than its $1$ cost. If the government promises the firm that discovers Panacea the entire social benefit of $25$, how many firms will enter? Hint: The solution to the equation in which expected social benefit divided by $n$ equals $1$ is $n = 25$.

c. Explain the intuition underlying this so-called common pool problem.