pour mon épouse Ewa et mes enfants (JH)
to the memory of my parents (GDM)
I PUBLIC ECONOMICS AND ECONOMIC EFFICIENCY

1 An Introduction to Public Economics 3
  1.1 Public Economics 3
  1.2 Methods 3
  1.3 Analyzing Policy 5
  1.4 Preview 6
  1.5 Scope 9

2 Equilibrium and Efficiency 13
  2.1 Introduction 13
  2.2 Economic Models 13
  2.3 Competitive Economies 14
  2.4 The Exchange Economy 15
  2.5 Production and Exchange 22
  2.6 Efficiency of Competition 26
    2.6.1 Single Consumer 27
    2.6.2 Pareto-Efficiency 31
    2.6.3 Efficiency in an Exchange Economy 33
    2.6.4 Extension to Production 37
  2.7 Lump-Sum Taxation 39
  2.8 Discussion of Assumptions 41
  2.9 Summary 43

3 Behavioral Economics 51
  3.1 Introduction 51
  3.2 Behavioral Individuals 53
    3.2.1 Simple Example: How Much to Save? 53
    3.2.2 Present-Bias 54
    3.2.3 The (β, δ) Model of Self-Control 55
    3.2.4 Reference-Dependence Bias 57
    3.2.5 The Gambler’s Fallacy 58
## Contents

3.2.6 Confirmation Bias 59  
3.2.7 Confidence Bias 60  
3.2.8 Framing Bias 60  
3.2.9 Conformism Bias 62  
3.2.10 Identity and Social Norms 63  
3.3 Behavioral Markets 64  
3.3.1 Money Pump 65  
3.3.2 Complementary Mistakes 65  
3.3.3 Rationality Tug-of-War 67  
3.4 Behavioral Policy 67  
3.4.1 Internalities versus Externalities 68  
3.4.2 Automatic Enrollment 69  
3.4.3 The SMarT Plan 69  
3.4.4 Complementarity 70  
3.5 Behavioral Welfare 71  
3.5.1 New Welfare Criterion 71  
3.5.2 Choice-Based Welfare Analysis 72  
3.5.3 Refinement and Structural Modeling 73  
3.5.4 Application: Global Warming 74  
3.6 Other-Regarding Preferences 76  
3.6.1 Ultimatum Game 76  
3.6.2 Social Preferences 77  
3.6.3 Market Impact 78  
3.7 Conclusions 80  

## II GOVERNMENT

4 Public Sector Statistics 89  
4.1 Introduction 89  
4.2 Historical Development 89  
4.3 Composition of Expenditure 94  
4.4 Revenue 97  
4.5 Government Debt 107  
4.6 Measuring the Government 112  
4.7 Conclusions 114  

---

viii Contents

3.2.6 Confirmation Bias 59  
3.2.7 Confidence Bias 60  
3.2.8 Framing Bias 60  
3.2.9 Conformism Bias 62  
3.2.10 Identity and Social Norms 63  
3.3 Behavioral Markets 64  
3.3.1 Money Pump 65  
3.3.2 Complementary Mistakes 65  
3.3.3 Rationality Tug-of-War 67  
3.4 Behavioral Policy 67  
3.4.1 Internalities versus Externalities 68  
3.4.2 Automatic Enrollment 69  
3.4.3 The SMarT Plan 69  
3.4.4 Complementarity 70  
3.5 Behavioral Welfare 71  
3.5.1 New Welfare Criterion 71  
3.5.2 Choice-Based Welfare Analysis 72  
3.5.3 Refinement and Structural Modeling 73  
3.5.4 Application: Global Warming 74  
3.6 Other-Regarding Preferences 76  
3.6.1 Ultimatum Game 76  
3.6.2 Social Preferences 77  
3.6.3 Market Impact 78  
3.7 Conclusions 80  

II GOVERNMENT

4 Public Sector Statistics 89  
4.1 Introduction 89  
4.2 Historical Development 89  
4.3 Composition of Expenditure 94  
4.4 Revenue 97  
4.5 Government Debt 107  
4.6 Measuring the Government 112  
4.7 Conclusions 114
## Contents

### 5 Theories of the Public Sector

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1 Introduction</td>
<td>119</td>
</tr>
<tr>
<td>5.2 Justification for the Public Sector</td>
<td>119</td>
</tr>
<tr>
<td>5.2.1 The Minimal State</td>
<td>119</td>
</tr>
<tr>
<td>5.2.2 Market versus Government</td>
<td>121</td>
</tr>
<tr>
<td>5.2.3 Equity</td>
<td>122</td>
</tr>
<tr>
<td>5.2.4 Efficiency and Equity</td>
<td>123</td>
</tr>
<tr>
<td>5.3 Public Sector Growth</td>
<td>123</td>
</tr>
<tr>
<td>5.3.1 Development Models</td>
<td>123</td>
</tr>
<tr>
<td>5.3.2 Wagner’s Law</td>
<td>124</td>
</tr>
<tr>
<td>5.3.3 Baumol’s Law</td>
<td>125</td>
</tr>
<tr>
<td>5.3.4 A Political Model</td>
<td>126</td>
</tr>
<tr>
<td>5.3.5 Ratchet Effect</td>
<td>128</td>
</tr>
<tr>
<td>5.4 Excessive Government</td>
<td>129</td>
</tr>
<tr>
<td>5.4.1 Bureaucracy</td>
<td>130</td>
</tr>
<tr>
<td>5.4.2 Budget-Setting</td>
<td>132</td>
</tr>
<tr>
<td>5.4.3 Monopoly Power</td>
<td>133</td>
</tr>
<tr>
<td>5.4.4 Corruption</td>
<td>134</td>
</tr>
<tr>
<td>5.4.5 Government Agency</td>
<td>135</td>
</tr>
<tr>
<td>5.4.6 Cost Diffusion</td>
<td>137</td>
</tr>
<tr>
<td>5.5 Conclusions</td>
<td>138</td>
</tr>
</tbody>
</table>

### III DEPARTURES FROM EFFICIENCY

### 6 Public Goods

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1 Introduction</td>
<td>147</td>
</tr>
<tr>
<td>6.2 Definitions</td>
<td>148</td>
</tr>
<tr>
<td>6.3 Private Provision</td>
<td>150</td>
</tr>
<tr>
<td>6.4 Efficient Provision</td>
<td>154</td>
</tr>
<tr>
<td>6.5 Voting</td>
<td>156</td>
</tr>
<tr>
<td>6.6 Personalized Prices</td>
<td>159</td>
</tr>
<tr>
<td>6.7 Mechanism Design</td>
<td>163</td>
</tr>
<tr>
<td>6.7.1 Examples of Preference Revelation</td>
<td>163</td>
</tr>
<tr>
<td>6.7.2 Clarke–Groves Mechanism</td>
<td>166</td>
</tr>
<tr>
<td>6.7.3 Clarke Tax</td>
<td>168</td>
</tr>
<tr>
<td>6.7.4 Further Comments</td>
<td>169</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>6.8 More on Private Provision</td>
<td>170</td>
</tr>
<tr>
<td>6.8.1 Neutrality and Population Size</td>
<td>170</td>
</tr>
<tr>
<td>6.8.2 Experimental Evidence</td>
<td>174</td>
</tr>
<tr>
<td>6.8.3 Modifications</td>
<td>176</td>
</tr>
<tr>
<td>6.9 Fund-Raising Campaigns</td>
<td>178</td>
</tr>
<tr>
<td>6.9.1 The Contribution Campaign</td>
<td>179</td>
</tr>
<tr>
<td>6.9.2 The Subscription Campaign</td>
<td>181</td>
</tr>
<tr>
<td>6.10 Conclusions</td>
<td>182</td>
</tr>
<tr>
<td>7 Club Goods and Local Public Goods</td>
<td>191</td>
</tr>
<tr>
<td>7.1 Introduction</td>
<td>191</td>
</tr>
<tr>
<td>7.2 Definitions</td>
<td>192</td>
</tr>
<tr>
<td>7.3 Single-Product Clubs</td>
<td>193</td>
</tr>
<tr>
<td>7.3.1 Fixed Utilization</td>
<td>194</td>
</tr>
<tr>
<td>7.3.2 Variable Utilization</td>
<td>196</td>
</tr>
<tr>
<td>7.3.3 Two-Part Tariff</td>
<td>197</td>
</tr>
<tr>
<td>7.4 Clubs and the Economy</td>
<td>199</td>
</tr>
<tr>
<td>7.4.1 Small Clubs</td>
<td>200</td>
</tr>
<tr>
<td>7.4.2 Large Clubs</td>
<td>200</td>
</tr>
<tr>
<td>7.4.3 Conclusion</td>
<td>207</td>
</tr>
<tr>
<td>7.5 Local Public Goods</td>
<td>208</td>
</tr>
<tr>
<td>7.6 The Tiebout Hypothesis</td>
<td>212</td>
</tr>
<tr>
<td>7.7 Empirical Tests</td>
<td>214</td>
</tr>
<tr>
<td>7.8 Conclusions</td>
<td>216</td>
</tr>
<tr>
<td>8 Externalities</td>
<td>223</td>
</tr>
<tr>
<td>8.1 Introduction</td>
<td>223</td>
</tr>
<tr>
<td>8.2 Externalities Defined</td>
<td>224</td>
</tr>
<tr>
<td>8.3 Market Inefficiency</td>
<td>225</td>
</tr>
<tr>
<td>8.4 Externality Examples</td>
<td>228</td>
</tr>
<tr>
<td>8.4.1 River Pollution</td>
<td>228</td>
</tr>
<tr>
<td>8.4.2 Traffic Jams</td>
<td>229</td>
</tr>
<tr>
<td>8.4.3 Pecuniary Externality</td>
<td>230</td>
</tr>
<tr>
<td>8.4.4 The Rat Race Problem</td>
<td>232</td>
</tr>
<tr>
<td>8.4.5 The Tragedy of the Commons</td>
<td>233</td>
</tr>
<tr>
<td>8.4.6 Bandwagon Effect</td>
<td>235</td>
</tr>
<tr>
<td>8.5 Pigouvian Taxation</td>
<td>236</td>
</tr>
<tr>
<td>8.6 Licenses</td>
<td>239</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>8.7</td>
<td>Internalization</td>
</tr>
<tr>
<td>8.8</td>
<td>The Coase Theorem</td>
</tr>
<tr>
<td>8.9</td>
<td>Nonconvexity</td>
</tr>
<tr>
<td>8.10</td>
<td>Conclusions</td>
</tr>
<tr>
<td>9</td>
<td>Imperfect Competition</td>
</tr>
<tr>
<td>9.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>9.2</td>
<td>Concepts of Competition</td>
</tr>
<tr>
<td>9.3</td>
<td>Market Structure</td>
</tr>
<tr>
<td>9.3.1</td>
<td>Defining the Market</td>
</tr>
<tr>
<td>9.3.2</td>
<td>Measuring Competition</td>
</tr>
<tr>
<td>9.4</td>
<td>Welfare</td>
</tr>
<tr>
<td>9.4.1</td>
<td>Inefficiency</td>
</tr>
<tr>
<td>9.4.2</td>
<td>Incomplete Information</td>
</tr>
<tr>
<td>9.4.3</td>
<td>Measures of Welfare Loss</td>
</tr>
<tr>
<td>9.5</td>
<td>Tax Incidence</td>
</tr>
<tr>
<td>9.6</td>
<td>Specific and Ad valorem Taxation</td>
</tr>
<tr>
<td>9.7</td>
<td>Regulation of Monopoly</td>
</tr>
<tr>
<td>9.8</td>
<td>Regulation of Oligopoly</td>
</tr>
<tr>
<td>9.8.1</td>
<td>Detecting Collusion</td>
</tr>
<tr>
<td>9.8.2</td>
<td>Merger Policy</td>
</tr>
<tr>
<td>9.9</td>
<td>Unions and Taxation</td>
</tr>
<tr>
<td>9.10</td>
<td>Monopsony</td>
</tr>
<tr>
<td>9.11</td>
<td>Conclusions</td>
</tr>
<tr>
<td>10</td>
<td>Asymmetric Information</td>
</tr>
<tr>
<td>10.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>10.2</td>
<td>Hidden Knowledge and Hidden Action</td>
</tr>
<tr>
<td>10.3</td>
<td>Actions or Knowledge?</td>
</tr>
<tr>
<td>10.4</td>
<td>Market Unraveling</td>
</tr>
<tr>
<td>10.4.1</td>
<td>Hazard Insurance</td>
</tr>
<tr>
<td>10.4.2</td>
<td>Government Intervention</td>
</tr>
<tr>
<td>10.5</td>
<td>Screening</td>
</tr>
<tr>
<td>10.5.1</td>
<td>Perfect Information Equilibrium</td>
</tr>
<tr>
<td>10.5.2</td>
<td>Imperfect Information Equilibrium</td>
</tr>
<tr>
<td>10.5.3</td>
<td>Government Intervention</td>
</tr>
<tr>
<td>10.6</td>
<td>Signaling</td>
</tr>
<tr>
<td>10.6.1</td>
<td>Educational Signaling</td>
</tr>
<tr>
<td>10.6.2</td>
<td>Implications</td>
</tr>
</tbody>
</table>
10.7 Moral Hazard (Hidden Action) 323
  10.7.1 Moral Hazard in Insurance 323
  10.7.2 Effort Observable 325
  10.7.3 Effort Unobservable 326
  10.7.4 Second-Best Contract 327
  10.7.5 Government Intervention 329
10.8 Public Provision of Health Care 330
  10.8.1 Efficiency 330
  10.8.2 Redistributive Politics 332
10.9 Evidence 334
10.10 Conclusions 336

IV POLITICAL ECONOMY

11 Voting 345
  11.1 Introduction 345
  11.2 Stability 345
  11.3 Impossibility 347
  11.4 Majority Rule 350
    11.4.1 May’s Theorem 350
    11.4.2 Condorcet Winner 351
    11.4.3 Median Voter Theorems 351
    11.4.4 Multidimensional Voting 356
    11.4.5 Agenda Manipulation 358
  11.5 Alternatives to Majority Rule 361
    11.5.1 Borda Voting 361
    11.5.2 Plurality Voting 363
    11.5.3 Approval Voting 363
    11.5.4 Runoff Voting 364
  11.6 The Paradox of Voting 365
  11.7 The “Alabama” Paradox 371
  11.8 Political Competition 372
    11.8.1 Downsian Model 372
    11.8.2 Policy Divergence 373
    11.8.3 Multidimensional Competition 376
    11.8.4 “Swing Voter” Politics 377
    11.8.5 Citizen-Candidate Model 379
  11.9 Conclusions 381
### 12 Rent-Seeking

12.1 Introduction 387
12.2 Definitions 388
12.3 Rent-Seeking Games
   12.3.1 Deterministic Game 391
   12.3.2 Probabilistic Game 394
   12.3.3 Free-Entry 396
   12.3.4 Risk Aversion 397
   12.3.5 Conclusions 398
12.4 Social Cost of Monopoly 398
12.5 Equilibrium Effects 401
12.6 Government Policy
   12.6.1 Lobbying 404
   12.6.2 Rent Creation 406
   12.6.3 Conclusions 408
12.7 Informative Lobbying 408
12.8 Controlling Rent-Seeking 413
12.9 Conclusions 414

### V EQUITY AND DISTRIBUTION

### 13 Optimality and Comparability

13.1 Introduction 423
13.2 Social Optimality 424
13.3 Lump-Sum Taxes 427
13.4 Impossibility of Lump-Sum Taxes 430
13.5 Redistribution In-Kind 434
13.6 Aspects of Pareto-Efficiency 436
13.7 Social Welfare Functions 440
13.8 Arrow’s Theorem 441
13.9 Interpersonal Comparability 443
13.10 Comparability and Social Welfare 446
13.11 Conclusions 450

### 14 Inequality and Poverty

14.1 Introduction 457
14.2 Measuring Income 458
14.3 Equivalence Scales 460
14.4 Inequality Measurement 466
14.4.1 The Setting 467
14.4.2 Statistical Measures 467
14.4.3 Inequality and Welfare 475
14.4.4 An Application 480
14.5 Poverty 481
14.5.1 Poverty and the Poverty Line 482
14.5.2 Poverty Measures 483
14.5.3 Two Applications 488
14.6 Unequal Opportunities 489
14.6.1 Defining Equality of Opportunity 490
14.6.2 Measuring Equality of Opportunity 491
14.6.3 Equal-Opportunity Policy 491
14.7 Intergenerational Inequality 492
14.7.1 Measuring Issues 492
14.7.2 Causal Mechanisms 493
14.8 Conclusions 496

VI TAXATION

15 Commodity Taxation 505
15.1 Introduction 505
15.2 Deadweight Loss 506
15.3 Optimal Taxation 509
15.4 Production Efficiency 513
15.5 Tax Rules 515
15.5.1 Inverse Elasticity Rule 516
15.5.2 Ramsey Rule 517
15.6 Equity Considerations 521
15.7 Applications 523
15.7.1 Reform 524
15.7.2 Optimality 526
15.8 Efficient Taxation 528
15.9 Public Sector Pricing 530
15.10 Conclusions 531

16 Income Taxation 537
16.1 Introduction 537
16.2 Equity and Efficiency 538
### Contents

#### 23 Social Security
- 23.1 Introduction
- 23.2 Types of System
- 23.3 The Pensions Crisis
- 23.4 The Simplest Program
- 23.5 Social Security and Production
- 23.6 Population Growth
- 23.7 Sustaining a Program
- 23.8 Ricardian Equivalence
- 23.9 Social Security Reform
- 23.10 Conclusions

#### 24 Economic Growth
- 24.1 Introduction
- 24.2 Exogenous Growth
  - 24.2.1 Constant Savings Rate
  - 24.2.2 Optimal Taxation
- 24.3 Endogenous Growth
  - 24.3.1 Models of Endogenous Growth
  - 24.3.2 Government Expenditure
- 24.4 Policy Reform
- 24.5 Empirical Evidence
- 24.6 Conclusions

#### IX Applications

#### 25 Cost–Benefit Analysis
- 25.1 Introduction
- 25.2 What Is Cost–Benefit Analysis?
  - 25.2.1 Simple Example
  - 25.2.2 Broader Approach
- 25.3 The Process of CBA
  - 25.3.1 Discounting Future Values
- 25.4 Principles of CBA
  - 25.4.1 Common Unit of Measurement
  - 25.4.2 Revealed Preferences
  - 25.4.3 Valuing Market Goods: Marshallian Surplus
We have been very pleased with the reception given to the first edition of *Intermediate Public Economics*. We wish to thank all our colleagues in universities across the world who have adopted the book to support their teaching. Their helpful comments and encouragement have provided the motivation to undertake this substantive revision.

The second edition of *Intermediate Public Economic* reflects recent developments in the academic literature and in policy debate. New chapters are included on behavioral economics, cost–benefit analysis, economics of climate policy, and international taxation. A chapter has also been added on the limits to redistribution that complements the existing chapters on commodity taxation and income taxation. The advice of numerous users of the first edition has resulted in many other parts of the book having been improved by the expansion of some arguments and the clarification of others.

The book is written with the intention that it should be accessible to anyone with a background of intermediate microeconomics and macroeconomics. The level and quantity of mathematics is kept as low as possible without sacrificing intellectual rigor, but the book remains analytical rather than discursive. Some fairly abstract arguments are included—particularly in the new chapter on redistribution—where it has been judged that these make sufficiently important points that the extra burden imposed on the reader will be amply rewarded.

Additional exercises have been added at the end of each chapter. Correspondingly, a new *Solutions Manual* is available to accompany the book. Powerpoint slides for each of the chapters are also available.

The list of people to whom we are indebted for directly or indirectly assisting with the preparation of this book continues to grow. For assistance with the second edition we wish to thank Nizar Allouch, Douglas Bernheim, Thierry Bréchet, James Davidson, Pascale Duran-Vigneron, Alessandro Ferrera, Christian Gollier, Roger Guesnerie, Andreea Halunga, Nigar Hashimzade, Tatiana Kirsanova, Christos Kotsogiannis, Etienne Lehmann, Cuong Le Van, Miguel-Angel Lopez-Garcia, Miltos Makris, François Maniquet, Ross McKitrick, Sushama Murthy, Pierre Pestieau, Matthew Rablen, Emmanuelle Taugourdeau, Henry Tulkens, Unal Zenginobuz, and the members of the Mirrlees Review of the UK tax system.

We also want to thank again the people we thanked for the first edition, many of whom have also contributed to this edition: Paul Belleflamme, Tim Besley, Chuck Blackorby, Christopher Bliss, Craig Brett, John Conley, Richard Cornes, Philippe De Donder, Sanjit Dhami, Peter Diamond, Jean Gabszewicz, Peter Hammond, Arye

Jean notes that the chapter on behavioral economics has benefited from the inspiring lectures given by Douglas Bernheim and Vincent Crawford at the 2011 ECORE Summer School in Louvain-la-Neuve, Belgium. He also wishes to thank his Master’s degree students at UCL who tested several of the chapters. Gareth wishes to thank ATAX (University of New South Wales), ENS Cachan, NYU in London, and Universitat Autònoma de Barcelona for generous hospitality while parts of the book were written and tested. He also wishes to thank the Association of Public Economic Theory for providing such excellent conferences that make it possible to keep in touch with the book’s users.

We observed in the Preface to the first edition that Public Economics is about the government and the economic effects of its policies. We should also observe that public economics continually absorbs new ideas from other fields and uses these whenever they can improve the understanding of policy. This book offers an extended insight into what Public Economics says and what it can do. We hope that you enjoy it.

Jean Hindriks
Louvain La Neuve
Gareth Myles
Exeter
February 2012
Preface to First Edition

This book has been prepared as the basis for a final-year undergraduate or first-year graduate course in Public Economics. It is based on lectures given by the authors at several institutions over many years. It covers the traditional topics of efficiency and equity but also emphasizes more recent developments in information, games, and, especially, political economy.

The book should be accessible to anyone with a background of intermediate microeconomics and macroeconomics. We have deliberately kept the quantity of math as low as we could without sacrificing intellectual rigor. Even so, the book remains analytical rather than discursive.

To support the content, further reading is given for each chapter. This reading is intended to offer a range of material from the classic papers in each area through recent contributions to surveys and critiques. Exercises are included for each chapter. Most of these should be possible for a good undergraduate but some may prove challenging.

There are many people who have contributed directly or indirectly to the preparation of this book. Nigar Hashimzade is entitled to special thanks for making incisive comments on the entire text and for assisting with the analysis in chapters 11 and 24. Thanks are also due to Paul Belleflamme, Tim Besley, Chuck Blackorby, Christopher Bliss, Craig Brett, John Conley, Richard Cornes, Philippe De Donder, Sanjit Dhami, Peter Diamond, Jean Gabszewicz, Peter Hammond, Arye Hillman, Norman Ireland, Michael Keen, Francois Maniquet, Jack Mintz, James Mirrlees, Frank Page Jr., Susana Peralta, Pierre Pestieau, Pierre M. Picard, Ian Preston, Maria Racionero, Antonio Rangel, Les Reinhorn, Elena del Rey, Todd Sandler, Kim Scharf, Hyun Shin, Michael Smart, Stephen Smith, Klaas Staal, Jacques Thisse, Harrie Verbon, John Weymark, David Wildasin, and Myrna Wooders.

Public Economics is about the government and the economic effects of its policies. This book offers an insight into what Public Economics says and what it can do. We hope that you enjoy it.

Jean Hindriks  
*Louvain La Neuve*  
Gareth Myles  
*Exeter*  
*February 2005*
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Typical Edgeworth box</td>
<td>17</td>
</tr>
<tr>
<td>2.2</td>
<td>Preferences and demand</td>
<td>18</td>
</tr>
<tr>
<td>2.3</td>
<td>Relative price change</td>
<td>19</td>
</tr>
<tr>
<td>2.4</td>
<td>Equilibrium</td>
<td>19</td>
</tr>
<tr>
<td>2.5</td>
<td>Equilibrium and excess demand</td>
<td>21</td>
</tr>
<tr>
<td>2.6</td>
<td>Typical production set</td>
<td>23</td>
</tr>
<tr>
<td>2.7</td>
<td>Profit maximization</td>
<td>24</td>
</tr>
<tr>
<td>2.8</td>
<td>Utility maximization</td>
<td>28</td>
</tr>
<tr>
<td>2.9</td>
<td>Efficient equilibrium</td>
<td>29</td>
</tr>
<tr>
<td>2.10</td>
<td>Constant returns to scale</td>
<td>31</td>
</tr>
<tr>
<td>2.11</td>
<td>Pareto-efficiency</td>
<td>33</td>
</tr>
<tr>
<td>2.12</td>
<td>The First Theorem</td>
<td>34</td>
</tr>
<tr>
<td>2.13</td>
<td>The Second Theorem</td>
<td>36</td>
</tr>
<tr>
<td>2.14</td>
<td>Proof of the Second Theorem</td>
<td>39</td>
</tr>
<tr>
<td>2.15</td>
<td>A lump-sum transfer</td>
<td>40</td>
</tr>
<tr>
<td>4.1</td>
<td>Total expenditure, 1870 to 2010 (% of GDP)</td>
<td>90</td>
</tr>
<tr>
<td>4.2</td>
<td>Total expenditure, 1970 to 2011 (% of GDP)</td>
<td>91</td>
</tr>
<tr>
<td>4.3</td>
<td>Individual expenditure items (% of GDP)</td>
<td>92</td>
</tr>
<tr>
<td>4.4</td>
<td>Government expenditure, 2010 (% of GDP)</td>
<td>94</td>
</tr>
<tr>
<td>4.5</td>
<td>Share of expenditure by levels of government, 2009</td>
<td>95</td>
</tr>
<tr>
<td>4.6</td>
<td>Composition of general spending, 2007</td>
<td>96</td>
</tr>
<tr>
<td>4.7</td>
<td>Composition of central spending</td>
<td>98</td>
</tr>
<tr>
<td>4.8</td>
<td>Composition of state spending</td>
<td>99</td>
</tr>
<tr>
<td>4.9</td>
<td>Composition of local spending</td>
<td>100</td>
</tr>
<tr>
<td>4.10</td>
<td>Tax revenues, 1965 to 2009 (% of GDP)</td>
<td>101</td>
</tr>
<tr>
<td>4.11</td>
<td>Tax revenue for category of taxation, 2009</td>
<td>101</td>
</tr>
<tr>
<td>4.12</td>
<td>Tax revenue by level of government, federal countries, 2009</td>
<td>102</td>
</tr>
<tr>
<td>4.13</td>
<td>Tax revenue by level of government, unitary countries, 2009</td>
<td>103</td>
</tr>
<tr>
<td>4.14</td>
<td>Tax shares at each level of government, United States, 2009</td>
<td>104</td>
</tr>
</tbody>
</table>
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.15</td>
<td>Tax shares at each level of government, Germany, 2009</td>
<td>105</td>
</tr>
<tr>
<td>4.16</td>
<td>Tax shares at each level of government, Japan, 2009</td>
<td>105</td>
</tr>
<tr>
<td>4.17</td>
<td>Tax shares at each level of government, United Kingdom, 2009</td>
<td>106</td>
</tr>
<tr>
<td>4.18</td>
<td>Budget deficit (% of GDP)</td>
<td>108</td>
</tr>
<tr>
<td>4.19</td>
<td>Corporate debt of nonfinancial corporations (% of GDP)</td>
<td>109</td>
</tr>
<tr>
<td>4.20</td>
<td>Government debt (% of GDP)</td>
<td>110</td>
</tr>
<tr>
<td>4.21</td>
<td>Household debt (% of GDP)</td>
<td>110</td>
</tr>
<tr>
<td>4.22</td>
<td>Aggregate debt of the 17 euro area countries (% of GDP)</td>
<td>111</td>
</tr>
<tr>
<td>4.23</td>
<td>Debt in euro area countries (% of GDP)</td>
<td>112</td>
</tr>
<tr>
<td>5.1</td>
<td>Excessive government</td>
<td>132</td>
</tr>
<tr>
<td>5.2</td>
<td>Government agency</td>
<td>136</td>
</tr>
<tr>
<td>6.1</td>
<td>Typology of goods</td>
<td>149</td>
</tr>
<tr>
<td>6.2</td>
<td>Preferences and choice</td>
<td>151</td>
</tr>
<tr>
<td>6.3</td>
<td>Best reaction for consumer 2</td>
<td>152</td>
</tr>
<tr>
<td>6.4</td>
<td>Nash equilibrium</td>
<td>153</td>
</tr>
<tr>
<td>6.5</td>
<td>Inefficiency of equilibrium</td>
<td>153</td>
</tr>
<tr>
<td>6.6</td>
<td>Allocation through voting</td>
<td>157</td>
</tr>
<tr>
<td>6.7</td>
<td>Lindahl equilibrium</td>
<td>161</td>
</tr>
<tr>
<td>6.8</td>
<td>Gaining by false announcement</td>
<td>162</td>
</tr>
<tr>
<td>6.9</td>
<td>Announcements and payoffs</td>
<td>164</td>
</tr>
<tr>
<td>6.10</td>
<td>Payoffs and overstatement</td>
<td>165</td>
</tr>
<tr>
<td>6.11</td>
<td>Clarke–Groves mechanism</td>
<td>167</td>
</tr>
<tr>
<td>6.12</td>
<td>Payoffs for player 1</td>
<td>167</td>
</tr>
<tr>
<td>6.13</td>
<td>Effect of a transfer</td>
<td>171</td>
</tr>
<tr>
<td>6.14</td>
<td>New equilibrium</td>
<td>172</td>
</tr>
<tr>
<td>6.15</td>
<td>Additional consumers</td>
<td>174</td>
</tr>
<tr>
<td>6.16</td>
<td>Public good experiment</td>
<td>175</td>
</tr>
<tr>
<td>6.17</td>
<td>Contribution campaign</td>
<td>180</td>
</tr>
<tr>
<td>7.1</td>
<td>Utility and club size</td>
<td>201</td>
</tr>
<tr>
<td>7.2</td>
<td>Optimum with unequally sized clubs</td>
<td>203</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>7.3</td>
<td>Optimum with equally sized clubs</td>
<td>205</td>
</tr>
<tr>
<td>7.4</td>
<td>Non-unique club size</td>
<td>206</td>
</tr>
<tr>
<td>7.5</td>
<td>Achieving efficiency</td>
<td>207</td>
</tr>
<tr>
<td>7.6</td>
<td>Stability of the symmetric equilibrium</td>
<td>210</td>
</tr>
<tr>
<td>7.7</td>
<td>Inefficient stable equilibrium</td>
<td>211</td>
</tr>
<tr>
<td>8.1</td>
<td>Deviation of private from social benefits</td>
<td>227</td>
</tr>
<tr>
<td>8.2</td>
<td>Equilibrium with river pollution</td>
<td>229</td>
</tr>
<tr>
<td>8.3</td>
<td>Choice of commuting mode</td>
<td>230</td>
</tr>
<tr>
<td>8.4</td>
<td>Job choice</td>
<td>231</td>
</tr>
<tr>
<td>8.5</td>
<td>Rat race</td>
<td>232</td>
</tr>
<tr>
<td>8.6</td>
<td>Tragedy of the commons</td>
<td>234</td>
</tr>
<tr>
<td>8.7</td>
<td>Equilibrium keyboard choice</td>
<td>235</td>
</tr>
<tr>
<td>8.8</td>
<td>Pigouvian taxation</td>
<td>237</td>
</tr>
<tr>
<td>8.9</td>
<td>Uncertain costs</td>
<td>241</td>
</tr>
<tr>
<td>8.10</td>
<td>Nonconvexity</td>
<td>247</td>
</tr>
<tr>
<td>9.1</td>
<td>Monopoly pricing</td>
<td>265</td>
</tr>
<tr>
<td>9.2</td>
<td>Deadweight loss with monopoly</td>
<td>266</td>
</tr>
<tr>
<td>9.3</td>
<td>Tax incidence with perfectly elastic supply</td>
<td>269</td>
</tr>
<tr>
<td>9.4</td>
<td>Tax incidence in the general case</td>
<td>269</td>
</tr>
<tr>
<td>9.5</td>
<td>Tax undershifting</td>
<td>270</td>
</tr>
<tr>
<td>9.6</td>
<td>Tax overshifting</td>
<td>271</td>
</tr>
<tr>
<td>9.7</td>
<td>Possibility of a profit increase</td>
<td>273</td>
</tr>
<tr>
<td>9.8</td>
<td>Contrasting taxes</td>
<td>276</td>
</tr>
<tr>
<td>9.9</td>
<td>Natural monopoly</td>
<td>279</td>
</tr>
<tr>
<td>9.10</td>
<td>Monopsony in the labor market</td>
<td>287</td>
</tr>
<tr>
<td>10.1</td>
<td>Equilibrium in the insurance market</td>
<td>304</td>
</tr>
<tr>
<td>10.2</td>
<td>Multiple equilibria</td>
<td>306</td>
</tr>
<tr>
<td>10.3</td>
<td>Perfect information equilibrium</td>
<td>309</td>
</tr>
<tr>
<td>10.4</td>
<td>Separating contracts</td>
<td>311</td>
</tr>
<tr>
<td>10.5</td>
<td>Separating and pooling contracts</td>
<td>312</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>--------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>10.6</td>
<td>Nonexistence of pooling equilibrium</td>
<td>313</td>
</tr>
<tr>
<td>10.7</td>
<td>Market intervention</td>
<td>314</td>
</tr>
<tr>
<td>10.8</td>
<td>Single-crossing property</td>
<td>317</td>
</tr>
<tr>
<td>10.9</td>
<td>Separating equilibrium</td>
<td>319</td>
</tr>
<tr>
<td>10.10</td>
<td>Unreasonable beliefs</td>
<td>322</td>
</tr>
<tr>
<td>10.11</td>
<td>First-best contract</td>
<td>326</td>
</tr>
<tr>
<td>10.12</td>
<td>Switching line</td>
<td>327</td>
</tr>
<tr>
<td>10.13</td>
<td>Second-best contract</td>
<td>328</td>
</tr>
<tr>
<td>11.1</td>
<td>Location of households</td>
<td>352</td>
</tr>
<tr>
<td>11.2</td>
<td>Single-peaked preferences</td>
<td>353</td>
</tr>
<tr>
<td>11.3</td>
<td>Single-crossing without single-peakedness</td>
<td>355</td>
</tr>
<tr>
<td>11.4</td>
<td>Single-peakedness in multi-dimensions</td>
<td>357</td>
</tr>
<tr>
<td>11.5</td>
<td>Binary agenda</td>
<td>359</td>
</tr>
<tr>
<td>11.6</td>
<td>Probabilities of election</td>
<td>367</td>
</tr>
<tr>
<td>11.7</td>
<td>Participation and the probability of being pivotal</td>
<td>368</td>
</tr>
<tr>
<td>11.8</td>
<td>Closeness and the probability of being pivotal</td>
<td>369</td>
</tr>
<tr>
<td>12.1</td>
<td>Monopoly deadweight loss</td>
<td>399</td>
</tr>
<tr>
<td>12.2</td>
<td>Competitive and monopoly equilibria</td>
<td>402</td>
</tr>
<tr>
<td>12.3</td>
<td>Monopoly and lobbying</td>
<td>402</td>
</tr>
<tr>
<td>12.4</td>
<td>Threat of lobbying</td>
<td>403</td>
</tr>
<tr>
<td>13.1</td>
<td>Utility possibility frontier</td>
<td>425</td>
</tr>
<tr>
<td>13.2</td>
<td>Social indifference curves</td>
<td>426</td>
</tr>
<tr>
<td>13.3</td>
<td>Social optimality</td>
<td>427</td>
</tr>
<tr>
<td>13.4</td>
<td>Optimal lump-sum taxes and incentive compatibility</td>
<td>433</td>
</tr>
<tr>
<td>13.5</td>
<td>Efficiency and inequity</td>
<td>438</td>
</tr>
<tr>
<td>13.6</td>
<td>Incompleteness of Pareto ranking</td>
<td>439</td>
</tr>
<tr>
<td>14.1</td>
<td>Construction of Engel scale</td>
<td>463</td>
</tr>
<tr>
<td>14.2</td>
<td>Adult good equivalence scale</td>
<td>464</td>
</tr>
<tr>
<td>14.3</td>
<td>General equivalence scale</td>
<td>465</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td>------</td>
</tr>
<tr>
<td>14.4</td>
<td>Construction of a Lorenz curve</td>
<td>472</td>
</tr>
<tr>
<td>14.5</td>
<td>Lorenz curves as an incomplete ranking</td>
<td>473</td>
</tr>
<tr>
<td>14.6</td>
<td>Relating Gini to Lorenz</td>
<td>474</td>
</tr>
<tr>
<td>14.7</td>
<td>Gini social welfare</td>
<td>477</td>
</tr>
<tr>
<td>14.8</td>
<td>Equally distributed equivalent income</td>
<td>479</td>
</tr>
<tr>
<td>15.1</td>
<td>Deadweight loss</td>
<td>506</td>
</tr>
<tr>
<td>15.2</td>
<td>Income and substitution effects</td>
<td>507</td>
</tr>
<tr>
<td>15.3</td>
<td>Absence of deadweight loss</td>
<td>508</td>
</tr>
<tr>
<td>15.4</td>
<td>Revenue and production possibilities</td>
<td>510</td>
</tr>
<tr>
<td>15.5</td>
<td>Consumer choice</td>
<td>511</td>
</tr>
<tr>
<td>15.6</td>
<td>Optimal commodity taxation</td>
<td>512</td>
</tr>
<tr>
<td>15.7</td>
<td>Production efficiency</td>
<td>514</td>
</tr>
<tr>
<td>16.1</td>
<td>Labor supply decision</td>
<td>540</td>
</tr>
<tr>
<td>16.2</td>
<td>Effect of a wage increase</td>
<td>541</td>
</tr>
<tr>
<td>16.3</td>
<td>Effect of a tax increase</td>
<td>542</td>
</tr>
<tr>
<td>16.4</td>
<td>A tax threshold</td>
<td>542</td>
</tr>
<tr>
<td>16.5</td>
<td>Several thresholds</td>
<td>543</td>
</tr>
<tr>
<td>16.6</td>
<td>Taxation and the participation decision</td>
<td>544</td>
</tr>
<tr>
<td>16.7</td>
<td>Taxation and the consumption function</td>
<td>548</td>
</tr>
<tr>
<td>16.8</td>
<td>Agent monotonicity</td>
<td>549</td>
</tr>
<tr>
<td>16.9</td>
<td>Income and skill</td>
<td>549</td>
</tr>
<tr>
<td>16.10</td>
<td>Upper limit on tax rate</td>
<td>550</td>
</tr>
<tr>
<td>16.11</td>
<td>Initial position and new tax function</td>
<td>551</td>
</tr>
<tr>
<td>16.12</td>
<td>Transfer of consumption</td>
<td>552</td>
</tr>
<tr>
<td>16.13</td>
<td>Allowing relocation</td>
<td>552</td>
</tr>
<tr>
<td>16.14</td>
<td>Zero marginal rate of tax</td>
<td>553</td>
</tr>
<tr>
<td>16.15</td>
<td>Allocations and the consumption function</td>
<td>556</td>
</tr>
<tr>
<td>16.16</td>
<td>Binding incentive compatibility</td>
<td>557</td>
</tr>
<tr>
<td>17.1</td>
<td>Interior choice: (0 &lt; X^* &lt; Y)</td>
<td>578</td>
</tr>
<tr>
<td>17.2</td>
<td>Corner solutions</td>
<td>578</td>
</tr>
</tbody>
</table>
### List of Figures

#### 17.3 Increase in detection probability 581
#### 17.4 Increase in the fine rate 581
#### 17.5 Income increase 582
#### 17.6 Tax rate increase 583
#### 17.7 Audit game 591
#### 17.8 Equilibrium compliance 600

#### 18.1 Mechanism design 612
#### 18.2 Direct and indirect mechanisms 613
#### 18.3 Allocation with no redistribution 617
#### 18.4 Maximal redistribution 618
#### 18.5 Decentralization by taxation 619
#### 18.6 Differential taxation and nonseparability 623

#### 19.1 The costs of uniformity 638
#### 19.2 Centralization and decentralization 641
#### 19.3 Political accountability and voter welfare 645
#### 19.4 Regional distribution of income 647

#### 20.1 Allocation of capital 668
#### 20.2 Symmetric Nash equilibrium 671
#### 20.3 Advantage of smallness 674
#### 20.4 Tax overlap 677
#### 20.5 Matching versus lump-sum grant 688

#### 21.1 Profits and effective corporate tax rates 712
#### 21.2 Effect of a tariff on a small country 734
#### 21.3 Effect of a tariff on a large country 735

#### 22.1 Generational structure 754
#### 22.2 Consumer choice 756
#### 22.3 Consumption possibilities 761
#### 22.4 Steady-state equilibrium 762
#### 22.5 Golden Rule and competitive equilibrium 765
List of Figures

23.1 Annual income and cost forecast for OASI 780
23.2 Pareto improvement and social security 782
23.3 Population growth and consumption 788
23.4 Population growth and consumption 789
23.5 Population growth and social security 789
23.6 Social security game 792

24.1 Dynamics of the capital stock 812
24.2 Steady state 813
24.3 Golden Rule 815
24.4 Consumption and the saving rate 816
24.5 Taxation and consumption growth 829
24.6 Level and growth effects 830
24.7 Growth effects of tax reform 832
24.8 US tax and growth rates 833
24.9 UK tax and growth rates 834
24.10 Average and marginal tax rates 835

25.1 Increase in gross benefit 858
25.2 Official values for a road accident fatality 869
25.3 Intertemporal equilibrium 872
25.4 Scitovsky reversal 879

26.1 The warming process 901
26.2 Accumulation of greenhouse gases, CO$_2$e ppm 902
26.3 Global average temperature (annual average difference from normal) 903
26.4 Hockey stick plot 903
26.5 Global average temperature (monthly, difference from normal) 905
26.6 Estimates of climate sensitivity 908
26.7 Predicted changes in climate patterns 910
26.8 Industrial CO$_2$ emissions: DICE and two IPCC scenarios 916
26.9 Discount rates 918
26.10 Ecological value 928
List of Figures

26.11 Coverage of the EU Emissions Trading Scheme 932
26.12 Allowance price evolution in the EU ETS, 2005 to 2008 933
26.13 Effect of a climate tax 934
26.14 CO\textsubscript{2} emissions 2007, metric tons 936
26.15 CO\textsubscript{2} per person 2007, metric tons 937
1 An Introduction to Public Economics

1.1 Public Economics

The study of public economics has a long tradition. It developed out of the original political economy of John Stuart Mill and David Ricardo, through the public finance tradition of tax analysis into public economics, and has now returned to its roots with the development of the new political economy. From the inception of economics as a scientific discipline, public economics has always been one of its core branches. The explanation for why it has always been so central is the foundation that it provides for practical policy analysis. This has always been the motivation of public economists, even if the issues studied and the analytical methods employed have evolved over time. We intend the theory described in this book to provide an organized and coherent structure for addressing economic policy in a rigorous and coherent way.

In the broadest interpretation, public economics is the study of economic efficiency, distribution, and government economic policy. The subject encompasses topics as diverse as responses to market failure due to the existence of externalities, the motives for tax evasion, and the explanation of bureaucratic decision-making. In order to reach into all of these areas, public economics has developed from its initial narrow focus on the collection and spending of government revenues to its present concern with every aspect of government interaction with the economy.

Public economics attempts to understand both how the government makes decisions and what decisions it should make. To understand how the government makes decisions, it is necessary to investigate the motives of the decision makers within government, how the decision makers are chosen, and how they are influenced by outside forces. Determining what decisions should be made involves studying the effects of the alternative policies that are available and evaluating the outcomes to which they lead. These aspects are interwoven throughout the text. By pulling them together, this book provides an accessible introduction to both of these aspects of public economics.

1.2 Methods

The feature that most characterizes modern public economics is the use made of economic models. These models are employed as a tool to ensure that arguments are conducted coherently with a rigorous logical basis. Models are used for analysis because
the possibilities for experimentation are limited and past experience cannot always be relied on to provide a guide to the consequences of new policies. Each model is intended to be a simplified description of the part of the economy that is relevant for the analysis. What distinguishes economic models from those in the natural sciences is the incorporation of independent decision-making by the firms, consumers, and politicians that populate the economy. These actors in the economy do not respond mechanically but are motivated by personal objectives and are strategic in their behavior. Capturing the implications of this complex behavior in a convincing manner is one of the key skills of a successful economic modeler.

Once a model has been chosen, its implications have to be derived. These implications are obtained by applying logical arguments that proceed from the assumptions of the model to a set of formally correct conclusions. Those conclusions then need to be given an interpretation in terms that can be related to the original question of interest. Policy recommendations can subsequently be derived but always with a recognition of the limitations of the model. A careful line has to be drawn between conclusions that are justifiable given the model and conclusions that go beyond what the model can support.

The institutional setting for the study of public economics is invariably the mixed economy where individual decisions are respected but the government attempts to affect these through the policies it implements. Within this environment many alternative objectives can be assigned to the government. For instance, the government can be assumed to care about the aggregate level of welfare in the economy and to act selflessly in attempting to increase this. Such a viewpoint is the foundation of optimal policy analysis that inquires how the government should behave. But there can be no presumption that actual governments act in this way. An alternative, and sometimes more compelling view, is that the government is composed of a set of individuals pursuing their own selfish agenda. Such a view provides a very different interpretation of the actions of the government and often provides a foundation for understanding how governments actually choose their policies. This perspective will also be considered in this book.

The focus on the mixed economy makes the analysis applicable to most developed and developing economies. It also permits the study of how the government behaves and how it should behave. To provide a benchmark from which to judge the outcome of the economy under alternative policies, the command economy with an omniscient planner is often employed. This, of course, is just an analytical abstraction.
1.3 Analyzing Policy

The method of policy analysis in public economics is to build a model of the economy and to find its equilibrium. Policy analysis is undertaken by tracing the effects of a policy through the ways in which it changes the equilibrium of the economy relative to some status quo. Alternative policies are contrasted by comparing the equilibria to which they lead.

In conducting the assessment of policy, it is often helpful to emphasize the distinction between positive and normative analysis. The positive analysis of government investigates topics such as why there is a public sector, where government objectives emerge, and how government policies are chosen. It is also about understanding what effects policies have on the economy. In contrast, normative analysis investigates what the best policies are, and aims to provide a guide to good government. These are not entirely disjoint activities. To proceed with a normative analysis, it is first necessary to conduct the positive analysis: it is not possible to say what is the best policy without knowing the effects of alternative policies upon the economy. It could also be argued that a positive analysis is of no value until used as part of a normative analysis of policy. Normative analysis is conducted under the assumption that the government has a specified set of objectives and its action are chosen in the way that best achieves these. Alternative policies (including the policy of laissez faire or, literally “leave to do”) are compared by using the results of the positive analysis. The optimal policy is that which best meets the government’s objective. Hence the equilibria for different policies are determined and the government’s objective is evaluated for each equilibrium.

In every case restrictions are placed on the set of policies from which the government may choose. These restrictions are usually intended to capture limits on the information that the government has available. The information the government can obtain on the consumers and firms in the economy restricts the degree of sophistication that policy can have. For example, the extent to which taxes can be differentiated among different taxpayers depends on the information the government can acquire about each individual. Administrative and compliance costs are also relevant in generating restrictions on possible policies.

When the government’s objective is taken to be some aggregate level of social welfare in the economy, important questions are raised as to how welfare can be measured. This issue is discussed in some detail in a later chapter, but it can be noted here that the answer involves invoking some degree of comparability between the welfare levels of different individuals. It has been the willingness to proceed on the basis that such comparisons
Part I: Public Economics and Economic Efficiency

can be made that has allowed the development of public economics. While differences of opinion exist on the extent to which these comparisons are valid, it is still scientifically justifiable to investigate what they would imply if they could be made. Furthermore, general principles can be established that apply to any degree of comparability.

1.4 Preview

Part I of the book, which includes this chapter, introduces public economics and provides a review of basic results that are used repeatedly in later chapters. The discussion of the methodology of public economics has shown that a necessary starting point for the development of the theory of policy analysis is an introduction to economic modeling. This represents the content of chapter 2 in which the basic model of a competitive economy is introduced. The chapter describes the agents involved in the economy and characterizes economic equilibrium. An emphasis is placed upon the assumptions on which the analysis is based since much of the subject matter of public economics follows from looking at how the government should respond if these are not satisfied. Having established the basic model, the chapter then investigates the efficiency of the competitive equilibrium. This leads into some fundamental results in welfare economics. The analysis of competitive equilibrium is based on the standard assumption that individuals act rationally in their pursuit of narrow self-interest. This assumption has been increasingly challenged by mounting evidence of observed behavior (so-called anomalies) that is inconsistent with rationality and self-interest. The literature that explores these inconsistencies is typically called behavioral economics. The anomalies raise important questions in many areas of economic policy, so a review of the most relevant parts of behavioral economics is undertaken in chapter 3.

The analysis of government begins in part II. Chapter 4 provides an overview of the public sector. It first charts the historical growth of public sector expenditure over the previous century and then reviews statistics on the present size of the public sector in several of the major developed economies. The division of expenditure and the composition of income are then considered. The level of public debt has recently become an issue of major policy importance, so data are given to demonstrate the increase in debt over time and its allocation across countries. Finally, issues involved in measuring the size of the public sector are addressed. The issues raised by the statistics of chapter 4 are addressed by the discussion of theories of the public sector in chapter 5. Reasons for the existence of the public sector are considered, as are theories that attempt to explain its growth. A positive analysis of how the government may have
Chapter 1: An Introduction to Public Economics

its objectives and actions determined is undertaken. An emphasis is given to arguments for why the observed size of government may be excessive.

The focus of part III is on the consequences of market failure. Chapter 6 introduces public goods into the economy and contrasts the allocation that is achieved when these are privately provided with the optimal allocation. Mechanisms for improving the allocation are considered and methods of preference revelation are also addressed. This is followed by an analysis of clubs and local public goods, which are special cases of public goods in general, in chapter 7. The focus in this chapter returns to an assessment of the success of market provision. The treatment of externalities in chapter 8 relaxes another of the assumptions. It is shown why market failure occurs when externalities are present and reviews alternative policy schemes designed to improve efficiency. Imperfect competition and its consequences for taxation is the subject of chapter 9. The measurement of welfare loss is discussed and emphasis is given to the incidence of taxation. A distinction is also drawn between the effects of specific and ad valorem taxes. Symmetry of information between trading parties is required to sustain efficiency. When it is absent, inefficiency can arise. The implications of informational asymmetries and potential policy responses are considered in chapter 10.

Part IV provides an analysis of the public sector and its decision-making processes. This can be seen as a dose of healthy scepticism before proceeding to the body of normative analysis. An important practical method for making decisions and choosing governments is voting. Chapter 11 analyzes the success of voting as a decision mechanism and the tactical and strategic issues it involves. The main results that emerge are the Median Voter Theorem and the shortcomings of majority voting. These results are then used to develop the analysis of political competition. The consequences of rent-seeking are analyzed in chapter 12. The theory of rent-seeking provides an alternative perspective upon the policy-making process that is highly critical of the actions of government.

Part III focuses on economic efficiency. Part V complements this by considering issues of equity. Chapter 13 analyzes the policy implications of equity considerations and addresses the important restrictions placed on government actions by limited information. Several other fundamental results in welfare economics are also developed including the implications of alternative degrees of interpersonal comparability. Chapter 14 considers the measurement of economic inequality and poverty. The economics of these measures ultimately re-emphasizes the fundamental importance of utility theory. The chapter also stresses the importance of the intergenerational transmission of poverty.
Part VI is concerned with taxation. It analyzes the basic tax instruments and the economics of tax evasion. Chapters 15 and 16 consider commodity taxation and income taxation, which are the two main taxes levied on consumers. In both of these chapters the economic effects of the instruments are considered and rules for setting the taxes optimally are derived. The results illustrate the resolution of the equity/efficiency trade-off in the design of policy and the consequences of the limited information available to the government. In addition to the theoretical analysis the results of application of the methods to data are considered. The numerical results are useful, since the theoretical analysis leads only to characterizations of optimal taxes rather than explicit solutions. These chapters all assume that the taxes that are levied are paid honestly and in full. This empirically doubtful assumption is corrected in chapter 17, which looks at the extent of the hidden economy and analyzes the motives for tax evasion and its consequences. The final chapter of this section, chapter 18, adopts a very abstract approach to investigate the question of whether restricting attention to mixed economies limits what can be achieved by policy intervention. It provides the comforting result that there is no allocation mechanism that can out-perform competitive trade with tax intervention.

Part VII studies public economics when there is more than one decision-making body. Chapter 19 on fiscal federalism addresses why there should be multiple levels of government and discusses the optimal division of responsibilities between different levels. The concept of tax competition is studied in chapter 20. It is shown how tax competition can limit the success of delegating tax-setting powers to independent jurisdictions. The globalization of the world economy and the closer integration of the European economies has emphasized the importance of the international dimension of taxation. A range of issues involved in the design of tax systems in a globalized world are reviewed in chapter 21.

Part VIII concentrates on intertemporal issues in public economics. Chapter 22 describes the overlapping generations economy that is the main analytical tool of this part. The concept of the Golden Rule is introduced for economies with production and capital accumulation, and the potential for economic inefficiency is discussed. Chapter 23 analyzes social security policy and relates this to the potential inefficiency of the competitive equilibrium. Both the motivation for the existence of social security programs and the determination of the level of benefits are addressed. Ricardian equivalence is linked to the existence of gifts and bequests. This part is completed by chapter 24, which considers the effects of taxation and public expenditure on economic growth. Alternative models of economic growth are introduced, and the evidence linking government policy to the level of growth is discussed.
Chapter 1: An Introduction to Public Economics

The final part of the book, part IX, considers two applications of the theory. The discussion of cost–benefit analysis in chapter 25 draws on many of the earlier results to detail a methodology for evaluating when the implementation of a project—which can be interpreted as any change in policy—will increase social welfare. Chapter 26 completes the book with a discussion of the economics of climate policy. The focus in the chapter is on how the tools of public economics can be applied to policy in the presence of pervasive uncertainty.

1.5 Scope

This book is essentially an introduction to the theory of public economics. It presents a unified view of this theory and introduces the most significant results of the analysis. As such, it provides a broad review of what constitutes the present state of public economics.

What will not be found in the book are many details of actual institutions for the collection of taxes or discussion of existing tax codes and other economic policies, although relevant data are used to illuminate arguments. There are several reasons for this. This book is much broader than a text focusing on taxation, and to extend the coverage in this way, something else has to be lost. Primarily, however, the book is about understanding the effects of public policy and how economists think about the analysis of policy. This should give an understanding of the consequences of existing policies, and to benefit from the discussion does not require detailed institutional knowledge.

Furthermore tax codes and tax law are country-specific, and pages spent discussing in detail the rules of one particular country will have little value for those resident elsewhere. In writing the book we wished it to appeal to as broad an audience as possible. From this perspective, the method of reasoning and the analytical results described here have value independent of country-specific detail. Finally there are many alternative texts available that describe tax law and tax codes in detail. These are usually written for accountants and lawyers and have a focus rather distinct from that adopted by economists.

Further Reading

The history of political economy is described in the classic volume:


Two classic references on economic modeling are:
Part I: Public Economics and Economic Efficiency

Two books that show how the tools of public economics can be applied to practical policy analysis are:
The issues involved in comparing individual welfare levels are explored in:

Exercises

1.1 Should an economic model be judged on the basis of its assumptions or its conclusions?
1.2 Explain the economic implications of the imposition of quality standards for drinking water.
1.3 Can economics contribute to an understanding of how government decisions are made?
1.4 In the introduction to “The Economics of Welfare” (1928) Pigou states that:

When a man sets out upon any course of inquiry, the object of his search may be either light or fruit—either knowledge for its own sake or knowledge for the sake of good things to which it leads.

What should be the balance between light and fruit in public economics?
1.5 What should guide the choice of economic policy?
1.6 Are bureaucrats motivated by different factors than entrepreneurs?
1.7 What restricts the policies that a government can choose? Are there any arguments for imposing additional restrictions?
1.8 The UK government scheme Sure Start helps children in disadvantaged areas to have an equal start in life with children from better-off backgrounds, by developing a “one-stop-shop” approach to local family support services. A government minister commented “In the years ahead, children who are healthier and more ready to learn, stronger families and communities, less crime and fewer school-age pregnancies will benefit us all.” Examine the efficiency and equity components of the minister’s comments.
1.9 “Physics is a simpler discipline than economics. This is because the objects of its study are bound by physical laws.” Do you agree?
1.10 If individual welfare levels cannot be compared, how can it be possible to make social judgments?
1.11 “Poverty should be reduced to lessen the extent of malnutrition and raise economic growth.”
Distinguish the positive and normative components of this statement.

1.12 “It is economically efficient to maintain a pool of unemployed labor.” Is this claim based on positive or normative reasoning?

1.13 “High-income earners should pay a high rate of tax because their labor supply is inelastic and the revenue raised can be used to assist those on low incomes.” Distinguish between the positive and normative components of this statement.

1.14 Bob Wailer was a talented singer from a very early age. With no training or practice he built a successful career as the enigmatic frontman of a punk revivalist band. From the day she could hold a bow Emma Libro was trained in classical violin. After many years of study and practice she finally obtained lead violin for the Exeter Philharmonic and a lucrative recording contract. Assume that Bob and Emma have the same income. Is it equitable for them to pay the same amount of tax?

1.15 Consider two methods of dividing a cake between two people. Method 1 is to throw some of the cake away, and share what is left equally. Method 2 is to give one person 75 percent of the cake and the other 25 percent. Which method do you prefer, and why?

1.16 A cake has to be apportioned between two people. One is well-nourished, and the other is not. If the well-nourished person receives a share $x$, $0 \leq x \leq 1$, a share $y = [1 - x]^2$ is left for the other person (some is lost when the cake is divided). Plot the possible shares that the two people can have. What allocation of shares would you choose? How would your answer change if $y = [1 - x]^4$?

1.17 A cake is to be divided between two people. Unfortunately, when the cake is sliced some is lost. Specifically, if the sizes of the two portions are denoted $x$ and $y$, then the size of the slices satisfy the relation $x^{1/2} + y^{1/2} = 1$. Graph the possible divisions of the cake and discuss how the balance between efficiency and equity will affect the choice of division.

1.18 Can an economic model be acceptable if it assumes that consumers solve computationally complex maximization problems? Does your answer imply that Tiger Woods can derive the law of motion for a golf ball?

1.19 To analyze the effect of a subsidy to rice production, would you employ a partial equilibrium or a general equilibrium model?

1.20 If the European Union considered replacing the income tax with an increase in VAT, would you model this using partial equilibrium?

1.21 What proportion of the world's economies (by number, population, and wealth) can be described as “mixed”?

1.22 What problems may arise in setting economic policy if consumers know the economic model?

1.23 Should firms maximize profit?

1.24 To what extent is it possible to view the government as having a single objective?

1.25 Are you happier than your neighbor? How many times happier or less happy?
1.26 Assume that consumers are randomly allocated to either earn income $M_\ell$ or income $M_h$, where $M_h > M_\ell$. The probability of being allocated to $M_\ell$ is $\pi$. Prior to being allocated to an income level, consumers wish to maximize their expected income level. If it is possible to redistribute income costlessly, show that prior to allocation to income levels, no consumer would object to a transfer scheme. Now assume that there is a cost $\Delta$ for each consumer of income $M_h$ from whom income is taken. Find the maximum value of $\pi$ for which there is still unanimous agreement that transfers should take place.
2 Equilibrium and Efficiency

2.1 Introduction

The link between competition and efficiency can be traced back, at least, to Adam Smith’s eighteenth-century description of the working of the invisible hand. Smith’s description of individually motivated decisions being coordinated to produce a socially efficient outcome is a powerful one that has found resonance in policy circles ever since. The expression of the efficiency argument in the language of formal economics, and the deeper understanding that comes with it, is a more recent innovation.

The focus of this chapter is to review what is meant by competition and to describe equilibrium in a competitive economy. The model of competition combines independent decision-making of consumers and firms into a complete model of the economy. Equilibrium is shown to be achieved in the economy by prices adjusting to equate demand and supply. Most important, the chapter employs the competitive model to demonstrate the efficiency theorems.

Surprisingly, equilibrium prices can always be found that simultaneously equate demand and supply for all goods. What is even more remarkable is that the equilibrium so obtained also has properties of efficiency. Why this is remarkable is that individual households and firms pursue their independent objectives with no concern other than their own welfare. Even so, the final state that emerges achieves efficiency solely through the coordinating role played by prices.

2.2 Economic Models

Prior to starting the analysis, it is worth reflecting on why economists employ models to make predictions about the effects of economic policies. Models are used essentially because of problems of conducting experiments on economic systems and because the system is too large and complex to analyze in its entirety. Moreover formal modeling ensures that arguments are logically consistent with all the underlying assumptions exposed.

The models used, while inevitably being simplifications of the real economy, are designed to capture the essential aspects of the problem under study. Although many different models will be studied in this book, there are important common features that apply to all. Most models in public economics specify the objectives of the individual
agents (e.g., firms and consumers) in the economy, and the constraints they face, and then aggregate individual decisions to arrive at market demand and supply. The equilibrium of the economy is next determined, and in a policy analysis the effects of government choice variables on this are calculated. This is done with various degrees of detail. Sometimes only a single market is studied—this is the case of *partial equilibrium* analysis. At other times *general equilibrium* analysis is used with many markets analyzed simultaneously. Similarly the number of firms and consumers varies from one or two to very many.

An essential consideration in the choice of the level of detail for a model is that its equilibrium must demonstrate a dependence on policy that gives insight into the functioning of the actual economy. On the one hand, if the model is too highly specified, it may not be capable of capturing important forms of response. On the other hand, if it is too general, it may not be able to provide any clear prediction. The theory described in this book will show how this trade-off can be successfully resolved. Achieving a successful compromise between these competing objectives is the “art” of economic modeling.

### 2.3 Competitive Economies

The essential feature of competition is that the consumers and firms in the economy do not consider their actions to have any effect on prices. Consequently, in making decisions, they treat the prices they observe in the market place as fixed (or *parametric*). This assumption can be justified when all consumers and firms are truly negligible in size relative to the market. In such a case the quantity traded by an individual consumer or firm is not sufficient to change the market price. But the assumption that the agents view prices as parametric can also be imposed as a modeling tool, even in an economy with a single consumer and a single firm.

This defining characteristic of competition places a focus on the role of prices, as is maintained throughout the chapter. Prices measure values and are the signals that guide the decisions of firms and consumers. It was the exploration of what determined the relative values of different goods and services that led to the formulation of the competitive model. The adjustment of prices equates supply and demand to ensure that equilibrium is achieved. The role of prices in coordinating the decisions of independent economic agents is also crucial for the attainment of economic efficiency.

The secondary feature of the economies in this chapter is that all agents have access to the *same* information, or in formal terminology, that information is *symmetric*. This
does not imply that there cannot be uncertainty, but only that when there is uncertainty, all agents are equally uninformed. Put differently, no agent is permitted to have an informational advantage. For example, by this assumption, the future profit levels of firms are allowed to be uncertain and shares in the firms to be traded on the basis of individual assessments of future profits. What the assumption does not allow is for the directors of the firms to be better informed than other shareholders about future prospects and to trade profitably on the basis of this information advantage.

Two forms of the competitive model are introduced in this chapter. The first form is an exchange economy in which there is no production. Initial stocks of goods are held by consumers and economic activity occurs through the trade of these stocks to mutual advantage. The second form of competitive economy introduces production. This is undertaken by firms with given production technologies who use inputs to produce outputs and distribute their profits as dividends to consumers.

2.4 The Exchange Economy

The exchange economy models the simplest form of economic activity: the trade of commodities between two parties in order to obtain mutual advantage. Despite the simplicity of this model, it is a surprisingly instructive tool for obtaining fundamental insights about taxation and tax policy. This will become evident as we proceed. This section presents a description of a two-consumer, two-good exchange economy. The restriction on the number of goods and consumers does not alter any of the conclusions that will be derived; they will all extend to larger numbers. What restricting the numbers does is allow the economy to be displayed and analyzed in a simple diagram.

Each of the two consumers has an initial stock, or endowment, of the economy’s two goods. The endowments can be interpreted literally as stocks of goods, or less literally as human capital, and are the quantities that are available for trade. Given the absence of production, these quantities remain constant. The consumers exchange quantities of the two commodities in order to achieve consumption plans that are preferred to their initial endowments. The rate at which one commodity can be exchanged for the other is given by the market prices. Both consumers believe that their behavior cannot affect these prices. This is the fundamental assumption of competitive price-taking behavior. More will be said about the validity and interpretation of this in section 2.6.

A consumer is described by their endowments and their preferences. The endowment of consumer $h$ is denoted by $\omega^h = (\omega^h_1, \omega^h_2)$, where $\omega^h_i \geq 0$ is $h$’s initial stock of good $i$.  

When prices are \( p_1 \) and \( p_2 \), a consumption plan for consumer \( h \), \( x^h = (x^h_1, x^h_2) \), is affordable if it satisfies the budget constraint

\[
p_1 x^h_1 + p_2 x^h_2 = p_1 \omega^h_1 + p_2 \omega^h_2.
\] (2.1)

The preferences of each consumer are described by their utility function. This function should be seen as a representation of the consumer’s indifference curves and does not imply any comparability of utility levels between consumers—the issue of comparability is taken up in chapter 13. The utility function for consumer \( h \) is denoted by

\[
U^h = U^h(x^h_1, x^h_2).
\] (2.2)

It is assumed that the consumers enjoy the goods (so the marginal utility of consumption is positive for both goods) and that the indifference curves have the standard convex shape.

This economy can be pictured in a simple diagram that allows the role of prices in achieving equilibrium to be explored. The diagram is constructed by noting that the total consumption of the two consumers must equal the available stock of the goods, where the stock is determined by the endowments. Any pair of consumption plans that satisfies this requirement is called a feasible plan for the economy. A plan for the economy is feasible if the consumption levels can be met from the endowments, so that

\[
x^i_1 + x^i_2 = \omega^i_1 + \omega^i_2, \quad i = 1, 2.
\] (2.3)

The consumption plans satisfying (2.3) can be represented as points in a rectangle with sides of length \( \omega^i_1 + \omega^i_2 \) and \( \omega^i_1 + \omega^i_2 \). In this rectangle the southwest corner can be treated as the zero consumption point for consumer 1 and the northeast corner as the zero consumption point for consumer 2. The consumption of good 1 for consumer 1 is then measured horizontally from the southwest corner and for consumer 2 horizontally from the northeast corner. Measurements for good 2 are made vertically.

The diagram constructed in this way is called an Edgeworth box and a typical box is shown in figure 2.1. It should be noted that the method of construction results in the endowment point, marked \( \omega \), being the initial endowment point for both consumers.

The Edgeworth box is completed by adding the preferences and budget constraints of the consumers. The indifference curves of consumer 1 are drawn relative to the southwest corner and those of consumer 2 relative to the northeast corner. From (2.1) it can be seen that the budget constraint for both consumers must pass through the endowment point, since consumers can always afford their endowment. The endowment point is
common to both consumers, so a single budget line through the endowment point with gradient \(-\frac{p_1}{p_2}\) captures the market opportunities of the two consumers. Thus, viewed from the southwest, it is the budget line of consumer 1, and viewed from the northeast, the budget line of consumer 2. Given the budget line determined by the prices \(p_1\) and \(p_2\), the utility-maximizing choices for the two consumers are characterized by the standard tangency condition between the highest attainable indifference curve and the budget line. This is illustrated in figure 2.2, where \(x^1\) denotes the choice of consumer 1 and \(x^2\) that of 2.

At an equilibrium of the economy, supply is equal to demand. This is assumed to be achieved via the adjustment of prices. The prices at which supply is equal to demand are called equilibrium prices. How such prices are arrived at will be discussed later. For the present the focus will be placed on the nature of equilibrium and its properties. The consumer choices shown in figure 2.2 do not constitute an equilibrium for the economy. This can be seen by summing the demands and comparing these to the level of the endowments. Doing this shows that the demand for good 1 exceeds the endowment but the demand for good 2 falls short. To achieve an equilibrium position, the relative prices of the goods must change. An increase in the relative price of good 1 raises the absolute value of the gradient \(-\frac{p_1}{p_2}\) of the budget line, making the budget line steeper. It becomes flatter if the relative price of good 1 falls. At all prices the budget
line continues to pass through the endowment point so that a change in relative prices sees the line pivot about the endowment point.

The effect of a relative price change on the budget constraint is shown in figure 2.3. In the figure the price of good 1 has increased relative to the price of good 2. This causes the budget constraint to pivot upward around the endowment point. As a consequence of this change the consumers will now select consumption plans on this new budget constraint.

The dependence of the consumption levels on prices is summarized in the consumers’ demand functions. Taking the prices as given, the consumers choose their consumption plans to reach the highest attainable utility level subject to their budget constraints. The level of demand for good $i$ from consumer $h$ is $x^h_i = x^h_i(p_1, p_2)$. Using the demand functions, we see that demand is equal to supply for the economy when the prices are such that

$$x^1_1(p_1, p_2) + x^2_1(p_1, p_2) = \omega^1_1 + \omega^2_1, \quad i = 1, 2. \tag{2.4}$$

A study of the Edgeworth box shows that an equilibrium is achieved when the prices lead to a budget line on which the indifference curves of the consumers have a point of common tangency. Such an equilibrium is shown in figure 2.4.

Having illustrated an equilibrium, we raise the question of whether an equilibrium is guaranteed to exist. As it happens, under reasonable assumptions, it will always do so. More important for public economics is the issue of whether the equilibrium has any desirable features from a welfare perspective. This is discussed in depth in section 2.6 where the Edgeworth box is put to substantial use.
Two further points now need to be made that are important for understanding the functioning of the model. These concern the number of prices that can be determined and the number of independent equilibrium equations. In the equilibrium conditions (2.4) there are two equations to be satisfied by the two equilibrium prices. It is now argued that the model can determine only the ratio of prices and not the actual prices. Accepting this, it would seem that there is one price ratio attempting to solve two equations. If this were the case, a solution would be unlikely, and we would be in the position of having a model that generally did not have an equilibrium. This situation is resolved by noting that the relationship between the two equilibrium conditions ensures that there is only one independent equation. The single price ratio then has to be solved by a single equation, making it possible for there to be always a solution.
The first point is developed by observing that the budget constraint always passes through the endowment point and its gradient is determined by the price ratio. The consequence of this is that only the value of $p_1$ relative to $p_2$ matters in determining demands and supplies rather than the absolute values. The economic explanation for this fact is that consumers are only concerned with the real purchasing power embodied in their endowment, and not with the level of prices. Since their nominal income is equal to the value of the endowment, any change in the level of prices raises nominal income just as much as it raises the cost of purchases. This leaves real incomes unchanged.

The fact that only relative prices matter is also reflected in the demand functions. If $x^h(p_1, p_2)$ is the level of demand at prices $p_1$ and $p_2$, then it must be the case that $x^h(p_1, p_2) = x^h(\lambda p_1, \lambda p_2)$ for $\lambda > 0$. A demand function having this property is said to be homogeneous of degree 0. In terms of what can be learned from the model, the homogeneity shows that only relative prices can be determined at equilibrium and not the level of prices. So, given a set of equilibrium prices, any scaling up or down of these prices will also be equilibrium prices because the change will not alter the level of demand. This is as it should be, since all that matters for the consumers is the rate at which they can exchange one commodity for another, and this is measured by the relative prices. This can be seen in the Edgeworth box. The budget constraint always goes through the endowment point so only its gradient can change, and this is determined by the relative prices.

In order to analyze the model, the indeterminacy of the level of prices needs to be removed. This is achieved by adopting a price normalization, which is simply a method of fixing a scale for prices. There are numerous ways to do this. The simplest way is to select a commodity as the numéraire, which means that its price is fixed at one, and measure all other prices relative to this. The numéraire chosen in this way can be thought of as the unit of account for the economy. This is the role usually played by money but, formally, there is no money in this economy.

The second point is to demonstrate the dependence between the two equilibrium equations. It can be seen that at the disequilibrium position shown in figure 2.2, the demand for good 1 exceeds its supply whereas the supply of good 2 exceeds demand. Considering other budget lines and indifference curves in the Edgeworth box will show that whenever there is an excess of demand for one good, there is a corresponding deficit of demand for the other. There is actually a very precise relationship between the excess and the deficit that can be captured in the following way: The level of excess demand for good $i$ is the difference between demand and supply and is defined by $Z_i = x^1_i + x^2_i - \omega^1_i - \omega^2_i$. By this definition, the value of excess demand can be
21 Chapter 2: Equilibrium and Efficiency

calculated as

\[ p_1 Z_1 + p_2 Z_2 = \sum_{i=1}^{2} p_i \left[ x_1^i + x_2^i - \omega_1^i - \omega_2^i \right] \]

\[ = \sum_{h=1}^{2} \left[ p_1 x_1^h + p_2 x_2^h - p_1 \omega_1^h - p_2 \omega_2^h \right] \]

\[ = 0, \quad (2.5) \]

where the second equality is a consequence of the budget constraints in (2.1). The relationship in (2.5) is known as Walras’s law and states that the value of excess demand is zero. This must hold for any set of prices, so it provides a connection between the extent of disequilibrium and prices. In essence, Walras’s law is simply an aggregate budget constraint for the economy. Since all consumers are equating their expenditure to their income, so must the economy as a whole.

Walras’s law implies that the equilibrium equations are interdependent. Since \( p_1 Z_1 + p_2 Z_2 = 0 \), if \( Z_1 = 0 \), then \( Z_2 = 0 \) (and vice versa). That is, if demand is equal to supply for good 1, then demand must also equal supply for good 2. Equilibrium in one market necessarily implies equilibrium in the other. This observation allows the construction of a useful diagram to illustrate equilibrium. Choose good 1 as the numéraire (so \( p_1 = 1 \)) and plot the excess demand for good 2 as a function of \( p_2 \). The equilibrium for the economy is then found where the graph of excess demand crosses the horizontal axis. At this point excess demand for good 2 is zero, so by Walras’s law, it must also be zero for good 1. An excess demand function is illustrated in figure 2.5 for an economy that has three equilibria. This excess demand function demonstrates

\[ Z_2(1, p_2) \]

\[ p_2 \]

Figure 2.5
Equilibrium and excess demand
why at least one equilibrium will exist. As \( p_2 \) falls toward zero, demand will exceed supply (good 2 becomes increasingly attractive to purchase), making excess demand positive. Conversely, as the price of good 2 rises, it will become increasingly attractive to sell, resulting in a negative value of excess demand for high values of \( p_2 \). Since excess demand is positive for small values of \( p_2 \) and negative for high values, there must be at least one point in between where it is zero.

Finally, it should be noted that the arguments made above can be extended to include additional consumers and additional goods. Income, in terms of an endowment of many goods, and expenditure, defined in the same way, must remain equal for each consumer. The demand functions that result from the maximization of utility are homogeneous of degree zero in prices. Walras’s law continues to hold, so the value of excess demand remains zero. The number of price ratios and the number of independent equilibrium conditions are always one less than the number of goods.

### 2.5 Production and Exchange

The addition of production to the exchange economy provides a complete model of economic activity. Such an economy allows a wealth of detail to be included. Some goods can be present as initial endowments (e.g., labor); others can be consumption goods produced from the initial endowments, while some goods, intermediates, can be produced by one productive process and used as inputs into another. The fully developed model of competition is called the *Arrow–Debreu economy* in honor of its original constructors.

An economy with production consists of consumers (or households) and producers (or firms). The firms use inputs to produce outputs with the intention of maximizing their profits. Each firm has available a production technology that describes the ways in which it can use inputs to produce outputs. The consumers have preferences and initial endowments as they did in the exchange economy, but they now also hold shares in the firms. The firms’ profits are distributed as dividends in proportion to the shareholdings. The consumers receive income from the sale of their initial endowments (e.g., their labor time) and from the dividend payments.

Each firm is characterized by its production set, which summarizes the production technology it has available. A production technology can be thought of as a complete list of ways that the firm can turn inputs into outputs. In other words, it catalogs all the production methods of which the firm has knowledge. For firm \( j \) operating in an economy with two goods a typical production set, denoted \( Y^j \), is illustrated in figure 2.6.
This figure employs the standard convention of measuring inputs as negative numbers and outputs as positive. The reason for adopting this convention is that the use of a unit of a good as an input represents a subtraction from the stock of that good available for consumption.

Consider the firm shown in figure 2.6 choosing the production plan $y_1^j = -2, y_2^j = 3$. When faced with prices $p_1 = 2, p_2 = 2$, the firm’s profit is

$$\pi^j = p_1y_1^j + p_2y_2^j = 2 \times (-2) + 2 \times 3 = 2.$$  (2.6)

The positive part of this sum can be given the interpretation of sales revenue, and the negative part that of production costs. This is equivalent to writing profit as the difference between revenue and cost. Written in this way, (2.6) gives a simple expression of the relation between prices and production choices.

The process of profit maximization is illustrated in figure 2.7. Under the competitive assumption the firm takes the prices $p_1$ and $p_2$ as given. These prices are used to construct isoprofit curves, which show all production plans that give a specific level of profit. For example, all the production plans on the isoprofit curve labeled $\pi = 0$ will lead to a profit level of 0. Production plans on higher isoprofit curves lead to progressively larger profits, and those on lower curves to negative profits. Since doing nothing (which means choosing $y_1^j = y_2^j = 0$) earns zero profit, the $\pi = 0$ isoprofit curve always passes through the origin.

The profit-maximizing firm will choose a production plan that places it upon the highest attainable isoprofit curve. What restricts the choice is the technology that is available as described by the production set. In figure 2.7 the production plan that maximizes
profit is shown by $y^*$, which is located at a point of tangency between the highest attainable isoprofit curve and the production set. There is no other technologically feasible plan that can attain higher profit.

It should be noted how the isoprofit curves are determined by the prices. The geometry in fact is that the isoprofit curves are at right angles to the price vector. The angle of the price vector is determined by the price ratio, $\frac{p_2}{p_1}$, so a change in relative prices will alter the gradient of the isoprofit curves. The figure can be used to predict the effect of relative price changes. For instance, if $p_1$ increases relative to $p_2$, which can be interpreted as the price of the input (good 1) rising in comparison to the price of the output (good 2), then the price vector becomes flatter. This makes the isoprofit curves steeper, so the optimal choice must move round the boundary of the production set toward the origin. The use of input and the production of output both fall.

Now consider an economy with $n$ goods. The price of good $i$ is denoted $p_i$. Production is carried out by $m$ firms. Each firm uses inputs to produce outputs and maximizes profits given the market prices. Demand comes from the $H$ consumers in the economy. They aim to maximize their utility. The total supply of each good is the sum of the production of it by firms and the initial endowment of it held by the consumers.

Each firm chooses a production plan $y^j = (y^j_1, \ldots, y^j_n)$. This production plan is chosen to maximize profits subject to the constraint that the chosen plan must be in the production set. From this maximization can be determined firm $j$’s supply function for good $i$ as $y^j_i = y^j_i(p)$, where $p = (p_1, \ldots, p_n)$. The level of profit is $\pi^j = \sum_{i=1}^n p_i y^j_i(p) = \pi^j(p)$, which also depends on prices.
Aggregate supply from the production sector of the economy is obtained from the supply decisions of the individual firms by summing across the firms. This gives the aggregate supply of good $i$ as

$$Y_i(p) = \sum_{j=1}^{m} y_i^j(p). \quad (2.7)$$

Since some goods must be inputs, and others outputs, aggregate supply can be positive (the total activity of the firms adds to the stock of the good) or negative (the total activity of the firms subtracts from the stock).

Each consumer has an initial endowment of commodities and also a set of shareholdings in firms. The latter assumption makes this a private ownership economy in which the means of production are ultimately owned by individuals. In the present version of the model, these shareholdings are exogenously given and remain fixed. A more developed version would introduce a stock market and allow them to be traded. For consumer $h$ the initial endowment is denoted $\omega^h$ and the shareholding in firm $j$ is $\theta^h_j$. The firms must be fully owned by the consumers, so $\sum_{h=1}^{H} \theta^h_j = 1$. That is, the shares in the firms must sum to one. Consumer $h$ chooses a consumption plan $x^h$ to maximize utility subject to the budget constraint

$$\sum_{i=1}^{n} p_i x^h_i = \sum_{i=1}^{n} p_i \omega^h_i + \sum_{j=1}^{m} \theta^h_j \pi^j. \quad (2.8)$$

This budget constraint requires that the value of expenditure be not more than the value of the endowment plus income received as dividends from firms. Since firms always have the option of going out of business (and hence earning zero profit), the dividend income must be nonnegative. The profit level of each firm is dependent on prices. A change in prices therefore affects a consumer’s budget constraint through a change in the value of their endowment and through a change in dividend income. The maximization of utility by the consumer results in demand for good $i$ from consumer $h$ of the form $x^h_i = x^h_i(p)$. The level of aggregate demand is found by summing the individual demands of the consumers to give

$$X_i(p) = \sum_{h=1}^{H} x^h_i(p). \quad (2.9)$$
The same notion of equilibrium that was used for the exchange economy can be applied in this economy with production. That is, equilibrium occurs when supply is equal to demand. The distinction between the two is that supply, which was fixed in the exchange economy, is now variable and dependent on the production decisions of firms. Although this adds a further dimension to the question of the existence of equilibrium, the basic argument why such an equilibrium always exists is essentially the same as that for the exchange economy.

As already noted, the equilibrium of the economy occurs when demand is equal to supply or, equivalently, when excess demand is zero. Excess demand for good $i$, $Z_i(p)$, can be defined by

$$Z_i(p) = X_i(p) - Y_i(p) - \sum_{h=1}^{H} \omega_h^i.$$  \hspace{1cm} (2.10)

Here excess demand is the difference between demand and the sum of initial endowment and firms’ supply. The equilibrium occurs when $Z_i(p) = 0$ for all of the goods $i = 1, \ldots, n$. There are standard theorems that prove such an equilibrium must exist under fairly weak conditions.

The properties established for the exchange economy also apply to this economy with production. Demand is determined only by relative prices (so it is homogeneous of degree zero). Supply is also determined by relative prices. Together, these imply that excess demand is homogeneous of degree zero. To determine the equilibrium prices that equate supply to demand, a normalization must again be used. Typically one of the goods will be chosen as numéraire, and its price set to one. Equilibrium prices are then those that equate excess demand to zero.

### 2.6 Efficiency of Competition

Economics is often defined as the study of scarcity. This viewpoint is reflected in the concern with the efficient use of resources that runs throughout the core of the subject. Efficiency would seem to be a simple concept to characterize: if more cannot be achieved, then the outcome is efficient. This is certainly the case when an individual decision maker is considered. The individual will employ their resources to maximize utility subject to the constraints they face. When utility is maximized, the efficient outcome has been achieved.

Problems arise when there is more than one decision maker. To be unambiguous about efficiency, it is necessary to resolve the potentially competing needs of different
Chapter 2: Equilibrium and Efficiency

decision makers. This requires efficiency to be defined with respect to a set of aggregate preferences. Methods of progressing from individual to aggregate preferences will be discussed in chapters 11 and 13. The conclusions obtained there are that the determination of aggregate preferences is not a simple task. There are two routes we can use to navigate around this difficulty. The first is to look at a single-consumer economy so that there is no conflict between competing preferences. But with more than one consumer some creativity has to be used to describe efficiency. The second route is met in section 2.6.2 where the concept of Pareto-efficiency is introduced. The trouble with such creativity is that it leaves the definition of efficiency open to debate. We will postpone further discussion of this until chapter 13.

Before we proceed further, some definitions are needed. A first-best outcome is achieved when only the production technology and the limited endowments restrict the choice of the decision maker. The first-best is essentially what would be chosen by an omniscient planner with complete command over resources. A second-best outcome arises whenever constraints other than technology and resources are placed on what the planner can do. Such constraints could be limits on income redistribution, an inability to remove monopoly power, or a lack of information.

2.6.1 Single Consumer

With a single consumer there is no doubt as to what is good and bad from a social perspective: the single individual’s preferences can be taken as the social preferences. To do otherwise would be to deny the validity of the consumer’s judgments. Hence, if the individual prefers one outcome to another, then so must society. The unambiguous nature of preferences provides significant simplification of the discussion of efficiency in the single-consumer economy. In this case the “best” outcome must be first-best because no constraints on policy choices have been invoked nor is there an issue of income distribution to consider.

If there is a single firm and a single consumer, the economy with production can be illustrated in a helpful diagram. This is constructed by superimposing the profit-maximization diagram for the firm over the choice diagram for the consumer. Such a model is often called the Robinson Crusoe economy. The interpretation is that Robinson acts as a firm carrying out production and as a consumer of the product of the firm. It is then possible to think of Robinson as a social planner who can coordinate the activities of the firm and producer. It is also possible (though in this case less compelling!) to think of Robinson as having a split personality and acting as a profit-maximizing firm on one side of the market and as a utility-maximizing consumer on the other. In the
latter interpretation the two sides of Robinson’s personality are reconciled through the prices on the competitive markets. The important fact is that these two interpretations lead to exactly the same levels of production and consumption.

The budget constraint of the consumer needs to include the dividend received from the firm. With two goods, the budget constraint is

\[ p_1[x_1 - \omega_1] + p_2[x_2 - \omega_2] = \pi, \]  

(2.11)
or

\[ p_1\tilde{x}_1 + p_2\tilde{x}_2 = \pi, \]  

(2.12)

where \( \tilde{x}_i \), the change from the endowment point, is the net consumption of good \( i \). This is illustrated in figure 2.8 with good 2 chosen as the numéraire. The budget constraint (2.12) is always at a right angle to the price vector and is displaced above the origin by the value of profit. Utility maximization occurs where the highest indifference curve is reached given the budget constraint. This results in net consumption plan \( \tilde{x}^* \).

The equilibrium for the economy is shown in figure 2.9, which superimposes figure 2.7 onto 2.8. At the equilibrium the net consumption plan from the consumer must match the supply from the firm. The feature that makes this diagram work is the fact that the consumer receives the entire profit of the firm, so the budget constraint and the isoprofit curve are one and the same. The height above the origin of both is the level

![Figure 2.8](Utility maximization)
of profit earned by the firm and received by the consumer. Equilibrium can only arise when the point on the economy’s production set that equates to profit maximization is the same as that of utility maximization. This is point $\tilde{x}^* = y^*$ in figure 2.9.

It should be noted that the equilibrium is on the boundary of the production set so that it is efficient: it is not possible for a better outcome to be found in which more is produced with the same level of input. This captures the efficiency of production at the competitive equilibrium, about which much more is said soon. The equilibrium is also the first-best outcome for the single-consumer economy, since it achieves the highest indifference curve possible subject to the restriction that it is feasible under the technology. This is illustrated in figure 2.9 where $\tilde{x}^*$ is the net level of consumption relative to the endowment point in the first-best and at the competitive equilibrium.

A simple characterization of this first-best allocation can be given by using the fact that it is at a tangency point between two curves. The gradient of the indifference curve is equal to the ratio of the marginal utilities of the two goods and is called the *marginal rate of substitution*. This measures the rate at which good 1 can be traded for good 2 while maintaining constant utility. The marginal rate of substitution is given by $MRS_{1,2} = \frac{U_1}{U_2}$, with subscripts used to denote the marginal utilities of the two goods. Similarly the gradient of the production possibility set is termed the *marginal rate of transformation* and denoted $MRT_{1,2}$. The $MRT_{1,2}$ measures the rate at which good 1 has to be given up to allow an increase in production of good 2. At the tangency point the two gradients are equal, so
The reason why this equality characterizes the first-best equilibrium can be explained as follows: The \( MRS \) captures the marginal value of good 1 to the consumer relative to the marginal value of good 2, while the \( MRT \) measures the marginal cost of good 1 relative to the marginal cost of good 2. The first-best is achieved when the marginal value is equal to the marginal cost.

The market achieves efficiency through the coordinating role of prices. The consumer maximizes utility subject to their budget constraint. The optimal choice occurs when the budget constraint is tangential to the highest attainable indifference curve. The condition describing this is that ratio of marginal utilities is equal to the ratio of prices. Expressed in terms of the \( MRS \), this is

\[
MRS_{1,2} = \frac{p_1}{p_2}. \tag{2.14}
\]

Similarly profit maximization by the firm occurs when the production possibility set is tangential to the highest isoprofit curve. Using the \( MRT \), we write the profit-maximization condition as

\[
MRT_{1,2} = \frac{p_1}{p_2}. \tag{2.15}
\]

Combining these conditions, we find that the competitive equilibrium satisfies

\[
MRS_{1,2} = \frac{p_1}{p_2} = MRT_{1,2}. \tag{2.16}
\]

The condition in (2.16) demonstrates that the competitive equilibrium satisfies the same condition as the first-best and reveals the essential role of prices. By the competitive assumption, both the consumer and the producer are guided in their decisions by the same price ratio. Each optimizes relative to the price ratio; hence their decisions are mutually efficient.

There is one special case that is worth noting before moving on. When the firm has constant returns to scale, the efficient production frontier is a straight line through the origin. The only equilibrium can be when the firm makes zero profits. If profit was positive at some output level, then the constant returns to scale allows the firm to double profit by doubling output. Since this argument can be repeated, there is no limit to the profit that the firm can make. Hence we have the claim that equilibrium profit must be zero. Now the isoprofit curve at zero profit is also a straight line through the origin.
Chapter 2: Equilibrium and Efficiency

The zero-profit equilibrium can only arise when this is coincident with the efficient production frontier. At this equilibrium the price vector is at right angles to both the isoprofit curve and the production frontier. This is illustrated in figure 2.10.

There are two further implications of constant returns. First, the equilibrium price ratio is determined by the zero-profit condition alone and is independent of demand. Second, the profit income of the consumer is zero, so the consumer’s budget constraint also passes through the origin. As this is determined by the same prices as the isoprofit curve, the budget constraint must be coincident with the production frontier.

In this single-consumer context the equilibrium reached by the market simply cannot be bettered. Such a strong statement cannot be made when more consumers are introduced because issues of distribution among consumers then arise. However, what will remain is the finding that the competitive market ensures that firms produce at an efficient point on the frontier of the production set and that the chosen production plan is what is demanded at the equilibrium prices by the consumers. The key to this coordination are the prices that provide the signals guiding choices.

2.6.2 Pareto-Efficiency

When there is more than one consumer, the simple analysis of the Robinson Crusoe economy does not apply. Since consumers can have differing views about the success of an allocation, there is no single, simple measure of efficiency. The essence of the
problem is that of judging among allocations with different distributional properties. What is needed is some process that can take account of the potentially diverse views of the consumers and separate efficiency from distribution.

To achieve this, economists employ the concept of \textit{Pareto-efficiency}. The philosophy behind this concept is to interpret efficiency as meaning that there must be no unexploited economic gains. Testing the efficiency of an allocation then involves checking whether there are any such gains available. More specifically, Pareto-efficiency judges an allocation by considering whether it is possible to undertake a reallocation of resources that can benefit at least one consumer without harming any other. If it is possible to do so, then there will exist unexploited gains. When no improving reallocation can be found, then the initial position is deemed to be Pareto-efficient. An allocation that satisfies this test can be viewed as having achieved an efficient distribution of resources. For the present chapter this concept will be used uncritically. The interpretations and limitations of this form of efficiency will be discussed in chapter 13.

To provide a precise statement of Pareto-efficiency that applies in a competitive economy, it is first necessary to extend the idea of feasible allocations of resources that was used in (2.3) to define the Edgeworth box. When production is included, an allocation of consumption is feasible if it can be produced given the economy’s initial endowments and production technology. Given the initial endowment, \( \omega \), the consumption allocation \( x \) is feasible if there is production plan \( y \) such that
\[
x = y + \omega.
\] (2.17)
Pareto-efficiency is then tested using the feasible allocations. A precise definition follows.

\textbf{Definition 2.1} A feasible consumption allocation \( \hat{x} \) is Pareto-efficient if there does not exist an alternative feasible allocation \( \tilde{x} \) such that:

\begin{enumerate}
\item allocation \( \tilde{x} \) gives all consumers at least as much utility as \( \hat{x} \), and
\item allocation \( \tilde{x} \) gives at least one consumer more utility than \( \hat{x} \).
\end{enumerate}

These two conditions can be summarized as saying that allocation \( \hat{x} \) is Pareto efficient if there is no alternative allocation (a move from \( \hat{x} \) to \( \tilde{x} \)) that can make someone better off without making anyone worse off. It is this idea of being able to make someone better off without making someone else worse off that represents the unexploited economic gains in an inefficient position.
It should be noted even at this stage how Pareto-efficiency is defined by the negative property of being unable to find anything better than the allocation. This is somewhat different from a definition of efficiency that looks for some positive property of the allocation. Pareto-efficiency also sidesteps issues of distribution rather than confronting them. This is why it works with many consumers. More will be said about this in chapter 13 when the construction of social welfare indicators is discussed.

### 2.6.3 Efficiency in an Exchange Economy

The welfare properties of the economy, which are commonly known as the *Two Theorems of Welfare Economics*, are the basis for claims concerning the desirability of the competitive outcome. In brief, the First Theorem states that a competitive equilibrium is Pareto-efficient and the Second Theorem that any Pareto-efficient allocation can be decentralized as a competitive equilibrium. Taken together, they have significant implications for policy and, at face value, seem to make a compelling case for the encouragement of competition.

The Two Theorems are easily demonstrated for a two-consumer exchange economy by using the Edgeworth box diagram. The first step is to isolate the Pareto-efficient allocations. Consider figure 2.11 and the allocation at point $a$. To show that $a$ is not a Pareto-efficient allocation, it is necessary to find an alternative allocation that gives at least one of the consumers a higher utility level and neither consumer a lower level. In this case, moving to the allocation at point $b$ raises the utility of both consumers when compared to point $a$—we say in such a case that $b$ is *Pareto-preferred* to $a$. This

---

**Figure 2.11**

Pareto-efficiency
establishes that \( a \) is not Pareto-efficient. Although \( b \) improves on \( a \), it is not Pareto-efficient either: the allocation at \( c \) provides higher utility for both consumers than \( b \).

The allocation at \( c \) is Pareto-efficient. Beginning at \( c \), any change in the allocation must lower the utility of at least one of the consumers. The special property of point \( c \) is that it lies at a point of tangency between the indifference curves of the two consumers. As it is a point of tangency, moving away from it must lead to a lower indifference curve for one of the consumers if not both. Since the indifference curves are tangential, their gradients are equal, so

\[
MRS_{1,2}^1 = MRS_{1,2}^2. \tag{2.18}
\]

This equality ensures that the rate at which consumer 1 will want to exchange good 1 for good 2 is equal to the rate at which consumer 2 will want to exchange the two goods. It is this equality of the marginal valuations of the two consumers at the tangency point that results in there being no further unexploited gains and so makes \( c \) Pareto efficient.

The Pareto-efficient allocation at \( c \) is not unique. There are in fact many points of tangency between the two consumers’ indifference curves. A second Pareto-efficient allocation is at point \( d \) in figure 2.11. Taken together, all the Pareto-efficient allocations form a locus in the Edgeworth box that is called the contract curve. This is illustrated in figure 2.12. With this construction it is now possible to demonstrate the First Theorem.

A competitive equilibrium is given by a price line through the initial endowment point, \( \omega \), that is tangential to both indifference curves at the same point. The common point of tangency results in consumer choices that lead to the equilibrium levels of demand. Such an equilibrium is indicated by point \( e \) in figure 2.12. As the equilibrium
Chapter 2: Equilibrium and Efficiency

is a point of tangency of indifference curves, it must also be Pareto-efficient. For the Edgeworth box, this completes the demonstration that a competitive equilibrium is Pareto-efficient.

The alternative way of seeing this result is to recall that each consumer maximizes utility at the point where their budget constraint is tangential to the highest indifference curve. Using the \( MRS \), we can write this condition for consumer \( h \) as \( MRS^h_{1,2} = \frac{p_1}{p_2} \). The competitive assumption is that both consumers react to the same set of prices, so it follows that

\[
MRS^1_{1,2} = \frac{p_1}{p_2} = MRS^2_{1,2}. \tag{2.19}
\]

Comparing this condition with (2.18) provides an alternative demonstration that the competitive equilibrium is Pareto-efficient. It also shows again the role of prices in coordinating the independent decisions of different economic agents to ensure efficiency.

This discussion can be summarized in the precise statement of the theorem.

**Theorem 2.2** (First Theorem of Welfare Economics) The allocation of commodities at a competitive equilibrium is Pareto-efficient.

This theorem can be formally proved by assuming that the competitive equilibrium is not Pareto-efficient and by then deriving a contradiction. Assuming that the competitive equilibrium is not Pareto-efficient implies that there is a feasible alternative that is at least as good for all consumers and strictly better for at least one. Now take the consumer who is made strictly better off. Why did that consumer not choose the alternative consumption plan at the competitive equilibrium? The answer has to be because it was more expensive than the choice at the competitive equilibrium and not affordable with that consumer’s budget. Similarly for all other consumers the new allocation has to be at least as expensive as their choice at the competitive equilibrium. (If it were cheaper, they could afford an even better consumption plan that made them strictly better off than at the competitive equilibrium.) Summing across the consumers, the alternative allocation has to be strictly more expensive than the competitive allocation. But the value of consumption at the competitive equilibrium must equal the value of the endowment. Therefore the new allocation must have greater value than the endowment, which implies it cannot be feasible. This contradiction establishes that the competitive equilibrium must be Pareto-efficient.
The theorem demonstrates that the competitive equilibrium is Pareto-efficient, but it is not the only Pareto-efficient allocation. Referring back to figure 2.12, we have that any point on the contract curve is also Pareto-efficient because all are defined by a tangency between indifference curves. The only special feature of \( e \) is that it is the allocation reached through competitive trading from the initial endowment point \( \omega \). If \( \omega \) were different, then another Pareto-efficient allocation would be achieved. There is in fact an infinity of Pareto-efficient allocations. Observing these points motivates the Second Theorem of Welfare Economics.

The Second Theorem is concerned with whether any chosen Pareto-efficient allocation can be made into a competitive equilibrium by choosing a suitable location for the initial endowment. Expressed differently, can a competitive economy be constructed that has a selected Pareto-efficient allocation as its competitive equilibrium? In the Edgeworth box this involves being able to choose any point on the contract curve and turning it into a competitive equilibrium.

From the Edgeworth box diagram it can be seen that this is possible in the exchange economy if the households’ indifference curves are convex. The common tangent to the indifference curves at the Pareto-efficient allocation provides the budget constraint that each consumer must face if they are to afford the chosen point. The convexity ensures that given this budget line, the Pareto-efficient point will also be the optimal choice of the consumers. The construction is completed by choosing a point on this budget line as the initial endowment point. This process of constructing a competitive economy to obtain a selected Pareto-efficient allocation is termed decentralization.

This process is illustrated in figure 2.13 where the Pareto-efficient allocation \( e' \) is made a competitive equilibrium by selecting \( \omega' \) as the endowment point. Starting from \( \omega' \),

\[\begin{array}{c}
\text{Figure 2.13} \\
The Second Theorem
\end{array}\]
trading by consumers will take the economy to its equilibrium allocation $e'$. This is the Pareto-efficient allocation that was intended to be reached. Note that if the endowments of the households are initially given by $\omega$ and the equilibrium at $e'$ is to be decentralized, it is necessary to redistribute the initial endowments of the consumers in order to begin from $\omega'$. The construction described above can be given a formal statement as the Second Theorem of Welfare Economics.

Theorem 2.3 (Second Theorem of Welfare Economics) With convex preferences, any Pareto-efficient allocation can be made a competitive equilibrium.

The statement of the Second Theorem provides a conclusion but does not describe the mechanism involved in the decentralization. The important step in decentralizing a chosen Pareto-efficient allocation is placing the economy at the correct starting point. For now it is sufficient to observe that behind the Second Theorem lies a process of redistribution of initial wealth. How this can be achieved is discussed later. Furthermore the Second Theorem determines a set of prices that make the chosen allocation an equilibrium. These prices may well be very different from those that would have been obtained in the absence of the wealth redistribution.

2.6.4 Extension to Production

The extension of the Two Theorems to an economy with production is straightforward. The major effect of production is to make supply variable: it is now the sum of the initial endowment plus the net outputs of the firms. In addition a consumer’s income includes the profit derived from their shareholdings in firms. Section 2.6.1 has already demonstrated efficiency for the Robinson Crusoe economy that included production. It was shown that the competitive equilibrium achieved the highest attainable indifference curve given the production possibilities of the economy. Since the single consumer cannot be made better off by any change, the equilibrium is Pareto-efficient and the First Theorem applies. The Second Theorem is of limited interest with a single consumer because there is only one Pareto-efficient allocation, and this is attained by the competitive economy.

When there is more than one consumer, the proof of the First Theorem follows the same lines as for the exchange economy. Given the equilibrium prices, each consumer is maximizing utility, so each consumer’s marginal rate of substitution is equated to the
same price ratio. This is true for all consumers and all goods, yielding

$$MRS_{i,j}^h = \frac{p_i}{p_j} = MRS_{i,j}^{h'}$$  \hspace{1cm} (2.20)$$

for any pair of consumers $h$ and $h'$ and any pair of goods $i$ and $j$. This is termed efficiency in consumption. In an economy with production this condition alone is not sufficient to guarantee efficiency; it is also necessary to consider production. The profit-maximization decision of each firm ensures that it equates its marginal rate of transformation between any two goods to the ratio of prices. For any two firms $m$ and $m'$,

$$MRT_{i,j}^m = \frac{p_i}{p_j} = MRT_{i,j}^{m'}$$  \hspace{1cm} (2.21)$$
a condition that characterizes efficiency in production. The price ratio also coordinates consumers and firms, giving the top-level condition

$$MRS_{i,j}^h = MRT_{i,j}^m$$  \hspace{1cm} (2.22)$$

for any consumer and any firm for all pairs of goods. As for the Robinson Crusoe economy, the interpretation of this condition is that it equates the relative marginal values to the relative marginal costs. Since (2.20) through (2.22) are the conditions required for efficiency, this shows that the First Theorem extends to the economy with production.

The formal proof of this claim mirrors that for the exchange economy, except for the fact that the value of production must also be taken into account. Given this fact, the basis of the argument remains that since the consumers chose the competitive equilibrium quantities, anything that is preferred must be more expensive and hence can be shown not to be feasible.

The extension of the Second Theorem to include production is illustrated in figure 2.14. The set $W$ describes the feasible output plans for the economy, with each point $w$ in $W$ being the sum of a production plan and the initial endowment; hence $w = y + \omega$. Set $Z$ describes the quantities of the two goods that would allow a Pareto improvement (a re-allocation that makes neither consumer worse off and makes one strictly better off) over the allocation $\hat{x}^1$ to consumer 1 and $\hat{x}^2$ to consumer 2. If $W$ and $Z$ are convex, which occurs when firms’ production sets and preferences are convex, then a common tangent to $W$ and $Z$ can be found. This makes $\hat{x}$ an equilibrium. Individual income allocations, the sum of the value of endowment plus profit income, can be placed anywhere on the budget lines tangent to the indifference curves at the individual allocations.
\( \hat{x}_1 \) and \( \hat{x}_2 \) provided that they sum to the total income of the economy. This decentralizes the consumption allocation \( \hat{x}_1, \hat{x}_2 \).

Before proceeding further, it is worth emphasizing that the proof of the Second Theorem requires more assumptions than the proof of the First, so there may be situations in which the First Theorem is applicable but the Second is not. The Second Theorem requires that a common tangent be found, which relies on preferences and production sets being convex. A competitive equilibrium can exist with some nonconvexity in the production sets of the individual firms or the preferences of the consumers, so the First Theorem will apply but the Second Theorem will not apply.

### 2.7 Lump-Sum Taxation

The discussion of the Second Theorem noted that it does not describe the mechanism through which the decentralization is achieved. It is instead implicit in the statement of the theorem that the consumers are given sufficient income to purchase the consumption plans forming the Pareto-efficient allocation. Any practical value of the Second Theorem depends on the government being able to allocate the required income levels. The way in which the theorem sees this as being done is by making what are called *lump-sum transfers* between consumers.
A transfer is defined as lump sum if no change in a consumer’s behavior can affect the size of the transfer. For example, a consumer choosing to work less hard or reducing the consumption of a commodity must not be able to affect the size of the transfer. This differentiates a lump-sum transfer from other taxes, such as income or commodity taxes, for which changes in behavior do affect the value of the tax payment. Lump-sum transfers have a very special role in the theoretical analysis of public economics because, as we will show, they are the idealized redistributive instrument.

The lump-sum transfers envisaged by the Second Theorem involve quantities of endowments and shares being transferred among consumers to ensure the necessary income levels. Some consumers would gain from the transfers; others would lose. Although the value of the transfer cannot be changed, lump-sum transfers do affect consumers’ behavior because their incomes are either reduced or increased by the transfers—the transfers have an income effect but do not lead to a substitution effect between commodities. Without recourse to such transfers, the decentralization of the selected allocation would not be possible.

The illustration of the Second Theorem in an exchange economy in figure 2.15 makes clear the role and nature of lump-sum transfers. The initial endowment point is denoted $\omega$, and this is the starting point for the economy. It we assume that the Pareto-efficient allocation at point $e$ is to be decentralized, then the income levels have to be modified to achieve the new budget constraint. At the initial point the income level of $h$ is $\hat{p}\omega^h$ when evaluated at the equilibrium prices $\hat{p}$. The value of the transfer to consumer $h$ that is necessary to achieve the new budget constraint is $M^h - \hat{p}\omega^h = \hat{p}\hat{x}^h - \hat{p}\omega^h$. One
way of ensuring this is to transfer a quantity $\bar{x}_1$ of good 1 from consumer 1 to consumer 2. But any transfer of commodities with the same value would work equally well.

There is a problem, though, if we attempt to interpret the model this literally. For most people, income is earned almost entirely from the sale of labor so that their endowment is simply lifetime labor supply. This makes it impossible to transfer the endowment since one person’s labor cannot be given to another. Responding to such difficulties leads to the reformulation of lump-sum transfers in terms of lump-sum taxes. Suppose that the two consumers both sell their entire endowments at prices $\hat{p}$. This generates incomes $\hat{p}\omega_1$ and $\hat{p}\omega_2$ for the two consumers. Now make consumer 1 pay a tax of amount $T^1 = \hat{p} \bar{x}_1$ and give this tax revenue to consumer 2. Consumer 2 therefore pays a negative tax (or, in simpler terms, receives a subsidy) of $T^2 = -\hat{p} \bar{x}_1 = -T^1$. The pair of taxes $(T^1, T^2)$ moves the budget constraint in exactly the same way as the lump-sum transfer of endowment. The pair of taxes and the transfer of endowment are therefore economically equivalent and have the same effect on the economy. The taxes are also lump sum because they are determined without reference to either consumers’ behavior and their values cannot be affected by any change in behavior.

Lump-sum taxes have a central role in public economics due to their success in achieving distributional objectives. It should be clear from the discussion above that the economy’s total endowment is not reduced by the application of the lump-sum taxes. This point applies to lump-sum taxes in general. As households cannot affect the level of the tax by changing their behavior, lump-sum taxes do not lead to any distortions in choice. There are also no resources lost due to the imposition of lump-sum taxes, so redistribution is achieved with no efficiency cost. In short, if they can be employed in the manner described they are the perfect taxes.

### 2.8 Discussion of Assumptions

The description of the competitive economy introduced a number of assumptions concerning the economic environment and how trade was conducted. These are important since they bear directly on the efficiency properties of competition. The interpretation and limitation of these assumptions is now discussed. This should provide a better context for evaluating the practical relevance of the efficiency theorems.

The most fundamental assumption was that of competitive behavior. This is the assumption that both consumers and firms view prices as fixed when they make their decisions. The natural interpretation of this assumption is that the individual economic agents are small relative to the total economy. When they are small, they naturally
have no consequence. This assumption rules out any kind of market power such as monopolistic firms or trade unions in labor markets.

Competitive behavior leads to the problem of who actually sets prices in the economy. In the exchange model it is possible for equilibrium prices to be achieved via a process of barter and negotiation between the trading parties. However, barter cannot be a credible explanation of price determination in an advanced economic environment. One theoretical route out of this difficulty is to assume the existence of a fictitious “Walrasian auctioneer” who literally calls out prices until equilibrium is achieved. Only at this point trade is allowed to take place. Obviously this does not provide a credible explanation of reality. Although there are other theoretical explanations of price-setting, none is entirely consistent with the competitive assumption. How to integrate the two remains an unsolved puzzle.

The second assumption was symmetry of information. In a complex world there are many situations in which this does not apply. For instance, some qualities of a product, such as reliability (I do not know when my computer will next crash, but I expect it will be soon), are not immediately observable but are discovered only through experience. When it comes to re-sale, this causes an asymmetry of information between the existing owner and potential purchasers. The same can be true in labor markets where workers may know more about their attitudes toward work and potential productivity than a prospective employer. An asymmetry of information provides a poor basis for trade because the caution of those transacting prevents the full gains from trade being realized.

When any of the assumptions underlying the competitive economy fail to be met, and as a consequence efficiency is not achieved, we say that there is market failure. Situations of market failure are of interest to public economics because they provide a potential role for government policy to enhance efficiency. A large section of this book is in fact devoted to a detailed analysis of the sources of market failure and the scope for policy response.

As a final observation, notice that the focus has been on positions of equilibrium. Several explanations can be given for this emphasis. Historically economists viewed the economy as self-correcting so that, if it were ever away from equilibrium, forces exist that move it back toward equilibrium. In the long run, equilibrium would then always be attained. Although such adjustment can be justified in simple single-market contexts, both the practical experience of sustained high levels of unemployment and the theoretical study of the stability of the price adjustment process have shown that the self-adjusting equilibrium view is not generally justified. The present justifications for focusing on equilibrium are more pragmatic. The analysis of a model must begin somewhere, and the equilibrium has much merit as a starting point.
In addition, even if the final focus is on disequilibrium, there is much to be gained from comparing the properties of points of disequilibrium with those of the equilibrium. Finally, no positions other than those of equilibrium have any obvious claim to prominence.

2.9 Summary

This chapter described competitive economies and demonstrated the Two Theorems of Welfare Economics. To do this, it was necessary to introduce the concept of Pareto-efficiency. While Pareto-efficiency was simply accepted in this chapter, it will be considered very critically in chapter 13. The Two Theorems characterize the efficiency properties of the competitive economy and show how a selected Pareto-efficient allocation can be decentralized. It was also shown how prices are central to the achievement of efficiency through their role in coordinating the choices of individual agents. The role of lump-sum transfers or taxes in supporting the Second Theorem was highlighted. These transfers constitute the ideal tax system because they cause no distortions in choice and have no resource costs.

The subject matter of this chapter has very strong implications that are investigated fully in later chapters. An understanding of the welfare theorems, and of their limitations, is fundamental to appreciating many of the developments of public economics. Since claims about the efficiency of competition feature routinely in economic debate, it is important to subject it to the most careful scrutiny.

Further Reading

The two fundamental texts on the competitive economy are:

A textbook treatment can be found in:

The competitive economy has frequently been used as a practical tool for policy analysis. A survey of applications is in:

A historical survey of the development of the model is given in:

Some questions concerning the foundations of the model are addressed in:


The classic proofs of the Two Theorems are in:


A formal analysis of lump-sum taxation can be found in:


An extensive textbook treatment of Pareto-efficiency is:


---

**Exercises**

2.1  Distinguish between partial equilibrium analysis and general equilibrium analysis. Briefly describe a model of each kind.

2.2  Keynesian models in macroeconomics are identified by the assumption of a fixed price for output. Are such models partial or general equilibrium?

2.3  You are requested to construct a model to predict the effect on the economy of the discovery of new oil reserves. How would you model the discovery? Discuss the number of goods that should be included in the model.

2.4  Let a consumer have preferences described by the utility function

\[ U = \log(x_1) + \log(x_2), \]

and an endowment of 2 units of good 1 and 2 units of good 2.

a. Construct and sketch the consumer’s budget constraint. Show what happens when the price of good 1 increases.

b. By maximizing utility, determine the consumer’s demands.

c. What effect does increasing the endowment of good 1 have on the demand for good 2? Explain your finding.

2.5  How would you model an endowment of labor?

2.6  Let two consumers have preferences described by the utility function

\[ U^h = \log(x^h_1) + \log(x^h_2), \quad h = 1, 2, \]
and the endowments described below:

<table>
<thead>
<tr>
<th></th>
<th>Good 1</th>
<th>Good 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consumer 1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Consumer 2</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

a. Calculate the consumers’ demand functions.
b. Selecting good 2 as the numéraire, find the equilibrium price of good 1. Hence find the equilibrium levels of consumption.
c. Show that the consumers’ indifference curves are tangential at the equilibrium.

2.7 Consider an economy with two goods and two consumers with preferences

\[ U^h = \min \left( x^h_1, x^h_2 \right), \quad h = 1, 2. \]

Assume that the endowments are as follows:

<table>
<thead>
<tr>
<th></th>
<th>Good 1</th>
<th>Good 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consumer 1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Consumer 2</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

a. Draw the Edgeworth box for the economy.
b. Display the equilibrium in the Edgeworth box.
c. What is the effect on the equilibrium price of good 2 relative to good 1 of an increase in each consumer’s endowment of good 1 by 1 unit?

2.8 Consumer 1 obtains no pleasure from good 1, and consumer 2 obtains no pleasure from good 2. At the initial endowment point both consumers have endowments of both goods.

a. Draw the preferences of the consumers in an Edgeworth box.
b. By determining the trades that improve both consumers’ utilities, find the equilibrium of the economy.
c. Display the equilibrium budget constraint.

2.9 Demonstrate that the demands obtained in exercise 2.4 are homogeneous of degree zero in prices. Show that doubling prices does not affect the graph of the budget constraint.

2.10 It has been argued that equilibrium generally exists on the basis that there must be a point where excess demand for good 2 is zero if excess demand is positive as the price of good 2 tends to zero and negative as it tends to infinity.

a. Select good 1 as the numéraire and show that these properties hold when preferences are given by the utility function
and the consumer’s endowment of both goods is positive.

b. Show that they do not hold if the consumer has no endowment of good 2.

c. Consider the implications of the answer to part b for proving the existence of equilibrium.

2.11 Consider an economy with 2 consumers, A and B, and 2 goods, 1 and 2. The utility function of A is

\[ U^A = \gamma \log(x^A_1) + (1 - \gamma) \log(x^A_2) \]

where \( x^A_i \) is consumption of good i by A. A has endowments \( \omega^A = (\omega^A_1, \omega^A_2) = (2, 1) \). For B,

\[ U^B = \gamma \log(x^B_1) + (1 - \gamma) \log(x^B_2) \]

and \( \omega^B = (3, 2) \).

a. Use the budget constraint of A to substitute for \( x^A_2 \) in \( U^A \), and by maximizing over \( x^A_1 \), calculate the demands of A. Repeat for B.

b. Choosing good 2 as the numéraire, graph the excess demand for good 1 as a function of \( p_1 \).

c. Calculate the competitive equilibrium allocation by equating the demand for good 1 to the supply and then substituting for \( M^A \) and \( M^B \). Verify that this is the point where excess demand is zero.

d. Show how the equilibrium price of good 2 is affected by a change in \( \gamma \) and in \( \omega^A_1 \). Explain the results.

2.12 A firm has a production technology that permits it to turn 1 unit of good 1 into 2 units of good 2. If the price of good 1 is 1, at what price for good 2 will the firm just break even? Graph the firm’s profit as a function of the price of good 2.

2.13 Consider the production process described by

\[ F(x_1, x_2, x_3) \equiv x_1 - (x_2)^\beta (-x_3)^{1 - \beta} = 0, \]

where \( x_1 \geq 0 \) is the output, and \( x_2 \leq 0, x_3 \leq 0 \) are the inputs.

a. If good 1 is the numéraire, what prices of goods 2 and 3 are consistent with zero profit?

b. Discuss the observation that, in a model with constant returns to scale, equilibrium prices are determined by technology and not preferences.

2.14 How can the existence of fixed costs be incorporated into the production set diagram? After paying its fixed costs, a firm has constant returns to scale. Can it earn zero profits in a competitive economy?

2.15 Consider a two good exchange economy. Let the excess demand function for good 1 be given by

\[ Z_1(p_1, p_2) = 3.426 + \delta - 10p_1 + 8(p_1)^2 - (p_1)^3 + \log(p_2). \]

a. Select good 2 as the numéraire and plot the excess demand function for \( \delta = 0 \). Show that the economy has two equilibria.
b. How many equilibria are there if $\delta = 0.1$? If $\delta = -0.1$? What can you conclude about the possibility of observing an economy with an even number of equilibria?

c. Compute the excess demand function for good 2 using Walras’s law. Can these excess demand functions result from utility maximization?

2.16 Consider a consumer with utility function $U = \log(x) + \log(1 - \ell)$ and a firm with production function $x = (\ell)^{1/2}$, where $x$ denotes output of a consumption good and $\ell$ denotes labor supply. Assume that the consumer receives the profit from the firm as a dividend and that both the firm and consumer act competitively. Choosing labor as the numéraire, find the maximized utility of the consumer and the maximized profit of the firm as functions of the price, $p$, of output. What value of $p$ maximizes utility? What is special about this value of $p$?

2.17 Consider an economy with 2 goods, $H$ consumers, and $m$ firms. Each consumer, $h$, has an endowment of 2 units of good 1 and none of good 2, with the preferences described by $U^h = x^h_1 x^h_2$, and a share $\theta^h_j = \frac{1}{H}$ in firms $j = 1, \ldots, m$. Each firm has a technology characterized by the production function $y^j_2 = \left[-y^j_1\right]^{1/2}$.

a. Calculate a firm’s profit-maximizing choices, a consumer’s demands, and the competitive equilibrium of the economy.

b. What happens to $\frac{p_2}{p_1}$ as (i) $m$ increases; (ii) $H$ increases? Why?

c. Suppose that each consumer’s endowment of good 1 increases to $2 + 2\delta$. Explain the change in relative prices.

d. What is the effect of changing
   i. the distribution of endowments among consumers;
   ii. the consumers’ preferences to $U^h = \alpha \log(x^h_1) + \beta \log(x^h_2)$?

2.18 Reproduce the diagram for the Robinson Crusoe economy for a firm that has constant returns to scale. Under what conditions will it be efficient for the firm not to produce? What is the consumption level of the consumer in such a case? Provide an interpretation of this possibility.

2.19 After the payment of costs, fishing boat captains distribute the surplus to the owner and crew. Typically the owner receives 50 percent, the captain 30 percent, and the remaining 20 percent is distributed to crew according to status. (See The Perfect Storm: A True Story of Man against the Sea by Sebastian Junger Norton 1997.) Is this distribution Pareto-efficient? Is it equitable?

2.20 A box of chocolates is to be shared by two children. The box contains ten milk chocolates and ten dark chocolates. Neither child likes dark chocolates. Describe the Pareto-efficient allocations.

2.21 As economists are experts in resource allocation, you are invited by two friends to resolve a dispute about the shared use of a car. By applying Pareto-efficiency, how are you able to advise them? Are they impressed with your advice?

2.22 Two consumers have utility functions

$$U^h = \ln(x^h_1) + \ln(x^h_2).$$

a. Calculate the marginal rate of substitution between good 1 and good 2 in terms of consumption levels.
b. By equating the marginal rates of substitution for the two consumers, characterize a Pareto-efficient allocation.

c. Using the solution to part b, construct the contract curve for an economy with 2 units of good 1 and 3 units of good 2.

2.23 A university has a fixed sum of money to allocate in bonus payments between two professors. Each professor appreciates receiving the bonus but resents that the other professor receives a bonus as well. Let the preferences of the two professors be given by

\[ U_G = b_G - cb_J, \]

and

\[ U_J = b_J - cb_G. \]

a. Describe the Pareto-efficient allocations when \( c = 0 \).

b. Is there a Pareto-efficient allocation when \( 0 < c < 1 \)?

c. What happens if \( c \geq 1 \)?

2.24 A consumer views two goods as perfect substitutes.

a. Sketch the indifference curves of the consumer.

b. If an economy is composed of two consumers with these preferences, demonstrate that any allocation is Pareto-efficient.

c. If an economy has one consumer who views its two goods as perfect substitutes and a second that considers each unit of good 1 to be worth 2 units of good 2, find the Pareto-efficient allocations.

2.25 Consider an economy in which preferences are given by

\[ U^1 = x_1^1 + x_2^1, \]

and

\[ U^2 = \min \{x_1^2, x_2^2\}. \]

Given the endowments \( \omega^1 = (1, 2) \) and \( \omega^2 = (3, 1) \), construct the set of Pareto-efficient allocations and the contract curve. Which allocations are also competitive equilibria?

2.26 Take the economy in the exercise above, but change the preferences of consumer 2 to

\[ U^2 = \max \{x_1^2, x_2^2\}. \]

Is there a Pareto-efficient allocation?

2.27 Consider an economy with two consumers, A and B, and two goods, 1 and 2. Using \( x_i^h \) to denote the consumption of good i by consumer h, assume that both consumers have the utility function \( U^h = \min \{x_1^h, x_2^h\} \).

a. By drawing an Edgeworth box, display the Pareto-efficient allocations if the economy has an endowment of 1 unit of each good.
Chapter 2: Equilibrium and Efficiency

2.28 Consider the economy in exercise 2.11.
   a. Calculate the endowments required to make the equal-utility allocation a competitive equilibrium.
   b. Discuss the transfer of endowment necessary to support this equilibrium.

2.29 Provide an example of a Pareto-efficient allocation that cannot be decentralized.

2.30 Let an economy have a total endowment of two units of the two available goods. If the two consumers have preferences

\[ U_h = \alpha \log(x_{1h}) + [1 - \alpha] \log(x_{2h}), \]

find the ratio of equilibrium prices at the allocation where \( U^1 = U^2 \). Hence find the value of the lump-sum transfer that is needed to decentralize the allocation if the initial endowments are \( \left( \frac{1}{2}, \frac{3}{4} \right) \) and \( \left( \frac{3}{2}, \frac{5}{4} \right) \).

2.31 Are the following statements true or false? Explain why in each case.
   a. If one consumer gains from a trade, the other consumer involved in the trade must lose.
   b. The gains from trade are based on comparative advantage, not absolute advantage.
   c. The person who can produces the good with less input has an absolute advantage in producing this good.
   d. The person who has the smaller opportunity cost of producing the good has a comparative advantage in producing this good.
   e. The competitive equilibrium is the only allocation where the gains from trade are exhausted.
3 Behavioral Economics

3.1 Introduction

Behavioral economics involves both the introduction and the development within economic theory of insights about behavior drawn from different domains of psychology. It has been widely recognized for a long time (since at least the early contributions of Allais and Ellsberg) that some of the central assumptions of standard economic analysis may reflect an unrealistic representation of human behavior. To be sure, the standard approach can produce wrong predictions on some occasions. Empirical and experimental research has cataloged a wide range of anomalies: observed choices that do not match the predictions of standard analysis. Behavioral economics is an attempt to improve the predictive power of economics by building in more realistic descriptions of individual behavior. It is useful for behavioral analysis to view people as departing from standard economic behavior in three distinct ways: limited rationality, limited will power, and limited self-interest.

In this chapter we consider public intervention motivated by the fact that people can make invalid choices or mistakes. This form of public intervention is aimed at protecting people against the consequences of their own decisions. There is a long history of opposition to any form of paternalistic public intervention where the state claims to know better what is “truly” good for people than they know for themselves. The nature of paternalism is frequently illustrated by the analogy to the interference of parents in their children’s choices and the limitation of those choices. We may readily accept that parents can limit the freedom of choices of their children and make some choices on their behalf, but the same argument is rarely accepted with adults and state intervention. Libertarians are the most opposed to that kind of public intervention by asking “are you too incompetent to know what’s best for yourself?” Traditional (public) economics assumes that people know what is best for themselves and they can get themselves to act according to their own best interest. In that context there is no need for public intervention beyond redistribution and the correction of market failure.

To provide the motivation for the approach we take, it is necessary to clarify some basic issues. We can distinguish three forms of public policies: paternalism, welfarism, and behavioralism. Paternalism refers to policies aimed at benefiting individuals who cannot be relied upon to pursue self-interest. It is most appropriate for children and others with behavioral disorders deemed unable to make rational choice. Welfarism, as
Part I: Public Economics and Economic Efficiency

We have noted many times already, denotes government policy aimed at resolving the trade-off between efficiency and equity in order to maximize a social objective function. Improving efficiency is uncontroversial, since it implies a Pareto improvement and as such will be unanimously supported. In contrast, attaining aims of equity implies redistribution, so there will be some consumers that lose from this (but others that gain). Welfarism represents policies intended to benefit individuals when self-interest (broadly defined) cannot be relied upon because of the presence of market failures or the need for redistribution. Behavioralism represents policies intended to benefit individuals when self-interest (again broadly defined) cannot be relied upon because of the presence of internal conflicts (internalities). The central idea is that if people make systematic mistakes or biased choices they may regret later on, public policy can manipulate those mistakes and biases that hurt people (1) to help them to protect themselves and at the same time (2) to respect their autonomy of choice.

There are two types of mistakes people can make and that can motivate public intervention. First, people do not know what is best for themselves through a mere lack of knowledge of the needed information. Suppose that people cannot easily distinguish healthy and nonhealthy food. Then a policy enforcing effective “healthy” food labeling is beneficial to individuals (if not to the food industry). Second, people do know what is best for themselves, but they cannot get themselves to act in the correct way because of a lack of self-discipline. For example, I may plan to stop smoking tomorrow or plan to start jogging tomorrow, but when tomorrow arrives I cannot get myself to do it. Public policy aimed at helping me to do it will be beneficial.

The purpose of this chapter is to describe how limited rationality, which is the possibility that individuals can make biased choices or invalid choices, opens up the scope for a wide range of policies intended to benefit individuals who cannot be relied upon because of internal conflicts. The policies must help people help themselves, and at the same time the policies must respect their autonomy of choice. In that sense the policies are not paternalistic. We will start by describing several examples of invalid choices. We will then describe policies that can effectively help people help themselves to do what is best for them. We will subsequently discuss the welfare evaluation of those policies, and more important, what welfare criteria to adopt when revealed preferences through individual choices cannot be relied upon to make welfare judgments, since people can make invalid choices that do not represent their “true” preferences. The final section of the chapter discusses another important deviation from the standard model, namely when people are not selfish optimizing agents but display concern over the material consumption of other agents. This section reflects the general idea that people do care about giving and receiving fair treatment in a wide set of circumstances.
3.2 Behavioral Individuals

In chapter 2 we described an economy in which people had only economic motives and were also fully rational. Those restrictions in the standard approach to public economics are useful as they impose a research structure and discipline in all analysis of the role and benefit of public intervention. But the immediate question is: How does the economy behave if people have noneconomic motives and rational responses (e.g., the other-regarding-preference model)? Economic motives but irrational responses (e.g., the behavioral model)? Noneconomic motives and irrational responses (psychological model)? It is obvious that the answers to important questions about how the economy behaves, and what public interventions are needed when it misbehaves, vary significantly from one model of the economy to another. Each model corresponds to a different vision of behavior in the economy. The role of the government is to set the conditions under which irrational people can be harnessed creatively (without harming the rational people) to serve the greater good. Some forms of irrational behavior may give the government new opportunities to step in with policy interventions.

Behavioral economics is interested in three forms of imperfections in decision-making due to imperfect rationality, imperfect will power, and imperfect self-interest. We now explore these imperfections and their consequences.

3.2.1 Simple Example: How Much to Save?

The United States Consumer Expenditure Survey suggests that households spend more out of dividend income than out of income from capital gains (Baker, Nagel, and Wurgler 2006). Rational economic theory implies that the same proportion of the extra money (when the extra is the same) should be spent in both cases. To address this issue, Shefrin and Thaler (1988) conducted an experiment that asked subjects how much they would spend out of an unexpected extra money gain of $2,400 in three possible situations (or framings).

- **First framing** The extra money is a bonus at work paid out at constant rate of $200 a month over a year (so that the total amount is $2,400). The median saving out of the extra money was $100 monthly for a total saving of $1,200.

- **Second framing** The extra money is a lump-sum payment of $2,400 this month. The median response was that $400 would be spent immediately, and $35 a month for the rest of the year. That makes total spending of $785, and so a total saving of $1,615.
Part I: Public Economics and Economic Efficiency

- **Third framing** The extra money is invested in an interest-bearing account for five years and the subjects receive at the end of five years $2,400 plus interest (so that the present value of the payment is $2,400). The median response is that none of the (future) capital would be spent during the first year.

Rational people should spend the same portion of the extra money for all the frames. Shefrin and Thaler interpret the deviation from the rational choice outcome as suggesting that people place different kinds of income in different *mental accounts*: the current income account (in frame 1), the asset income account (in frame 2), and the future income account (in frame 3), and that they spend differently from each of these mental accounts (spending much less out of the future income account). This is one of the many facets of the so-called framing effect to which we return later. A framing effect is usually said to occur when equivalent descriptions of a decision problem lead to systematically different decisions. Framing effects are commonly taken as evidence for incoherence in human decision-making, and for the empirical inapplicability of the rational choice models used by economists.

### 3.2.2 Present Bias

Present bias is an explanation of the self-control problem. Strotz (1957) analyzed how preferences that change over time would affect the saving decision, but present bias is a much more general problem where people face a choice that is liable to change their own preferences in the future. This is comparable to Odysseus when passing the Sirens and their enchantingly seductive voices. Odysseus faced an interesting decision on how to resist the temptation (either sailing close to the island to hear the Sirens but then running the risk of directing his ship onto the rocks, or sailing past the island and missing the chance to hear the Sirens).

A similar situation is the decision to initially engage in an addictive activity (gaming, smoking, drinking, etc.). There is first a decision to start the addictive activity (option *Smoke*) or not (option *No*). An initial decision to start the addictive activity, like smoking, leads to a future decision of whether to continue the addictive activity or to stop (option *Quit*). Because of the addictive nature of the activity, future preferences are such that option *Smoke* is preferred to option *Quit*: you are hooked to this addictive activity in some sense. But initial preferences (before addiction kicks in) are such that people may want to give it a try, so that option *Quit* seems better than option *No*, which in turn is better than option *Smoke*, which is the awful option. So the initial preferences at time $t$ are

\[ Quit \succ_t No \succ_t Smoke, \]

but future preferences at time $t+1$ when the addiction has set in are

$$\text{Smoke} >_{t+1} \text{Quit}.$$  

Facing this preference reversal, a naive individual ignores the issue entirely, and plans to reach what is initially the best option Quit by starting the addictive activity. But then in the next period, when the addiction has set in, the awful option Smoke is the final result. On the contrary, a sophisticated individual foresees that Smoke will be preferred to Quit in the future so that Quit is not really a feasible option. The choices boil down to either Smoke or No, and the sophisticated agent chooses No, avoiding any risk of addiction. In this self-control problem there is, in effect, one individual self today and an entirely different individual self in the future. The today self and the future self have different preferences. The naive individual who ignores this duality seems obviously irrational. The sophisticated individual is rational, in a sense, but achieves rationality only by realizing the truth that there is a preference reversal between today self and future self.

The self-control problem is a dynamic consistency problem in which the naive individual fails to take into account the future preference reversal at the initial stage. There is an inconsistency between what he would like to do tomorrow, and what he would do in effect tomorrow. Odysseus solved the present-bias problem in a different way: by precommitment. He created an extra option by being tied to the mast of the ship that allowed him to hear the Sirens without being able to direct his ship onto the rocks. The introduction of methods of precommitment is always a useful public policy intervention.

### 3.2.3 The ($\beta, \delta$) Model of Self-Control

Rational choice implies time consistency, namely that decisions are not sensitive to timing. Time consistency means that an initial consumption plan can be constructed, and that this plan will not need to be revised as times passes. In many domains early decisions are not carried out because consumption plans change over time. There are many illustrations such as: next month I will quit smoking, next week I will study and catch up on my homework, tomorrow morning I will wake up early and exercise; after the Christmas vacation I will start eating better; next weekend I will send in this form, next month I will start saving, and so forth. Early plans tend to have gratification up front and the “good” behavior to follow later (lie in today, but get up at 6 am tomorrow to finish the problem set), but when tomorrow comes instant gratification is again chosen and the “good” part of the plan is delayed. It is as if people are inhabited by multiple selves that disagree (internal conflict). Early selves make plans and choices that later
selves will not want to follow. Plans made at a distance tend to be more patient than choices made in the present. Dynamic inconsistency implies a conflict between early and late selves (i.e., a preference reversal). It is in essence a self-control problem (like procrastination, laziness, addiction, or compulsive consumption) where people cannot act according to plans.

There is a simple way to model the self-control problem. As it involves intertemporal choice, we need to use discounting factors for future utility levels. The self-control problem can be represented by the \((\beta, \delta)\) model (or quasi-hyperbolic utility)

\[
U = u_0 + \beta \delta u_1 + \beta \delta^2 u_2 + \ldots + \beta \delta^T u_T,
\]

(3.1)

where \(\delta < 1\) is the standard discount rate and \(\beta \leq 1\) is self-control discounting. For \(\beta = 1\), there is no self-control problem. For \(\beta < 1\), the immediate future is more heavily discounted than the more-distant future (this is the “present bias”): in the long-run we are relatively more patient than we are in the short run. The discount rate between any two periods in the future is \(\delta\) whereas the discount rate between the present (time 0) and the immediate future (time 1) is \(\beta \delta\).

To see how the model works, consider a consumption decision involving utility \(u_1\) at \(t = 1\) and delayed utility \(u_2\) at \(t = 2\). If it is an investment good (like exercising, studying, training, or savings), it has the feature that people must trade off the cost \(u_1 < 0\) against a future benefit \(u_2 > 0\). If it is a temptation good (like compulsive consumption, unhealthy food, surfing on the web, or credit card usage), it has the feature that people must trade off the reward \(u_1 > 0\) against a future cost \(u_2 < 0\). What is the consumption decision from an ex ante perspective?

If the consumer could commit to a choice in advance, say at time \(t = 0\), she would consume if \(\beta \delta u_1 + \beta \delta^2 u_2 \geq 0\) or, equivalently,

\[
u_1 + \delta u_2 \geq 0.
\]

(3.2)

Note that the parameter \(\beta\) cancels out in the desired (future) consumption choice. However, the consumer actually consumes at time \(t = 1\) if

\[
u_1 + \beta \delta u_2 \geq 0.
\]

(3.3)

Compared to the desired level of consumption set in advance, the naive individual (with parameter \(\beta < 1\)) underconsumes investment goods (with delayed benefit \(u_2 > 0\)) and overconsumes temptation goods (with delayed cost \(u_2 < 0\)), since \(\beta \delta < \delta\). This is the self-control problem. Compared to the actual consumption, the naive individual overestimates the consumption of investment good \((u_2 > 0)\) and underestimates the consumption of temptation good \((u_2 < 0)\). Conversely, a sophisticated agent (without
self-control problem, so $\beta = 1$) will actually consume according to the plan since $\beta \delta = \delta$.

### 3.2.4 Reference-Dependence Bias

One explanation for some of the observed anomalies is that people assess alternative options by comparing them to a reference point. The reference point might be a previous level of consumption or a target level of consumption. Whatever the explanation, the key assumption is that utility is measured relative to the reference point.

Denote the reference point by $r$. Utility depends on a combination of an absolute (possibly stochastic) consumption utility, $m(x)$, and a penalty function $\mu(m(x) - m(r))$, which is reference dependent since it is determined by the deviation $m(x) - m(r)$ between consumption utility and reference utility. When the reference point is the same as the bundle chosen, then $x = r$, and the penalty term disappears, so the model reduces to standard consumer choice. This approach creates multiple equilibria, opening up a role for marketing, advertising, and sales prices to influence preferences by creating and changing the reference point. This approach also helps explain conformism and the effect of experience on adaptive preference. Reference effects can be seen by comparing the situation of owners and dealers in either the housing market or the car market. Dealers do not expect to hold on to goods they receive. Since their reference point does not include the goods, they do not experience a loss when selling them. In contrast, the owner of a car and the owner of a house will exhibit some endowment effect according to which their reference points include those goods, and so they feel more of a loss when selling them. This is a general difference you feel when you buy a good for resale rather than for utilization.

Another illustration of reference-dependent preferences is the fact that people value income changes as well as income levels. Standard preferences involve valuing only income and consumption levels. Reference-dependent preferences assume that the value function $v(x; r)$ is defined over differences from a reference point $r$ instead of over the overall income level. A simplified version involves the following reference-dependent preferences over income, $x$:

$$v(x; r) = \begin{cases} x - r & \text{for } x \geq r, \\ \lambda(x - r) & \text{otherwise}. \end{cases}$$

(3.4)

The parameter $\lambda > 1$ denotes a loss aversion parameter that overweights losses: the value function is steeper for losses below the reference point ($x < r$) than for gains.
This simple formulation can explain the asymmetry in the valuation of small equal-sized losses and gains exhibited in many experiments.

Loss aversion can also explain the endowment effect. Consider the housing market. Homeowners willing to sell their houses are likely to fix the sale prices with reference to the initial purchase prices. A homeowner values the sale according to the extent of deviation from the purchase price. If there is loss aversion, the homeowner will overweight a price loss compared to a price gain. The homeowner who fears selling at a loss is willing to ask a higher sale price. Obviously a higher sale price will increase the utility of a sale, but it will also reduce the probability of a sale. The homeowner will trade off these two opposing effects. The loss-averse homeowner will sell above the purchase price. We observe similar application in the stock market where the tendency is to sell “winners” and hold back “losers.”

The reference-dependent model can also explain some anomalies in response to price changes. Consider the labor supply response to an hourly wage increase. For a rational worker a higher hourly wage induces longer working hours (if the substitution effect dominates the income effect). However, for a reference-dependent worker, the wage increase may well reduce her labor supply if the wage increase shifts income above the target income used as the reference point used to value labor choice. The worker achieves her target income by working less. Similar reasoning applies to the response of saving rates to interest rates: the household achieves the target, saving income, by saving less when the interest rate is higher.

3.2.5 The Gambler’s Fallacy

The gambler’s fallacy is the first of three behavioral anomalies we consider that involve the mistakes people make when forming beliefs, and these mistakes distort their decisions.

The least controversial type of mistake about objective, real world facts is non-Bayesian statistical reasoning. Such errors are likely to affect investment decisions and many other economic decisions under uncertainty. Consider the simple example of coin flips. Let \{h, t\} represents a lottery that pays $h if the next flip of a coin comes up heads (H) and $t if the next flip comes up tails (T). Consider the following situation: If the person observes HHH, she chooses a lottery with \( h < t \); that is, she chooses the lottery that pays more if the next flip is T. If \( h = t - \Delta \), this is equivalent to a bet of the amount \( \Delta \) that T will come up next. If instead she observes that the previous flips are TTT, she chooses a lottery with \( h > t \); that is, the lottery that pays more if the next flip is H. If \( h = t + \Delta \), this is equivalent to a bet of the amount \( \Delta \) that H will come up.
next. And finally, if she has observed no flips before, she chooses a lottery such that \( h = t \), considering that heads and tails are equally likely. These choices are wrong in the sense that no matter what previous outcomes have appeared the probability of \( h \) arriving next (or \( t \)) is always one-half.

The specific pattern of mistakes in these choices are called the gambler’s fallacy: the person believes that if the same realization of the random process has occurred a number of times, the other realization is in some sense “due” for the next draw.

### 3.2.6 Confirmation Bias

Confirmation bias arises from inferring *less* than what is justified from the observation of a recent event. This is the tendency to perceive data as more consistent with a prior hypothesis than they truly are. The agent updates her information based on unfolding observations by overweighting information that confirms her initial opinion and underweighting information that contradicts her initial opinion. This is a Bayesian updating framework, except for the mistake (bias), in the encoding of data. The gambler’s fallacy is a special case with the misperception that successive samples are drawn without replacement.

Another deviation from Bayesian rationality is irreversibility. According to Bayesian rationality, if you discover that a piece of information is mistaken, the memory should erase it so that it will have no impact on future judgment (i.e., information reversibility). However, the brain is such that when new information merges with old information, it is impossible to undo the effect of the old information even when such information is mistaken. Information in the brain is long-lasting. For example, when juries are instructed to ignore certain statements after they have been heard, it is hard for them to fully ignore the statements when making their final judgment. Information sticks where it hits.

Hindsight bias is the opposite of confirmation bias: it is inferring *more* than is justified from the observation of a recent event. It reflects our tendency to rapidly rewrite our memory of the past to fit what we have just learned. Rapid rewriting creates “hindsight bias”; that is, the ex post recollection of the ex ante probability of an event will be biased in the direction of the event’s realization. The problem is that revising our beliefs rapidly reflects more about how little we knew before the event, and so we should not reach too quickly to make strong beliefs out of recent events. Hindsight bias is on display every day in sport events and news coverage. It is an important force in political life and in organizational life.
3.2.7 Confidence Bias

Many studies show that people are often overconfident. Think about cycling without a helmet or driving without a seat belt because we believe our skills help us avoid an accident. Overconfidence is the tendency to overestimate one’s own (relative) abilities and expect the resulting outcomes to be better than they will be. Similarly overoptimism is the overestimation of general prospects.

There are many ways to model overconfidence with different practical implications. One possibility is that people overestimate the output they can generate, or they overestimate the marginal productivity of their effort. In either case they may end up striving less hard than if they were not overconfident. A different possibility is that overconfident persons think they are more skilled and talented than they really are. Drivers are overconfident about overall driving ability. This is especially true for young drivers and men, but much less so for women and mature drivers.

Whether overconfident persons will exert too little effort or too much will depend on whether effort and skill are complementary. Obviously overconfidence is not uniform in the population. For example, some studies show that women are less (over)confident than men, which might explain why women feel the need to work harder at school to achieve success.

3.2.8 Framing Bias

Any theory of rational choice must stipulate that the same problem will be evaluated in the same way regardless of how the problem is described; thus different but equivalent descriptions should lead to the same choice. Framing effects violate this bedrock normative condition of “description invariance.” As already discussed in the saving example, a framing effect occurs when different but equivalent descriptions of a decision problem lead to systematically different decisions, and this is commonly taken as evidence for incoherence in individual choices.

The best-known framing problem is risk framing. It was first described by the so-called Asian disease problem (Tversky and Kahneman 1981). Participants to the experiment are first told the following story: “The United State is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. One possible program to combat the disease has been proposed.” Then some participants to the experiment are presented with two options. A: If this program is adopted, 200 people will be saved. B: If this program is adopted, there is a one-third probability that 600 people will be saved and a two-thirds probability that no people will be saved. Other
participants are presented with two other options. C: If this program is adopted, 400 people will die. D: If this program is adopted, there is a one-third probability that nobody will die and a two-thirds probability that 600 people will die. The robust experimental finding is that subjects tend to prefer the “sure thing” A when given options A and B, but tend to prefer the gamble D when offered options C and D. Note, however, that options A and C are equivalent, as are options B and D. Subjects thus appear to be risk-averse for gains preferring option A to B, and risk-seeking for losses, preferring option D to C. This is a central feature of prospect theory (Kahneman and Tversky 1979). In prospect theory, it is the decision maker’s private framing of the problem in terms of gains or losses that determines her evaluation of the options; the framing manipulation is thus viewed as a public tool for influencing this private frame.

There are many other framing problems. In attribute framing a single attribute of a single object is described in terms of either a positively valued attribute or an equivalent negatively valued attribute. The subject is then required to provide some evaluation of the described object. The typical and robust finding is that objects described in terms of a positive attribute are generally evaluated more favorably than equivalent objects described in terms of negative attribute. For example, in one study, beef described as “75 percent lean” was given higher ratings than beef described as “25 percent fat.” In goal framing subjects are urged to engage in some activity (e.g., wearing seat belts). This plea involves a description of either the advantages of participating in the activity or the corresponding disadvantages of not participating. The most common result is that subjects are more likely to engage in the activity when the disadvantages of not engaging, rather than the advantages of engaging, are emphasized.

The framing effects reflect more generally the fact that human perception and cognition is heavily influenced by contrast. This is clear in the Titchener illusion of circles, where a circle looks larger when surrounded by smaller circles than when it is surrounded by larger circles. Since choices involve basic perceptions, it would then be surprising if choices were not sensitive to contrast as well. Similarly the comparison of an outcome with unrealized outcomes (disappointment) or with outcomes from forgone choices (regret) imply that the appeal of choices depends on the set of choices they are part of. There are situations where the possibility for more choices implies that fewer choices are made, or even no choice altogether. It is well documented that offering a broader set of choices can lead consumers to buy less, whereas in the standard model more choices can only lead to more purchases. The explanation is that the broader opportunity set makes choice more difficult and stressful, so some individuals may prefer to avoid choosing. Similarly individuals
facing difficult choices may prefer to go for a default option such as the “menu of the day” (or the house wine) in a restaurant offering a very large choice of menus (or wines).

### 3.2.9 Conformism Bias

Conformism in social psychology refers to the inclination of an individual to change spontaneously, without any explicit order or request by someone else, her opinion (beliefs) and or action (choices) to conform to the socially prevailing opinions and actions. It is the individual tendency to change the intrinsic optimal choice toward the most prevailing choices within a group.

We could attribute this conformism to some general force of learning, such as imitation or suggestion, regarded as innate and instinctive. There is nothing wrong with that. However, it could also reflect a possible lack of autonomy leading to individual mistakes when making choices. The conformism effect was forcefully demonstrated in Asch’s famous experiment. The experiment consisted of many trials in which a subject was placed in a group of people who were secretly accomplices of the experimenter. The subject was asked to estimate the length of a line by matching it with one of three lines. This estimate was provided after the other accomplices had successively expressed their opinions. The reality of conformism emerged when the other accomplices announced a clearly wrong comparison line, as then about one-third of the tested subjects revised their own (correct) opinions to conform to the wrong judgment of the group. This finding is important because it violates one of the basic postulate of standard economics, that is, the idea of full autonomy of the individual. Indeed, despite the lack of any uncertainty about the correct judgment, agents may renounce their preferences (or judgments) and conform to an erroneous choice of the others. This is conforming to social pressures.

Unlike sociopsychologists, economists are reluctant to take conformism as a primitive assumption but prefer to derive conformism endogenously. A first possible explanation for this conformism bias is that individuals may suffer from being “different” and that conformism is a reaction to this. Another explanation assumes agents care about “status,” which can only be inferred from their actions. If all agents prefer to be perceived as “good” by others and if they all agree about what is a “good” type, then they will all make uniform choices through fear that they would appear “bad” by deviating from the norm.

A simple way of modeling conformism is to extend the utility function of an individual $i$ to include a penalty $p(x_i, m)$ that depends on the distance between her individual
choice \( x_i \) and the “normal” choice \( m \). The penalty function must be of the first order, for otherwise a small deviation away from the norm toward the intrinsic preferred choice would be beneficial as it would only cause a second-order loss in the penalty function and first-order gain in the intrinsic utility. Obviously the dispersion of choices is important because the smaller the dispersion, the greater the degree of conformism. A suitable penalty function that takes account of the dispersion of choices and the first-order condition is the index of individual conformity measuring the standardized distance from the mode:

\[
p(x_i, m) = \frac{|x_i - m(x_i, x_{-i})|}{\sigma(x_i, x_{-i})}, \tag{3.5}
\]

where \( x_i \) is the choice of individual \( i \), \( m(x_i, x_{-i}) \) is the mode and \( \sigma(x_i, x_{-i}) \) is the standard deviation around the mode (where \( x_{-i} \) denotes choice variables pertaining to anyone other than individual \( i \)). This index is a simple and perhaps realistic representation of how individuals synthetically compare themselves to others.

### 3.2.10 Identity and Social Norms

There is a substantial psychological literature on group identification that is important for social interaction and market allocation. A special case is the “identity economics” coined by Akerlof and Kranton (2010). To get a grasp of this approach, consider the relative performance of boys and girls at school. It is a general feature and well-documented fact across various PISA (Program for International Student Assessment) studies that girls perform better on average than boys at age 15. Why is it so? This cannot be due to better salary prospects, since women are paid less and are more likely to work part time. Pursuing the analysis is in fact intriguing because we cannot attribute such difference to either family background or migration status nor to gender difference in those factors. We cannot attribute such gender difference to school difference either (because they attend the same schools in general). Finally, we can hardly claim gender difference in cognitive ability. So there is something else. Something less visible and obvious but still very important. What could it be? This is where identity and norm come in.

When we examine people’s decisions from the perspective of their identities and social norms, we provide new insights into many different economic questions. Who people are and how they think of themselves is key to the decisions that they make. Their identities and norms are basic motivations. Tastes vary with social norms. This vision of tastes is important because norms are powerful sources of motivation. Norms
affect fine-grain decisions of the moment. Norms drive life-changing decisions as well: on matters as important as whether to quit school, whether to go to university, or whether to go to work.

The important determinant of whether an organization functions well is not only the monetary incentive system, as standard economic models would suggest, but also how well its members identify with the organization and with their activities within it. Their work must have some meaning for them to function effectively in a company. Workers may well be willing to trade a reduction in wages for “meaningful” work. This is probably a central feature in the occupational choice of many workers in public services with a real vocation of serving the public (e.g., in education, health, and justice services). This effect refers to intrinsic motivation: the satisfaction a worker gets from work for its own sake. An interesting phenomenon documented in psychology is the possibility that extrinsic incentives (e.g., money) can “crowd out” intrinsic motivation. Blood donation is a concrete illustration. In general, if workers do not identify with their job, they will seek to game the incentive system, rather than to meet the organization’s goals. Likewise good schooling occurs not as a result of monetary rewards and costs but because students, parents, and teachers identify with their schools, and because that identification is associated with learning. Given this, education policy should look at what some successful programs have done to establish a school identity that motivates students and teachers to work according to a common purpose. If we focus on training teachers in how to inspire their students to identify with their school rather than teaching students to take standardized tests, we just might be able to reproduce within these schools great results.

As economists and policy makers we could be content to continue looking only at prices, incomes, and related statistics to explain people’s decisions. In some situations, that might be enough to understand what is happening. But in other situations, we would miss major sources of motivation and thus could adopt useless, if not counterproductive, measures aimed at producing the outcomes we seek.

### 3.3 Behavioral Markets

In 1991 Vernon Smith attacked Daniel Kahneman (before both received Nobel Prize in 2002 for work in behavioral economics). Smith’s claim was that anomalies at the individual level play no role at the aggregate level, in particular in competitive markets. Introducing the possibility of systematic imperfections in individual rationality, as we all are willing to accept, raises new questions in the study of markets. Among them
the central question is whether the market will erase or exploit limits on consumer rationality? Does competitive pressure eliminate irrational choices and induce agents to make rational choices?

Individuals make mistakes that markets do not fully correct. There are three factors limiting the extent to which the market can remedy biased decisions. First, decisions are not frequent and do not deliver clear feedback. Second, individuals are not specialized in making those decisions. Third, individuals are protected from market pressure and competition.

3.3.1 Money Pump

Davidson, McKinsey, and Suppes (1955) use the money pump argument to justify rational choices. The argument is as follows. Suppose that a consumer has nontransitive preferences for consumption bundles \( a, b, c \), where \( a \succ b, b \succ c, \) and \( c \succ a \). The consumer has cyclic preferences, preferring \( a \) to \( b \), \( b \) to \( c \), and \( c \) to \( a \), meaning that the consumer is willing to trade \( a \) for \( c \), next \( c \) for \( b \), and then \( b \) for \( a \), so getting back to the initial bundle. That is, the cyclic consumer is always willing to pay a small amount \( \Delta \) of money to get \( a \) instead of \( b \), \( b \) instead of \( c \), and \( c \) instead of \( a \). So, by allowing the agent to cycle between the different bundles \( c, b, a, c, b, a, \ldots \) against successive small payments of \( \Delta \) units of money, the market will “pump” an indefinite amount of money out of the consumer. This argument establishes that an irrational consumer with intransitive preferences is doomed to bankruptcy when operating in the market. It suggests that all intransitivities should be removed from a consumer’s preference.

3.3.2 Complementary Mistakes

Whether individual mistakes are erased or exacerbated depends on whether behaviors are strategic substitutes or strategic complements. When behaviors are complements (like in the stock market with buying or selling decisions), a small number of irrational traders can force others to behave irrationally.

A good illustration is the guessing game. This guessing game is like a market bubble where market participants are not rational and the market not in equilibrium. Consider a classroom experiment where students are asked to pick a number between 0 and 100 and not to let others see their pick. The winner of the contest is the student who is closest to two-thirds of the average number picked by all students (hence the name of Guessing Game). Ties will be broken randomly. In this game there is no dominant
strategy (i.e., a unique best guess independently of the guess of others). However, there is a unique Nash equilibrium when every player is rational and expects each other to behave rationally. To compute it, ask what guess would be irrational and eliminate this guess. For instance, any guess above 66.67 is irrational for every player, since it cannot possibly be two-thirds of the average guess. These guesses can reasonably be eliminated for every player, but then since no player is expected to guess above 66.67 and two-thirds of 66.67 is approximately 44.45, any guess above 44.45 is also irrational. This process of iterated elimination of weakly dominated guesses will continue until all guesses above 0 have been eliminated.

Now when the experiment is performed among ordinary students, it is usually found that the winning guess is much higher than 0. Some students guessed close to 100 indicating they did not understand the game at all. A large number of students guessed 33.3 (i.e., two-thirds of 50), indicating an expectation that other players will guess randomly. A small but significant numbers of students guessed 22.2 (i.e., two-thirds of 33.3), indicating a second iteration based on an assumption that others would guess 33.3. In many experiments the average guess was around 33, and the winning guess was around 22. So we can see that the rational equilibrium does not predict well in this strategic environment. Interestingly, even perfectly rational players participating in such a game should not guess 0 unless they know that others players are rational. If a rational player believes that others are not rational, she will not follow the chain of elimination described above and she could rationally guess above 0.

Keynes (1936) believed that similar behavior was at work in the stock market and could explain a market bubble. This is the case when the price of shares is not based on what people think their fundamental value is; rather, the price is based on what they think everyone else thinks their value is, or what everybody else would predict the average value is.

The nonequilibrium approach assumes a “cognitive hierarchy” in which more “rational” players best-respond to the perception that others do less thinking. Nagel (1995), based on an original idea of Hervé Moulin, suggested that people based their guesses on different levels of rationality. She found many guesses on different levels of rationality: level 0 rationality (guessing 50); level 1 rationality with best response to level 0 rationality (guessing 33 in response to 50); level 2 rationality, with best response to level 1 rationality (guessing 22 in response to 33); and so on. These cognitive hierarchy approaches are more precise than Nash equilibrium because they always predict a single statistical distribution of play, and are generally more accurate than equilibrium in predicting behavior. Interestingly, if the experiment is re-run with the same players, the results will now come much closer to the rational prediction.
Chapter 3: Behavioral Economics

3.3.3 Rationality Tug-of-War

Will the market and firms take advantage of limited consumer rationality or will the market and firms help consumers? Whether markets will correct irrationality depends on factors like whether consumers know their own limits and hence are receptive to advice, and whether there is more profit in protecting consumers or taking advantage of them.

As an illustration consider the Gabaix–Laibson (2006) model. There is a market of products with “add-on” prices (e.g., bank transaction fees that can be easily hidden). If consumers do not know about the hidden add-on price, then competitive firms will offer a low price on base goods (below marginal cost) and will charge high markups on add-ons. Sophisticated consumers who know the add-on price, but can easily substitute away from the add-on (e.g., avoiding bank ATM fees), will prefer products with expensive add-ons to benefit from the low base-good price. The naive consumers are subsidizing the sophisticated consumers. As a consequence competition does not lead to add-on prices being revealed (to protect the consumers) because a firm that reveals its add-ons will not attract either naive consumer (mistakenly thinking price is too high) or sophisticated benefiting from low base-good price.

3.4 Behavioral Policy

The standard economic approach assumes that people make appropriate decisions when they are well informed. The role of the government is then to combat ignorance or misinformation. So, if the government can provide relevant and reliable information more effectively than private markets, it will be a potentially beneficial intervention for consumers to do so. Information and education campaigns to combat ignorance are desirable to help people make informed choices.

Assuming information is not an issue, the reasons for government intervention still involve market failure. It may be appropriate, for example, for the government to tax pollution and subsidize charitable contributions to ensure adequate levels of their provision (see part III on departures from efficiency). However, once market failures are corrected, under the standard approach there will be nothing wrong with the choices people make, given the constraints they face (obviously that does not rule out government intervention to enforce property rights and redistribute resources!).

In practice, benevolent policy makers worry that people make inappropriate choices. One important example is that governments are concerned that people save “too little”
and that this is a systematic mistake given their present bias. Such decision-making failures are another justification for intervention.

3.4.1 Internalities versus Externalities

Behavioral public policy permits the possibility of decision-making failures, and this opens up the possibility of enhancing individual welfare by correcting or preventing “bad” choices. Behavioral policy is an extension of standard public policy. The common feature is that public policy can change behavior by changing relative prices, budgets, and information. The distinctive feature of behavioral policy is that there are additional channels through which policy can change behavior and welfare, even if the policy leaves prices, budgets, and information unchanged!

The central idea is that if people make systematic mistakes, or biased decisions they regret later on, behavioral policy uses those mistakes and biases that hurt people (1) to help them protect themselves and, at the same time, (2) to respect their autonomy of choice.

What do all these “invalid” choices have in common? People are facing the wrong prices.

With traditional public policy, prices are wrong or misaligned because of externalities. Externalities are costs that people impose on others but do not internalize. So, in the presence of externalities, the prices people pay for things do not reflect the “true cost” to others: the market price is wrong. The point will be explored in depth in chapter 8.

With behavioral public policy, prices are wrong because of internalities. Internalities are costs that people impose on themselves but do not internalize. So, in the presence of internalities, the prices people pay for things do not reflect true costs to themselves. This is so because people face internal conflicts in their choices: the question “Do you want a piece of chocolate?” triggers an internal clash between temptation and reason for someone on a diet program. People might have present bias, reference-dependent bias, and so on. It is not enough simply begging people to do the right thing. The government must realign prices and incentives so that it is in people’s own interest to do the right thing.

Behavioral policy uses people’s biases to help them by making the healthy option the default (status quo bias) and by giving immediate rewards for healthy choices (present bias). Behavioral policy supercharges economic incentives with deposit contracts and regret lotteries to act as commitment devices against people’s time inconsistency. Behavioral public policy is consistent with the traditional justification for public intervention (the enforcement of property rights, correction of market failures, redistribution
of income). It also introduces justifications for intervention, notably by allowing public policy that raises welfare by limiting the possibilities for decision-making failure and its consequences.

### 3.4.2 Automatic Enrollment

Consider the public policy of automatic enrollment to a saving plan or organ donation scheme with a small cost to opt out. The default option is the outcome resulting from inaction. A rational consumer is not influenced by the default option because the cost to opt out is small. But behavioral consumers are influenced by the default option because of status quo bias. In practice, the choice of the default option matters: for organ donation schemes and retirement saving plans there is considerable evidence that the default option affects participation rates, even though such a default neither affects opportunities (low cost to opt out) nor provides new information. Johnson and Goldstein (Science 2003) report a significant effect of automatic enrollment into organ donation: the consent rate is 85.9 percent in Sweden with automatic enrollment against 4.3 percent in Denmark where there is no automatic enrollment. The introduction of automatic enrollment in 401(K) pension plans in the United State also provides an effective default option (costlessly changeable by employees). The default contribution rate is 3 percent of the salary into the 401(K) plan and 100 percent of allocation to the money market. Before the automatic enrollment, the participation rate was 40 percent with a dispersion of contribution rates. After the automatic enrollment, the participation rate was 88 percent without dispersion of contribution rates, which are mostly concentrated around the default rate of 3 percent. This automatic enrollment policy looks nonpaternalistic because the desirable default can improve the welfare of those who mindlessly adhere to the default without restricting the options of those who do not. The policy faces little opposition because it is beneficial to those who believe there is a status quo bias, and it is harmless to those who believe there is no status quo bias.

### 3.4.3 The SMarT Plan

People accumulate an inadequate level of savings because of the different biases involved. The present bias implies a willingness to save, but only tomorrow. The opportunity-cost bias implies a willingness to save but only out of pay increases. The opportunity-cost bias reflects the tendency for people to treat “out-of-pocket” costs differently from “opportunity costs.” People tend to underweight (or neglect)
opportunity costs relative to out-of-pocket costs. The status quo bias implies a willingness to adhere to a saving plan once implemented. Combining these different biases, Benartzi and Thaler (2004) designed a saving plan to help workers save more. The plan was implemented in a workplace with employees invited to join a saving plan. They could elect in advance a portion of their current income, and also a possibly different portion of future income, to be saved. This program induced large increases in saving. Although workers chose to save little out of current income, they committed to save a large portion of future income. In a short period of time, the average saving rate increased from 4.4 to 8.8 percent.

3.4.4 Complementarity

Behavioral public policy should complement, not replace, more substantive public policies. For example, if standard public policy suggests creating a price differential between healthy and nonhealthy food, or between bottled water and a soda drink, behavioral analysis could help us better understand the consumer response to various forms of public intervention. Behavioral analysis could suggest whether consumers would respond better to a subsidy on bottled water or to a tax on regular soda. It would also suggest how to complement a tax-subsidy policy with more effective labeling of healthy/nonhealthy food. But that’s the most it can do according to the behavioral economists themselves.

The limits to behavioral policy arise from social interaction. Indeed a key difference between psychologists and economists is that psychologists are interested in individual behavior while economists are interested in explaining the outcome of social interaction among many individuals. Behavioral economics focuses on human dysfunction (interpreted as making irrational choices) with a perspective to help people become more functional. But, in general, most people are sufficiently functional most of the time. Hence the focus of economists on people who are “rational” remains a useful benchmark to form policy choices. Behavioral analysis can contribute to the improvement of existing policies, but it offers no realistic prospect of replacing policies. The narrow and complex models of behavior used in psychology cannot easily be used to study the behavior of many people interacting. However, it is the social interaction of people that will eventually determine the final outcome of any policy intervention. The need to study groups that consist of large number of people requires constraints on economics that are not present for the psychologist. Economists need simple and broad models of behavior. Hence economists focus on rational and selfish behavior that provides a reasonable description over a broad range of social settings.
Chapter 3: Behavioral Economics

3.5 Behavioral Welfare

Standard welfare analysis is concerned about the evaluation of public policy. To evaluate whether a policy is good or bad, we must evaluate its effect on individual welfare levels. The effect of policies on welfare are calculated in two stages: (1) the effect of the policy on behavior and (2) the effect of the change in behavior on welfare.

The standard welfare approach to compute the effect of policy on behavior is to assume individuals will rationally respond to policy change. Then the preferences revealed by choices are used to compute the effect of change in behavior on welfare. The welfare effect of policy intervention is measured by the extent to which individual preferences are satisfied. The standard preference revelation axioms assume that the choices people make are valid and thus correctly reveal their true preferences. There is no conflict between actions and intentions. The assumption that preferences are revealed by choices can be traced back to Vilfredo Pareto (or Wilfried Pareto at birth). Pareto recommended the use of choices (“objective facts”) to reveal preferences (“subjective fact”). This is a philosophical stance and not a robust empirical regularity. Pareto justified his assumption by restricting attention only to repeated actions so that rational choices emerge as the consequence of learning. For instance, credit card account holders learn to pay their bills on time by first suffering the payment of a late fee when they do not. But because Pareto clearly limited the domain of revealed preference to “repeated actions” in which learning has taught people what they want, he leaves out important economic decisions that are rare, partly irreversible, or difficult to learn about from trial and error: educational choices, occupational choices, retirement and saving plans, fertility and mate choices, housing choices, and so forth. In principle, we could consider that people learn by observing others, but people are generally far more responsive to their own experiences than to the experience of others.

3.5.1 New Welfare Criterion

The real challenge with the behavioral approach is that people can make mistakes. Choice mistakes represent a conflict between actions and intentions. So the question is then how are (true) preferences revealed when choices reveal mistakes rather than preferences, that is when actions do not reveal true intentions? There are many cases, as described in earlier sections of this chapter, in which even the choices of mature consumers do not reveal a true preference but rather reflect the combined influence of true preferences and choice mistakes.
When people cannot act according to plan or when people do not save enough for their retirement, we cannot assume that they are acting in their own best interest, and that those decisions (or nondecisions) are revealing their true preferences. If preferences are only imperfectly revealed by choices, what yardstick can we use to evaluate policy recommendations? The danger is to replace individual preferences by some ad hoc external preferences to legitimate policy choices. This is the danger of paternalism such as when parents teach their children how to behave based on their own preferences. It may be legitimate to do this for children, but it is not legitimate for adults in full control of their autonomy (unless they voluntarily accept to abandon their freedom of choice). Often there exists a compromise, so economists may use choices to identify (true) preferences but take care to acknowledge the possible wedge between revealed preferences and true preferences.

3.5.2 Choice-Based Welfare Analysis

In standard theory, agents decide on choice, \( x \), from a set of possible choices, \( X \). The goal of policy is to identify the optimal choice \( x^*, x^* \in X \).

In behavioral models, agents choose from \textit{generalized choice sets} \( G = (X, d) \) where \( d \) is an \textit{ancillary condition} that affects choice but by assumption does not affect (true) preferences (e.g., salience, framing, default option). Let \( C(X, d) \) denote choice made in a given generalized choice sets \( G \). Choice mistakes and inconsistent choices imply that different ancillary conditions \( d \neq d' \) lead to different choices even if the choice set \( X \) is unchanged:

\[
C(X, d) \neq C(X, d') \quad \text{for } d \neq d'.
\]  

(3.6)

We can thus define the revealed preference relation \( P \) as \( x P x' \) if \( x \) is always chosen over \( x' \) for any ancillary conditions \( d \). Using the revealed preference relation \( P \), it is possible to identify the choice set that maximizes welfare instead of a single point.

With sufficiently many observed choices, it is effectively possible to obtain bounds on welfare. To illustrate this, consider three different saving plans with varying benefits and corresponding contributions rates: high (\( H \)), middle (\( M \)), and low (\( L \)). Suppose that we have collected observations from two different framing conditions, \( d \) and \( d' \). In frame \( d \), revealed preferences are \( H > M > L \) and the consumer chooses saving plan \( H \), whereas in frame \( d' \) preferences are \( M > H > L \) and the consumer chooses saving plan \( M \). We do not need to understand why the frame affects the choice of saving plan to make a welfare statement about the optimal policy. Indeed \( L \) cannot be optimal given the observed choices because it is never chosen no matter what the framing condition.
is. Therefore the optimal policy must be bounded between \( M \) and \( H \). That delivers bounds on welfare based entirely on choice observations.

The revealed preference relation can identify the set of choices that maximizes welfare (but not the unique optimal choice). Welfare bounds are tight when choices are less sensitive to framing conditions, that is, when behavioral problems are small. However, welfare bounds and the set of optimal policies are large when behavioral problems are large. In the previous example, if there exists another frame \( d'' \) such that \( L > M > H \), then \( L \) would be chosen and the welfare bounds would include \( L, M, \) and \( H \). That is, any saving plans could be the optimal policy. So this approach is not restrictive enough to generate policy prescriptions when ancillary conditions exist that lead to vast changes in choices. There are two alternatives solutions: preferences refinements and the structural model.

### 3.5.3 Refinement and Structural Modeling

The idea behind preference refinement is to discard certain ancillary conditions that have become too “contaminated” for welfare analysis. For instance, by dropping the nonvalid framing \( d'' \), we can eliminate plan \( L \) from the set of optimal policies. With fewer ancillary conditions we have more restrictive bounds on welfare and policy. There is a good argument that supports this alternative for redistributive policies and the importance of the status quo. According to reference dependence, people tend to concern themselves more with income change (gains and losses) than income levels. Moreover feeling the impact of loss is larger than feeling the impact of gain (loss aversion). As a result people will give higher subjective weight to avoiding a loss than experiencing a gain. Status quo comes into play if the redistributive policies cause the rich to lose and the poor to gain. Due to loss aversion more prominent weight is given to the rich relative to the poor when a redistributive policy is being evaluated.

Structural modeling takes a different approach. The idea of mapping observed choices directly into statements about welfare is abandoned. Instead, the observed choices are interpreted using a behavioral model that seeks to explain their deviations from rationality. The objective is to discover preferences by building a behavioral model that can explain how ancillary conditions affect the choices, and then to use this model to predict which choices reveal true preferences. By this approach, it is assumed that preferences can be identified from the observed choices on the basis of a structural decision-making model. For example, one could construct a present-bias \((\beta, \delta)\) model that explains saving plan choices, and then calculate the optimal policy within such a model using normative discounting with \( \beta = 1 \).
3.5.4 Application: Global Warming

Global warming involves intertemporal trade-offs that raise important normative questions for public policy. We discuss the economics of climate policy more fully in chapter 26. Here we focus on the fact that any discussion of policy must invoke trade-offs between different time periods (bear a cost now to mitigate emissions in order to benefit from lower temperatures later) and between different generations (should the current generation emit pollution that changes the climate for future generations?). Indeed, whether or not people are rational, there are good reasons why revealed choices cannot be used to infer the true intertemporal preferences of an agent, and why the choices cannot be relied upon to make normative judgments. We consider in turn the rational model and the irrational model.

For people who are rational, there is no reason for the policy maker to give any normative value to the discount rate revealed by individual choices. Basically, why should the payoff at time $t$ have less weight than payoffs at time $t + \Delta$? A good explanation may relate to the risk of mortality. But then how can we account for the fact that young adults are markedly less patient than middle-aged and older adults? Yet, discounting to account for the mortality risk is too small to account for the revealed exponential discounting rate of around 5 percent per year. This is a matter of considerable importance for curbing global warming. Indeed the policy recommendations for addressing global warming are heavily dependent on the choice of the normative discount rate. The Stern Report (2006) used a normative discount rate of 0.1 percent per year to make its recommendations, namely a normative discount rate about 1/50th of the revealed discount rate of 5 percent. It is therefore not surprising to find such a big gap between the call for immediate and massive actions to curb CO$_2$ emissions in the Stern Report and what people seem willing to accept. The rational approach to intertemporal choice assumes a constant discount rate whereby agents make choices that are consistent over time. The exponential discount function with constant discount rate, $U = u(x_0) + \delta u(x_1) + \delta^2 u(x_2) + \ldots$, as originally stated by Ramsey (1928), is the only discount function that generates dynamically consistent choices: preferences held at some point in time do not change with the passage of time (unless obviously new information arrives). No preference reversal is often invoked as a rationality requirement.

If people are irrational, their choices may contradict their intentions. As was the case with the self-control problem, nonconstant discount rates imply dynamically inconsistent choices. For the global warming problem, suppose that an agent can make some investment at a cost of $C$ (i.e., pollution abatement costs) to gain delayed benefits
of B (i.e., curb global warming). For simplicity, suppose that the benefit occurs one period after the investment cost. The agent has a \((\beta, \delta)\) preference, also called a quasi-hyperbolic discounting function whereby cost and benefits at times 0, 1, 2, \ldots, \(n\), are discounted respectively by rates \(1, \beta \delta, \beta \delta^2, \beta \delta^3, \ldots, \beta \delta^n\), with \(0 < \beta, \delta \leq 1\). When \(\beta = 1\), the model is identical to the constant exponential discounting model. When \(\beta < 1\), this model replicates the “hyperbolic discounting” pattern with more periodic discounting in the short run than in the long run. Consider the case \(\beta = 1/2\) and \(\delta = 1\). Let \(C = 100\) and \(B = 180\) so that the undiscounted benefit is larger than the undiscounted cost. When evaluated from an earlier perspective \(t\) periods before the cost has to be incurred, \(t \geq 1\), this investment is desirable because

\[-\beta \delta^t C + \beta \delta^{t+1} B = -\left(\frac{1}{2}\right)100 + \left(\frac{1}{2}\right)180 = 90 > 0.\] (3.7)

But the investment becomes undesirable when the agent is asked to act immediately:

\[-C + \beta \delta B = -100 + \left(\frac{1}{2}\right)180 = -10 < 0.\] (3.8)

So the agent is faced with conflicting preferences.

When asked to make a binding commitment in advance, the agent will choose to invest. When precommitment is not feasible, the agent will not invest, since she always reneges from her previous plan when the moment arrives to act. In such a case revealed choices cannot be used to infer the true intertemporal preference of the agent, and cannot be relied upon to make normative judgments. Additional normative assumptions are needed. For instance, we could assume that \(\beta = 1\) in order to evaluate the policy choice from the perspective of the rational self of this agent with a no–self-control problem. Bernheim and Rangel (2005) provide a formal justification of this normative criterion based on aggregation principles when the consumer’s horizon is sufficiently long. The problem is similar to the welfare aggregation involving many individuals to rank policies: here we aggregate over multiple selves. The idea is that person at time \(t\) is a different person at time \(t+1\) due to the preference reversal. As in the problem with multiple consumers, it is possible to apply a multi-person welfare analysis. If the consumers’ horizon is sufficiently long, the aggregation is over many different selves. Then the reason for using normative criterion \(\beta = 1\) is that the consumer evaluates trade-offs between any two periods \(t\) and \(t+1\) by exactly the same discount rate in all periods but one; then the influence of anyone self must decline to zero as the number of selves becomes large. To put it differently, if we aggregate preferences according to the frequency with which rationality prevails, we end up using rational preferences
for normative analysis because the “momentary lapses” of reason approach zero when considered over a very long horizon. Obviously momentary lapses of reason matter for positive analysis because they have long-lasting effects.

### 3.6 Other-Regarding Preferences

Standard economics assumes that agents are rational and selfish. We have already discussed numerous deviations from rationality. We now discussed deviations from selfish behavior. Experimental economics has confirmed the predictions of competitive markets, namely that even with a limited number of participants, experimental markets clear at competitive prices. The experimental results give support to the predictions of the competitive equilibrium model analyzed in chapter 2. The equilibrium outcome was based on selfish optimization by agents interested only in their own material consumption and profits. In contrast in a *strategic environment* (where an individual choice can affect someone else), agents do not seem to act according to the standard model of selfish optimization. For instance, in the simplest ultimatum game that we describe next, the deviations from the standard model are systematic. The deviation is not from rationality but rather from the standard assumption of selfish behavior.

#### 3.6.1 Ultimatum Game

In the ultimatum game two players bargain over the distribution of a surplus of fixed size 1. The first player (proposer) chooses any share of the surplus \( s \in [0, 1] \). The second player (responder) either accepts or rejects the proposal. If the responder accepts the proposal \( s \), then the responder’s payoff is \( r(s) = s \) and the proposer’s payoff is \( p(s) = 1 - s \). If the responder rejects the proposal, both receive nothing \( p = r = 0 \). It is a Nash equilibrium for the proposer to offer \( s > 0 \) and for the responder to accept any offer greater or equal to \( s \). However, such an equilibrium is not fully rational (i.e., subgame-perfect) because it relies on the (noncredible) threat that the responder will reject the positive offers less than \( s \). So the unique (subgame-perfect) Nash equilibrium involves the proposer making an offer \( s = 0 \) (if the set of possible proposals is continuous) and the responder will accept the offer as matter of indifference. So in equilibrium \( s^* = 0 \), and the payoffs are \( r(s^*) = 0 \) and \( p(s^*) = 1 \).

The ultimatum game is simple with sharp predictions: the proposer demands essentially everything and the responder accepts. However, those sharp predictions are
systemically wrong in the sense that they are violated in most experiments. Low proposals giving less than $\frac{1}{2}$ to the responder are rare ($s < 0.2$) and proposals giving more than $\frac{1}{2}$ to the responder are also rare ($s > 0.5$). Equal or almost equal splits are frequent ($s \simeq 0.5$). Proposals are also rejected in some experiments, with the probability of rejection increasing as the responder’s share of the surplus $s$ decreases.

In a competitive environment the results from the play of ultimatum game experiments are again compatible with the predictions of self-interest. Consider a variation of the ultimatum game with a unique responder but $n > 1$ proposers (i.e., competition among $n$ proposers). The proposers simultaneously make an offer to the unique responder. So the list of offers is $s = (s_1, \ldots, s_i, \ldots, s_n)$. If the responder accepts the offer $s_i$ from proposer $i$, then the responder earns $s_i$, the proposer $i$ earns $1 - s_i$, and other proposers earn nothing. If the responder rejects all offers, then everyone earns nothing. The subgame perfect (Nash) equilibrium involves the responder receiving almost all of the surplus with $s^* \simeq 1$. And experiments confirm this equilibrium prediction of self-interest.

### 3.6.2 Social Preferences

In social organizations, people make friends and enemies, and compare themselves to others. Workers may thus sacrifice some potential earnings to help their friends and harm their enemies, or to create better social comparisons. It is also well documented that the perception of fairness is a key determinant of strike action. If people compare their own wages to those who work in similar activities, then the results may create turnover costs or costs in a social organization.

There are many different forms of social preferences, displaying Selfishness, Altruism, Fairness, or Envy. They can collectively be called SAFE preferences and they depend on how the others’ material consumption enters your own utility function. For the sake of clarity, consider an exchange economy with only one good (which we call income) and two individuals. Each individual has preferences, not only over her own income but also over the income of the other. Preferences are complete, transitive, and continuous with the resulting utility for individual $i$ being $v_i(y_i, y_j)$ (where $i \neq j$). Utility is increasing in own income, so $\frac{\partial v_i}{\partial y_i} > 0$. By this formulation, four types of social preferences are possible:

1. **Selfishness** Utility is independent of the income of the other:

   $$\frac{\partial v_i(y_i, y_j)}{\partial y_j} = 0 \quad \text{for all } y_i, y_j.$$  
   (3.9)
2. **Altruism** Utility is increasing in the income of the other. Altruism is a form of unconditional kindness to others:

\[
\frac{\partial v_i(y_i, y_j)}{\partial y_j} > 0 \quad \text{for all } y_i, y_j. 
\]  

(3.10)

3. **Envy** Utility is decreasing in the income of the other. An envious person always values the material payoff of other negatively. It is a form of unconditional enviousness to others:

\[
\frac{\partial v_i(y_i, y_j)}{\partial y_j} < 0 \quad \text{for all } y_i, y_j. 
\]  

(3.11)

4. **Fairness** Utility is either increasing or decreasing in the income of the other, if the other is respectively poorer or richer. It is a form of inequity aversion that can exhibit both altruism or envy to other depending on relative position:

\[
\frac{\partial v_i(y_i, y_j)}{\partial y_j} \begin{cases} 
\leq 0 & \text{for all } y_i \leq y_j, \\
> 0 & \text{for all } y_i > y_j.
\end{cases} 
\]  

(3.12)

If agents have extended preferences depending both on their own monetary payoffs and on the payoffs of others, then the equilibrium outcomes can be reconciled with experimental findings. In fact such extended preferences do a good job in organizing experimental results that are at odds with standard predictions (e.g., as in the ultimatum game described above).

### 3.6.3 Market Impact

What is the impact of other-regarding preferences on the market behavior? Consider the competitive economy of chapter 2 with price-taking consumers and firms but with other-regarding preferences. There are two forms of preference interdependence: (1) consumers do care about the consumption levels of others (consumption externalities), and (2) they do care about the budget possibilities of others (income externalities).

Consequently the utility of individual \( i \) will depend both on her own material consumption \( x_i \) and on the consumption choices of others \( x_{-i} \) as well as her own budget possibilities \( b_i \) compared to the budget sets of others \( b_{-i} \). Denote by \( u_i(x_i, x_{-i}, b) \) the utility of individual \( i \) from consumption profile \( x = (x_i, x_{-i}) \) and the budget profile \( b = (b_i, b_{-i}) \). Preferences are assumed to be strictly convex and strictly monotone in own consumption. Price-taking behavior in competitive markets implies that own
consumption decisions have no impact on prices and nobody is rationed at the prevailing prices. Also the consumption decisions and budget possibilities of others are taken as given when making own consumption choices. So the price-taking consumer $i$ chooses her own consumption that solves

$$\max_{x_i \in b_i} u_i(x_i, x_{-i}, b). \quad (3.13)$$

The optimal consumption of individual $i$, $x_i^*(x_{-i}, b_i, b_{-i})$, is a function of her own budget $b_i$ and the consumption and budget sets of others $(x_{-i}, b_{-i})$. If the own consumption choices are independent of the consumption choices and budget sets of others, then we say that consumers behave as if they were selfish. This is the case under the following separability assumption:

**Definition 3.1** (Dufwenberg et al. 2011) The preferences of consumers $i$ are separable if for all $x$, $x'$ and all $b$, $b'$,

$$u_i(x_i, x_{-i}, b) \geq u(x'_i, x_{-i}, b) \iff u_i(x_i, x'_{-i}, b') \geq u(x'_i, x'_{-i}, b'). \quad (3.14)$$

This separability assumption is required to make a meaningful comparison between the competitive equilibria in an economy with and without other-regarding preferences. Hence agents who care directly about the consumption of others cannot be directly distinguished from selfish agents in their consumption behavior. They look as if they are selfish even though they are not. In a Walrasian equilibrium each firm $j$ maximizes its profits $\pi_j^*$ for given price $p^*$, each consumer $i$ chooses her utility maximizing consumption bundle $x_i^*$ for given budget sets $b_i^*$, and the budget sets are compatible with equilibrium price $p^*$ in the sense that $b_i^* = \{ x_i : p^* x_i \leq p^* \omega_i + \sum_j \theta_{ij} \pi_j^*(p^*) \}$ with $\omega_i$ the initial endowment of individual $i$ and $\theta_{ij}$ the profit share of individual $i$ in firm $j$. The major theorem concerning the comparison of equilibria now follows.

**Theorem 3.2** (Dufwenberg et al. 2011) If all agents have separable preferences that are strictly monotone in own consumption, any Walrasian equilibrium of an economy with other-regarding preferences is a Walrasian equilibrium of the standard economy with selfish preferences.

This result implies that the competitive market outcome is consistent with behavior far more general than selfish optimization. There is a simple intuition for this result. If an agent’s decision does not influence the market price or the volume of trade, then he has no opportunity to change the material consumptions of others in the economy. As a
result in a competitive environment agents typically behave as if they care only about
their own material consumption, even though they have more general preferences.

There is a dual observation that is more familiar. In certain noncompetitive environ-
ments people may act as if they were altruistic, even if they care only about their own
material consumption. This is the case in many environments intended to induce coop-
eration among selfish agents (e.g., think of the Kyoto Protocol on climate change). The
fact that market equilibrium may not be affected by other-regarding preferences does
not mean that the market’s outcome will be efficient (contrarily to the standard compet-
tive economy). In general, the market outcome will be inefficient. Also this result does
not exclude the possibility that the market outcome can be affected by other-regarding
preferences when there are some forms of market failures such as those described in
part III on departures from efficiency.

3.7 Conclusions

In this chapter we did not seek to be completely comprehensive, but only to provide an
introductory account of the main themes in behavioral economics, and to explore some
of the implications for public policy. For that purpose there was no need and no attempt
to pursue the theory into every one of its corners. There was no attempt either to be
sophisticated in bringing risk or expectation into the theory, except in a very simple
way. We also only introduced dynamic considerations in a nonformal manner because
this is an issue we analyze in more depth in part VIII of the book.

To conclude this review of behavioral approach to economics, there is one puzzle
we would like to share with the reader. This puzzle is reminiscent of the one raised by
Little (1956) about the so-called welfare economics revolution. The puzzle is that the
conclusions of behavioral economics are important and influential, especially among
economists and possibly policy makers, but very few economists are clear as to what the
word “behavioral” means, or what precisely the theory is about. Indeed it is only recently
that the word “behavioral” has been employed. Time inconsistency, habit formation,
satisficing, and social interaction figured in economic analysis long before they were
swept under the umbrella of behavioral economics. It is by no means clear what this
word means (except nonstandard preferences, beliefs, or choices). It is, to put it differ-
ently, not clear what behavioral economics is about. Despite this lack of clarity, the ideas
we have discussed have influenced the opinions of many people. It obviously could not
have had any such influence if its conclusions had been meaningless, or merely formal
and recognized as such. Its conclusions certainly have some real (nonformal) meaning.
It is also fair to say that it is rather difficult to test the theory simply because there are many competing explanations for the mistakes people can make in their decisions. Since the subject matter is often something that arouses peoples’ emotions such as eating disorders, illicit drug use, excessive alcohol consumption, smoking, gambling, compulsive consumption, and excessive debt, the result seems to be a lack of balance with the conclusions of the theory being either passionately attacked or passionately defended. Only the future will tell us if the theory is likely to have much direct influence on public policy. In the meantime the theory is likely to have a considerable indirect influence by molding the opinions of economics students and, possibly as a result, some of its fashionable conclusions passing into ordinary language and being taken for granted as though they were the most obvious scientific truth. The automatic enrollment policy is a good illustration of that.

Further Reading

The starting point of behavioral economics that forcefully brings psychology into the economics of consumer choice is:


For the money pump argument see:


The self-control problem is analyzed in:


The \((\beta, \delta)\) model is successively developed in:


The framing effects are in:

Part I: Public Economics and Economic Efficiency

An overview of the ultimatum game and other bargaining predictions and experiments is in:

Conformism is in:

The influence of social norms is in:

Two excellent reviews of the central issues that arise with deviations from rationality:

Further reading on time discounting:

The first experiment of the guessing game is in:

General equilibrium treatment of the other-regarding preference is in:

Overview of reciprocity and fairness is in:

**Exercises**

3.1 Distinguish between the behavioral model of choice and the rational model of choice. Briefly describe a model of each kind.

3.2 Intertemporal models in economics typically assume exponential discounting (a constant discount rate). Do such models predict an optimal consumption plan that does not change over time or a optimal consumption plan that changes over time?
3.3 You are requested to construct a model to predict the self-control problem of procrastination to start exercising. Assume that exercising has a cost today of $-6$ and a delayed benefit of $8$. How would you model the fact that your early plan involves exercising tomorrow but not today? Show that when tomorrow arrives you will again want to postpone action. Discuss the inconsistency problem.

3.4 Let $c$ represent calories (or cigarettes), with $u$ the strictly concave (immediate) utility of consumption and $v$ the strictly convex (delayed) cost of consumption. Let the consumer have preferences described by the (intertemporal) utility function

$$U(c_{t-1}, c_t, c_{t+1}, c_{t+2}, \ldots) = \left[u(c_t) - v(c_{t-1})\right] + \beta \delta [u(c_{t+1}) - v(c_t)] + \beta \delta^2 [u(c_{t+2}) - v(c_{t+1})],$$

where $0 < \beta \leq 1$ and $0 < \delta \leq 1$.

a. Calculate the optimal level of consumption from the perspective of date $t$.

b. Assuming that $\beta = 1$, show that $c_t = c_{t+1}$.

c. Assuming that $\beta < 1$, show that $c_t > c_{t+1}$.

d. For $\beta < 1$, what is the effect of increasing the self-control parameter $\beta$ upon the consumption $c_t$? Explain your finding.

3.5 Consider a naif with $\beta = \frac{1}{2}$ and $\delta = 1$. The naif has to finish a project by deadline $T$. At date $t$, the undiscounted project costs $\left(\frac{1}{2}\right)^t$ to implement. When will the naif undertake the project?

3.6 Consider the same project as in previous exercise but now with a sophisticate. When will a sophisticate undertake the project?

3.7 “A horse! A horse! My kingdom for a horse!” (Richard III, scene iv.) Is this an example of extreme present bias?

3.8 Search and procrastination (Carroll et al. 2010). Let a “naive” consumer have discount parameters $0 < \beta < 1$ and $\delta = 1$ (daily discounting so that $\delta \leq 1$). The daily loss from delay is $L$ (the daily benefit loss). The search cost $c_t$ is stochastic and drawn from a uniform distribution on the interval $[0, 1]$. Let $W$ describe the cost function today, with

$$W(c) = \begin{cases} c & \text{if act now}, \\ \beta [L + EV(c')] & \text{if wait}, \end{cases}$$

where $V$ represents the “exponentially discounted” ($\delta = 1$) cost function tomorrow

$$V(c) = \begin{cases} c & \text{if act tomorrow}, \\ L + EV(c') & \text{if wait tomorrow}. \end{cases}$$

a. Calculate the optimal stopping rule; that is, the equilibrium cost cutoff $c^*$ such that the “naive” agent is indifferent between acting and waiting at this cost cutoff. (Hint: Solve two equations with two unknowns $c^*$ and $EV$.)

b. How does $c^*$ change with $L$?

c. How does $c^*$ change with $\beta$, the short-term discount factor?
d. Is \( c^* > 0 \)? And is \( c^* < 1 \)?
e. If \( L = \frac{1}{2} \), what is the probability of procrastination?

### 3.9
Consider the same model as in previous exercise, but with a sophisticated agent. That is, assume that the short-term discount factor is \( \beta = 1 \).

a. What is the equilibrium cost cutoff \( c^{**} \)?
b. How does \( c^{**} \) compare with \( c^* \)?
c. If \( L = \frac{1}{2} \), what is the probability of procrastination?

### 3.10
(Overconfidence). A worker chooses effort \( e \), which has cost \( c(e) \) that is increasing and convex in \( e \). The productivity of effort also depends on a variable called skill \( s \) so that output is

\[
x = f(e, s) + \theta,
\]

where \( f(e, s) \) is increasing and concave in each component, and \( \theta \) is a random term that can be called “luck.” The random term \( \theta \) is distributed according to \( \mu(\theta) \). The firm pays a linear wage based on output

\[
w(x) = w_0 + ax.
\]

Assuming separability between effort cost and wage benefit, the expected utility is

\[
EV(e) = \int_\theta U(w(f(e, s) + \theta)) \mu(\theta) d\theta - c(e).
\]

a. Suppose that skill does not matter and that \( f(e, s) = e \). Compute the optimal effort choice.
b. Now assume that the worker overestimates the marginal productivity of effort by a factor \( \beta > 0 \) so that \( \hat{f}(e, s) = (1 + \beta)e \). Show that this overconfident worker may cut back on effort.

### 3.11
Consider the previous exercise with \( f(e, s) \) increasing and concave and with a nonzero cross derivatives \( f_{e,s} = \frac{\partial^2 f(e, s)}{\partial e \partial s} \neq 0 \). \( f_{e,s} > 0 \) if \( e \) and \( s \) are complements, and \( f_{e,s} < 0 \) if \( e \) and \( s \) are substitutes.

a. Compute the optimal effort choice conditional on the skill level.
b. Now assume the worker overestimates her skill value by a factor \( \Delta \), so that \( \hat{s} = s + \Delta \). Show that whether a worker who is overconfident in this way will cut back on work (or do the opposite) depends on whether effort and skill are complementary.

### 3.12
(Beauty contest). “…professional investment may be likened to those newspaper competitions in which the competitors have to pick out the six prettiest faces from a hundred photographs, the prize being awarded to the competitor whose choice most nearly corresponds to the average preferences of the competitors as a whole; so that each competitor has to pick, not those faces which he himself finds prettiest, but those which he thinks likeliest to catch the fancy of the other competitors, all of whom are looking at the problem from the same point of view. It is not a case of choosing those which, to the best of one’s judgment, are really the prettiest, nor even those which average opinion genuinely thinks the prettiest. We have reached the third degree where we devote our intelligences to anticipating what
average opinion expects the average opinion to be. And there are some, I believe, who practice the fourth, fifth and higher degrees.’” (John Maynard Keynes, *The General Theory of Employment, Interest, and Money*).

In reference to the quote above, consider the following guessing game. There are 100 participants, and each participant announces independently and simultaneously a number between 0 and 100 (only integers are allowed). There is a fee of $1 to participate to the guessing game, to be redistributed as a prize of $100 to the winner. The participant who guessed closest to a target wins the prize.

a. Suppose that the target is one-half of the average guess. Would you participate in this game? Explain why.

b. Suppose that the target is one-half of the average guess. What would be your best guess? Explain briefly.

c. What would we expect if everyone is rational and thinks everyone else is rational? Compare to your answer in part b.

3.13 (Ultimatum game). Consider a fixed income of \( y \) to share between two players. Player 1 is the proposer and demands \( x_1 \) for himself. Player 2 who is the responder accepts or rejects the demand. If the demand is accepted, the payoffs are \( v_1 = x_1 \) and player 2’s payoff is \( v_2 = y - x_1 \).

a. What is the subgame-perfect Nash equilibrium if both players are selfish?

b. What would happen if both players are fair?

3.14 The utility function of individual \( i \) is given by \( u_i(y_i, y_j) = \log y_i + \lambda \log y_j \), where \( y_i \) is individual \( i \)’s income, \( y_j \) is individual \( j \)’s income, and \( \lambda \) is a parameter representing the other-regarding preference.

a. For what values of \( \lambda \) is individual \( i \) selfish?

b. For what values of \( \lambda \) is individual \( i \) an altruist?

c. For what values of \( \lambda \) is individual \( i \) fair?

d. For what values of \( \lambda \) is individual \( i \) envious?

3.15 (Envy model). Consider a 2-player model of envy with the utility function of individual \( i \), given by \( u_i(y_i, y_j) \), where \( \frac{\partial u_i}{\partial y_i} > 0 \) and \( \frac{\partial u_i}{\partial y_j} < 0 \). Show that any division of an object of fixed size in which the responder gets nothing will be necessarily rejected by the responder.

3.16 Consider a two-stage version of the ultimatum game with a shrinking amount to be distributed among two selfish players. The amount to be distributed in stage 1 is 1. Player 1 demands \( x_1 \) and player 2 accepts or rejects the demand. If player 2 accepts, he receives \( 1 - x_1 \). If player 2 rejects, he can make a counterdemand \( x_2 \) and player 1 can accept or reject, but then the amount to be distributed is reduced to \( \lambda < 1 \). If player 1 accepts the counterdemand \( x_2 \), he receives \( \lambda - x_2 \). If player 1 rejects the counterdemand, both players receive 0.

a. Compute the subgame-perfect Nash equilibrium.

b. Show that the equilibrium demand by player 1 is decreasing in \( \lambda \). Explain.

3.17 Repeat the previous two-stage ultimatum game with a shrinking amount to be distributed, but now assume a linear model of envy with utility of player \( i \).
Part I: Public Economics and Economic Efficiency

\[ u_i(y_i, y_j) = y_i - \alpha y_j, \]

with the envy preference parameter \( \alpha > 0 \).

a. Compute the subgame-perfect Nash equilibrium

b. Show that the equilibrium demand by player 1 is decreasing in \( \lambda \) if \( \alpha < \lambda \). Explain.

c. Show that an unfavorable counteroffer is possible; that is, in equilibrium \( x_2 < 1 - x_1 \). Explain.

3.18 What is the status of behavioral models?
II

GOVERNMENT
4.1 Introduction

In 1913 the Sixteenth Amendment to the US Constitution gave Congress the legal authority to tax income. In so doing, it made income taxation a permanent feature of the US tax system and provided a significant source of additional tax revenues. Revenue collection passed the $1 billion mark in 1918, increased to $5.4 billion by 1920, and reached $43 billion in 1945. It was not until the tax cut of 1981 that this process of growth showed any marked sign of slowing. This growth in tax revenue was matched by an equal growth in government expenditure. The US experience is typical of similar developments in all industrialized economies.

This chapter provides a statistical overview of the public sector in modern market economies. Data are presented on government expenditure and revenue. The purpose is to give both a historical perspective and an insight into the current situation.

From the numerous items of expenditure and sources of revenue, we can observe the extent and range of activities in which the public sector is involved. A surprising feature that the data reveals is the similarity in public sector behavior in countries that are otherwise very diverse culturally. Specifically, the difference in the size of the public sector between the social-market economies of northern Europe and the free-market economies of North America and Asia is rather less than might be imagined.

4.2 Historical Development

The historical development of the public sector over the past century can be summarized as one of significant growth. For the typical industrially developed economy, government expenditure was only a small proportion of the gross domestic product (GDP) at the start of the twentieth century. Expenditure then rose steadily over the next sixty years, leveling out toward the end of the century. The details behind this broad-brush description are illustrated in the figures that follow.

Figure 4.1 shows the growth of public spending during the last century for five developed economies. This depicts expenditure as a percentage of gross domestic product to give an idea of the size of the public sector relative to the economy as a whole. Only a selection of years is plotted, but the figure provides a clear impression of the overall trend. Although there is a persistent difference in the levels of expenditure
among the three European countries (France, Germany, and United Kingdom) and the non-European countries (Japan and United States) the pattern of growth is the same for all. These five economies had a clear long-run upward path in public spending relative to the gross domestic product. Starting with a level of public spending around 10 percent of the gross domestic product in 1870, this increased markedly around 1910 and then continued to rise afterward. In 1996 the United States had the lowest public spending level of the five countries at 32.4 percent, but even this is one-third of the gross domestic product. France had the highest level at 55 percent. A number of explanations for this long-run increase have been proposed. These explanations are presented in chapter 5.

A more detailed view of the changes in the level of expenditure over the last thirty years is provided in figure 4.2. The picture displayed is of a slowing, or even a stagnation, of the growth in public sector expenditure, particularly over the past twenty years. Although expenditure was higher in 2007 than in 1970 for the six countries shown, the increases for the United Kingdom and the United States were very small (from 38.8 to 43.9 percent for the United Kingdom and from 31.7 to 36.9 percent for the United States). For the United Kingdom especially, expenditure was clearly higher in the early 1980s (peaking at 47.5 percent in 1981) than in 2007. This picture was changed significantly by the financial crisis. Expenditure rose sharply for all the countries between 2008 and 2009 as the governments increased spending to reflate their economies. There has been a small fall in expenditure since the peak in 2009, but it remains at a level about that of 2007. Overall, the figure suggests that there has been convergence in
the level of expenditure between the countries. For example, in 1970, expenditure in Japan was approximately half that in France, Germany, and the United Kingdom, but by 2002, it had reached 38.8 percent in Japan and almost matched that in the United Kingdom. The gap has widened again since (to 6.3 percentage points), as the United Kingdom has been affected more than Japan by the financial crisis.

Figure 4.3 shows the path of expenditure in selected subcategories of public spending during the last century, again expressed as a percentage of gross domestic product. This breakdown into categories is helpful in understanding the composition of the long-run increase in figure 4.1. Defense spending constituted one of the largest items of public spending in the late nineteenth century. It has since been somewhat erratic and driven in large part by the history of international relations. In all cases defense spending peaked at midcentury and has fallen continually since. In 1996 the United States spent the largest proportion its gross domestic product on defense (4 percent).

The most marked rises have come from social spending on items such as education, health, and pensions. Expenditure on education and pensions has risen sharply as a share of the gross domestic product in all five countries since the early twentieth century but particularly so since midcentury (and perhaps slightly earlier in the United Kingdom). In all five countries it is currently around 5 percent of the gross domestic product. Health expenditure has risen more rapidly. Even in the United States, which has a primarily private health care system, the public sector expenditure on health was 6.3 percent of the gross domestic product in 1994. The significant increase in expenditure on pensions is
important from a policy perspective. As discussed further in chapter 23, many countries are facing a “pensions crisis” in which the current rate of expenditure on state pensions is unsustainable. The basis of this is clearly apparent in the rate of expenditure increase in France and Germany.

Data on public sector expenditure for a wider range of countries in 2010 is given in figure 4.4. This includes developed, developing, and transition economies. The figure clearly justifies the claim that the public sector is significant in countries across the world. France has the highest level of public sector expenditure (at 56.6 percent) and
Figure 4.3 (Continued)

Korea the lowest (at 21 percent). All have “mixed economies” characterized by substantial government involvement. They are clearly not free-market economies with minimal government intervention. These values for the size of the public sector emphasize the importance of studying how government should best choose its means of revenue collection and its allocation of expenditure.

As a final point, it is worth noting that data on expenditures typically understate the full influence of the public sector on the economy. For instance, regulations such as employment laws and safety standards affect economic activity but do not directly
generate any measurable government expenditure or income. Analysis of statistics on government expenditures do not therefore capture the effects of such policies. This point is explored further in section 4.6.

4.3 Composition of Expenditure

The historical data display the broad trend in public expenditure. This section looks in more detail at the composition of expenditure. Expenditure is considered from the perspective of its allocation between various levels of government and its division into categories.

Figure 4.5 allocates expenditures among the different levels of government (net of all transfers between levels). The significant difference between the United Kingdom, which has no expenditures at the state level, and Germany and the United States is explained by their political structures. Germany and the United States are federal countries that have central government, state government, and local government. In contrast, the United Kingdom is a unitary country that has only central and local governments. The figures reveal that expenditure at the state level is similar in Germany and the United States (20 and 22 percent respectively), although local government is
larger in the United States (26 percent compared to 15 percent). Despite the different political structure in the United Kingdom, the proportion of expenditure at the local level is identical to that in the United States (26 percent). By definition, central expenditure in the United Kingdom (73 percent) is then equal to the proportions of central plus state in the United States.

Figure 4.6 displays the different compositions of general spending in the United States, United Kingdom, and Germany. The diversity of public sector activity is clear from the list of spending categories. Interestingly, spending on the goods associated with the core functions of the state—defense and public order—is relatively minor and forms about 10 percent of spending when averaged across the countries. Administrative and governmental costs are recorded under the heading of general public services and add no more than another 12 percent on average.

Health and education, despite providing benefits of an arguably largely private nature, are substantial in all three countries (e.g., education is 17 percent and health 21 percent
Figure 4.6
Composition of general spending, 2007
in the United States). Spending on housing and community amenities, on recreation and culture, and on transport and communications sectors are comparatively small. Subsidies to the agriculture, energy, mining, manufacturing, and construction sectors are brought together here under the heading of other economic affairs and also appear relatively minor. Social security and welfare spending is the largest single item in all countries under this classification. This is so even in the United States where, at 19 percent, it is noticeably smaller than in Germany and the United Kingdom (44 and 35 percent respectively). Averaged across the three countries it constitutes over a third of spending.

Figures 4.7 to 4.9 show how spending responsibilities are allocated among different tiers of government in the United States, United Kingdom, and Germany. This provides an interesting contrast between the two federal countries (Germany and the United States) as compared with the unitary country (United Kingdom). Yet, even though the political structures are significantly different, some common features can be observed. Certain items such as defense are always allocated to the center. Redistributive functions also tend to be concentrated centrally, for the good reason that redistribution between poor and rich regions is only possible in that way and also because attempts at redistribution at lower levels are vulnerable to frustration through migration of richer individuals away from localities with internally redistributive programs. Education, in contrast, is largely devolved to lower levels, either to the states or to local government. Public order is also typically dealt with at lower levels. Health spending is always substantial at the central level but can also be important at lower tiers, as in Germany.

The fact that spending is made at a lower level need not mean that it is financed from taxes levied locally. In most multiple-tier systems the central government partly finances lower tier functions by means of grants. These have many purposes, including correcting for imbalances in resources among localities and among tiers given the chosen allocation of tax instruments. Sometimes grants are lump sum, and sometimes they depend on the spending activities of the lower tiers. In the latter case the incentives of lower tiers to spend can be changed by the design of the grant formula, and central government can use this as a way to encourage recognition of externalities between localities.

4.4 Revenue

The discussion of public sector expenditure is now matched by a discussion of revenue. The following figures first trace the historical path of tax revenues and then relate revenues to different tax instruments and to alternative levels of government.
Figure 4.7
Composition of central spending
Figure 4.8
Composition of state spending

The first set of statistics consider the growth of total tax revenue from 1965 to 2009. Figure 4.10 charts total tax revenue for seven countries expressed as a percentage of the gross domestic product. The general picture that emerges from this mirrors that drawn from the expenditure data. All of the countries have witnessed some growth in tax revenue over the period, and there has also been a degree of convergence. The financial crisis and resulting recession have caused tax revenues to fall from 2007 to 2009. In 2009 government revenue in these countries ranged between 24 and 42 percent of the gross domestic product.

Looking more closely at the details, France (42 percent) and the United Kingdom (34 percent) have the highest percentage, closely followed by Canada (32 percent). Japan
(27 percent), Turkey (25 percent), and the United States (24 percent) are somewhat lower. The country with the most growth in tax revenue is Turkey, with revenue rising from 11 percent of the gross domestic product in 1965 to 33 percent in 2000 (but declining since then). Tax revenue also grew strongly in Japan between 1965, when it was 11 percent of the gross domestic product, and 1990, when it reached 30 percent, but has leveled off since. Overall, the data show surprising uniformity among these countries with all achieving a similar outcome. The figures that follow present the details behind these aggregates.

Figure 4.11 shows at the proportion of tax revenue raised by six categories of the tax instrument in 2009. Note that income and profit taxes raise the largest proportion of
Chapter 4: Public Sector Statistics

**Figure 4.10**
Tax revenues, 1965 to 2009 (% of GDP)

**Figure 4.11**
Tax revenue for category of taxation, 2009
revenue in Australia (56 percent), Canada (47 percent), the United States (41 percent), and the United Kingdom (39 percent). Social security taxes are the largest proportion in Japan (40 percent), France (39 percent), and Germany (39 percent). Taxes on goods and services are the most significant item in Turkey (46 percent) and Korea (32 percent). There is also noticeable division between the European countries, where taxes on goods and services are much more significant, and the United States. For instance, taxes on goods and services raise 29 percent of revenue in the United Kingdom, but only 18 percent in the United States. This is a reflection of the importance of value-added taxation (VAT) in Europe where it has been a significant element of EU tax policy. Property taxes are significant in the majority of countries (14 percent in the United States, 12 percent in the United Kingdom, and 12 percent in Korea). Payroll taxes are only really significant in Australia (5 percent).

The next two figures display the proportion of tax revenue raised by each level of government. Figure 4.12 shows the proportions in five federal countries. In contrast, figure 4.13 shows the five unitary countries. For all the federal countries the central government raises more revenue than state government. The two are closest in Canada, with the central government raising 42 percent and the provinces 39 percent, and in Germany, with the central government raising 31 percent and the provinces (Bundesländer) 22 percent. The federal governments in Australia and the United States raise considerably more revenue than the states (80 and 16 percent for Australia and 34 and

Figure 4.12
Tax revenue by level of government, federal countries, 2009
21 percent for the United States). In all countries local government raises the smallest proportion of revenue. The US local government raises 17 percent of revenue, which is the largest value among these countries. The smallest proportion of revenue raised by local government is 3.5 percent in Australia.

The unitary countries in figure 4.13 display the same general feature: that the central government raises significantly more revenue than local government. The largest value is 67 percent in Turkey and the smallest 31 percent in France. Local government is most significant in Japan (28 percent) and least significant in Turkey (9 percent).

From comparing the federal and unitary countries, it can be seen that local government raises slightly more revenue on average in the unitary countries than the federal countries. What really distinguishes them is the size of central government. The figures suggest that the revenue raised by central government in the unitary countries is almost the same on average as that of central plus state in the federal countries. The absence of state government does not therefore put more emphasis on local government in the unitary countries. Instead, the role of the state government is absorbed within central government.

The final set of figures presents the share of revenue raised by each category of tax instrument at each level of government for two federal countries, the United States and Germany, and two unitary countries, Japan and the United Kingdom. Most of the previous figures have shown remarkable similarities in the behavior of a range of
countries. In contrast, allocating revenues to tax instruments for the alternative levels of government reveals some interesting differences.

For the United States figure 4.14 shows that the importance of income and profits taxes falls as the progression is made from central to local government (90 percent for central, 5 percent for local). Their reduction is matched by an increase in the importance of property taxes from 2 percent for central government up to 73 percent for local government. It would be easy to argue that this is the natural outcome since property is easily identified with a local area but income is not. However, figure 4.15 for Germany shows that the opposite pattern can also arise with income and profit taxes becoming more important for local government (78 percent of revenue) than for central government (38 percent of revenue). Despite this difference Germany and the United States do share the common feature that property taxes are more important for local government than for central government.

The same data are now considered for two unitary countries. In Japan (figure 4.16) income and profits taxes are almost equally important for both central government (50 percent of revenue) and local government (51 percent). They are also more important for both levels of government than any other category of tax instrument. Where the difference arises is that property taxation is much more significant for local government (raising 30 percent of revenue) than for central (6 percent). For central government, general taxes (24 percent of revenue) make up the difference. The UK data, in figure 4.17,
Chapter 4: Public Sector Statistics

Figure 4.15
Tax shares at each level of government, Germany, 2009

Figure 4.16
Tax shares at each level of government, Japan, 2009
present an extreme version of the importance of property taxation for local government. As the figure shows, property taxes raised 100 percent of tax revenue for local government. No revenue is raised by local government in the United Kingdom from income and profit taxes.

As compared, the data for the unitary and federal countries do not reveal any standard pattern of revenues within each group. In fact the differences are as marked within the categories as they are across the categories. The one feature that is true for all four countries is that property taxes raise a larger proportion of revenue for local government than they do for central government.

This section has looked at data on tax revenues from an aggregate level down to the revenue raised from each category of tax instrument for different levels of government. What the figures show is that at an aggregate level there are limited differences among the countries. Those for which data are reported have converged on a mixed-economy solution with tax revenues at a similar percentage of the gross domestic product. The most significant differences emerge when the source of revenue for the various levels of government is analyzed. Even countries that have adopted the same form of government structure (either unitary or federal) can have very different proportions of revenue raised by the various categories of tax instrument.
4.5 Government Debt

When the level of government expenditure exceeds the value of revenue, the budget is in deficit. Deficits are financed by issuing debt, so the stock of debt represents an accumulation of past deficits. Debt can be viewed as the substitution of current government expenditure for future expenditure. It imposes a cost (the service and repayment of the debt) on future taxpayers in order to provide expenditure for current taxpayers. The level of government debt has become an issue of major policy importance due to the increasing debt levels in many countries. The increase in debt has occurred for two reasons. First, many governments have responded to the global financial crisis by boosting expenditure at a time of falling revenues. Second, the operation of the euro has provided poor incentives for fiscal prudence in the European Union countries that adopted the single currency.

The debt levels of governments are usually sustainable—in the sense that promised payments on the debt can be made—but there are occasions when governments cannot make the promised payments. In such cases the government defaults on the promises to pay. When a corporation defaults on debt, it is put into liquidation by the holders of the debt and the value of its assets is distributed among them. In contrast, a default by a government (a sovereign debt default) cannot be met with liquidation of the country. The only consequence for a country is future difficulty in borrowing, and the higher rates that will have to be paid on future debt to compensate for the perceived increase in risk. If a country issues debt in its own currency, then it need not default (although some countries, such as France in the sixteenth century, have chosen to repudiate debt). Instead, a country has the option of simply printing more money to repay the debt but then has to live with the inflationary consequences. The more significant problems arise when a country either issues debt denominated in a foreign currency (e.g., a South American country issuing bonds denominated in dollars) or, as in the unique case of the euro, issues debt in its own currency but cannot unilaterally print more money.

The roots of the global financial crisis are usually traced to the bursting of a bubble in the United States housing market after it peaked in 2007. The bubble had been financed by extensive mortgage borrowing at cheap rates. These mortgages were securitized (bundled together and sold in packages) and then traded as financial securities (securitized mortgage obligations) in their own right. As mortgage borrowers began to default on their loans, the securitized mortgage obligations became risky and collapsed in value. This undermined the financial position of the holders of these securities. The consequence was acute liquidity problems in the US banking system in 2008, and the failure
of major financial institutions in the United States and elsewhere. The level of credit available declined, causing a slowdown in real economic activity and the recession of 2009. The response of many governments was to boost their economies through fiscal stimulus and monetary expansion. Significant expenditure was also made to provide support and bailouts for financial institutions. These actions lead to rising government debt.

The story for the euro countries is partly related to the debt crisis and partly to the operational structure of the euro. Ireland, in particular, has incurred debts because of the financial crisis. The Irish economy had its own boom and became over-inflated through the 1990s. The onset of recession elsewhere in the world in 2008 hit the Irish economy hard. An alternative story is appropriate for Greece. After joining the euro Greece enjoyed much lower interest rates than it had previously faced. It chose to borrow heavily at these new low rates to finance higher levels of current consumption. This reflected a combination of a simple change in the intertemporal trade-off and the intention of the incumbent government to secure short-term popularity.

The time path of the budget deficit for several countries is shown in figure 4.18 for 1996 to 2009. A noticeable feature of the data is that the deficits in Germany and Italy do not show any marked variation over the time interval. The effect of the financial

Figure 4.18
Budget deficit (% of GDP)
crisis can be seen from 2008 but the effect does not push the deficits significantly below the long-term trend. In contrast, the other countries show strongly widening deficits from 2006. The largest increases were in Greece and Ireland, which explains why these countries have been at the center of the crisis in the euro area. The United Kingdom and the United States follow fairly similar paths that reflect the use of fiscal stimulus policies in both countries.

High and rising debt is a source of justifiable concern. It is not only government debt that has grown but also corporate and household debt. Over the past thirty years the ratio of total debt to gross domestic product in advanced economies has risen from 167 percent in 1980 to 314 percent in 2011. This level of debt is unprecedented. Figures 4.19 to 4.21 show the growth in this debt for a range of countries. With the exception of Japan, corporate debt has shown a general tendency to rise, and it has risen sharply in most countries from 2000 to 2010. The growth of government debt is particularly marked in Japan and Greece, but all the countries shown had higher levels of debt in 2010 than in 1980. The pattern for household debt is not so uniform. Italy, France, and Spain have had significant increases in household debt, but in Italy and France this has been from a very low level in 1980. In contrast, household debt fell between 2000 and 2010 in Germany and Japan.

Figure 4.19
Corporate debt of nonfinancial corporations (% of GDP)
Figure 4.20
Government debt (% of GDP)

Figure 4.21
Household debt (% of GDP)
The level of debt accumulated by some euro area countries has led to a crisis for the currency. The crisis has involved several countries (Greece, Ireland, and Italy being the notable examples) coming close to defaulting on their debts and requiring financial bailouts. The increase in aggregate debt of the 17 countries that have adopted the euro is shown in figure 4.22. The level was steady at around 70 percent of the gross domestic product until the financial crisis of 2008; then it rose sharply to 90 percent in 2011. This level of indebtedness has arisen despite the countries being notionally subject to the conditions of the Stability and Growth Pact, which requires the government deficit to be less than 3 percent of the gross domestic product and debt to be less than 60 percent of the gross domestic product. It is clear that these conditions were simply not enforced because of the lack of any credible enforcement mechanism.

Details of the debt levels in individual euro area countries is reported in figure 4.23. Over the period in which the euro had been in existence, the debt to the gross domestic product ratios have increased significantly (with the exception of Spain). Ireland had the greatest increase relative to the gross domestic product, which is why it required a financial bailout in 2010. Greece has the highest debt ratio in 2010 at 152.6 percent of the gross domestic product, followed by Italy at 119.5 percent of the gross domestic product. Both countries came close to defaulting on debt in 2011. The lowest levels of
The past two decades witnessed a massive accumulation of debt across the world. Corporations, governments, and households have contributed to this process. The figures have revealed how the level of debt was increased by the global financial crisis. Debt places a burden on the public finances of a country. Countries can usually print money to avoid default, but this is not an option for euro area countries where debt is causing some of them very severe problems.

4.6 Measuring the Government

The figures examined above have provided several different viewpoints on the public sector. They have traced both the division of expenditure and the level of expenditure. For the purpose of obtaining a broad picture of the public sector, these are interesting and informative statistics. However, they do raise two important questions that must be addressed in order to gain a proper perspective on their meaning.
The first issue revolves around the fact that the figures have expressed the size of the public sector relative to the size of the economy as a whole. To trace the implications of this, take as given that there exists an accurate measure of the expenditure level of the public sector. The basic question is then: What should this expenditure be expressed as a proportion of? The standard approach is to use the nominal gross domestic product (i.e., gross domestic product measured using each year’s own prices), but this is very much an arbitrary choice that can have a significant impact on the interpretation of the final expenditure figure.

Recall from basic national income accounting that the size of the economy can be measured in either nominal or real terms, using gross output or net output. Domestic or national product can be employed. Outputs can be valued at market prices or factor prices. For many purposes, as long as the basis of measurement is made clear, the choice of measure does not make much real difference. Where it can make a critical difference is in the impression it gives about the size of the public sector. By adopting the smallest measure of the size of the economy (which depends on a number of factors, e.g., the level of new investment relative to depreciation, the structure of the tax system, and income from abroad), the apparent size of the public sector can be increased by several percentage points over that when using the largest expenditure level.

While not changing anything of real economic significance, such manipulation of the figures can be very valuable in political debate. There is a degree of freedom for those who are supportive of the public sector, or are opponents of it, to present a figure that is more favorable for their purposes. This may be useful for those wishing to push a particular point of view, but it hinders informed discussion. Consequently, as long as the figures are calculated in a consistent way, it does not matter for comparative purposes which precise definition of output is used. In contrast, for an assessment of whether the public sector is “too large,” it can matter significantly.

The second issue of measurement concerns what should be included within the definition of government. To see what is involved here, consider the question of whether state-run industries should be included. Assume that these are allowed to function as if they were private firms, so that they follow the objective of profit maximization and simply remit their profits to the government. In this case they should certainly not be included, since the government is acting as if it were a private shareholder. The only difference between the state-run firm and any other private firm in which the government is a shareholder would be the extent of the shareholding. Conversely, assume that the state-run firm was directed by the government to follow a policy of investment in impoverished areas and to use cross-subsidization to lower the prices of some of its
products. In this case there are compelling reasons to include the activities of the firm within the measure of government.

What this example illustrates is that it is not government expenditure per se that is interesting to the economist. What is instead really relevant is the degree of influence the government has over the economy. When the government is simply a shareholder, it is not directly influencing the firm’s decisions. The converse is true when it directs the firm’s actions. Looked at in this way, measuring the size of government via its expenditure is a means of estimating government influence using an easily observable statistic. In fact the extent of government influence is somewhat broader than just its expenditure. What must also be included are the economic consequences of government-backed regulations and restrictions on economic behavior. Minimum wage laws, weights and measures regulation, health and safety laws, are all examples of government intervention in the economy. However, none of these would feature in any observation of government expenditure.

What this discussion shows is that there is a degree of flexibility in interpreting measures of government expenditure. Furthermore government influence on the economy is only approximately captured by the expenditure figure. The true extent, including all relevant laws and regulations, is most certainly much larger.

### 4.7 Conclusions

This chapter has reviewed the expenditures and revenues of the public sector using data from a range of countries. Despite their clear cultural differences the countries considered have all experienced the same phenomenon of significant public sector growth in the last century. From being only a minor part of the economy at the start of the last century, the public sector had grown to be significant proportion of the gross domestic product in all developed countries by the end of the century. There is some variation within the figures for the precise level of public expenditure, but the pattern of growth is the same for all. There is also evidence that the growth has now ceased, and unless there is some major upheaval, the size of the public sector will now remain fairly constant.

In terms of the composition of public sector revenue and expenditure, it can be noted that there are differences in the details among countries. However, there is common reliance on similar tax instruments. Spending patterns are also not too dissimilar. It is
these commonalities that make the ideas and concepts of public economics so broadly applicable.

Further Reading

Detailed evaluations of the different areas of public expenditure can be found in:


The data for figures 4.1 and 4.3 are taken from:


Figure 4.2 is compiled using data from:

OECD *Economic Outlook*, vols. 51 and 73.

The expenditure data in figures 4.5 to 4.9 are from:


Data on revenues in figures 4.10 to 4.17 are from:


The source of the data on debt in figures 4.19 to 4.21 is


Exercises

4.1 What factors might have been responsible for the growth of government expenditure between 1920 and 1940?

4.2 Obtain data on public sector expenditure and estimate the growth trend:
   a. Over the last 50 years.
   b. Over the last 20 years.

   Has there been a structural break (a point at which the rate of growth distinctly changed)?

4.3 Why may expenditure data underestimate the influence of the public sector on the economy?

4.4 Does recent experience suggest that the growth of expenditure has now ceased?
4.5 In the 1980s both the United Kingdom and the United States had governments that aimed to cut expenditure and reduce the role of government. Did they succeed? Could any government now cut expenditure?

4.6 Is expenditure to combat market failure greater than expenditure for redistributive purposes?

4.7 What is the “pensions crisis”? How can this be solved?

4.8 Comparing figure 4.2 to figure 4.10 shows that taxation is a smaller proportion of the gross domestic product than expenditure. How can this be so?

4.9 Why is income taxed rather than wealth?

4.10 What explains the limited revenue from property taxation?

4.11 Should social security taxes be viewed as a second component of income taxation?

4.12 (Malcolmson 1986) It is natural to wonder whether there is a limit to the size of the public sector. One way to think about this is to consider the Laffer curve. This curve graphs the amount of government revenue as a function of the tax rate. A limit to the size of government will be reflected in this curve first increasing, reaching a maximum, and then decreasing. The maximum represents the largest size of government that can be sustained.

Consider a consumer with preferences \( U = \ln(x) + \ln(1 - \ell) \), where \( x \) is consumption and \( \ell \) is labor supply. Let the budget constraint be \( x = (1 - t)\ell \) where \( t \) is the tax rate.

a. Find the labor supply of the consumer by eliminating \( x \) from the utility function.

b. Use the solution for labor supply to calculate tax revenue.

c. Plot revenue as a function of the tax rate. Comment on the form of the function.

4.13 Explain why defense spending is organized centrally and education locally.

4.14 Is there any logic to the division of spending responsibilities between different levels of government?

4.15 Does the division of political responsibility among different levels of government have any economic implications?

4.16 Provide an interpretation of the EU structure from the perspective of the division of tax collection.

4.17 Should the collection of taxes on consumption (VAT in the European Union, sales tax in the United States) be centralized?

4.18 How could a minimum wage law be evaluated as government intervention?

4.19 Do increases in public expenditure cause an increase in national income, or vice versa? How would you test which is the case?

4.20 The value of the gross domestic product for several measures is given in the table below. If public expenditure is $10 billion, what are the largest and smallest proportional measures of the public sector? Does the difference matter?
### Chapter 4: Public Sector Statistics

<table>
<thead>
<tr>
<th>Measure</th>
<th>Factor prices</th>
<th>Market prices</th>
<th>Domestic product</th>
<th>National product</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value ($billion)</td>
<td>30.2</td>
<td>32.3</td>
<td>31.2</td>
<td>31.5</td>
</tr>
</tbody>
</table>

There is good evidence that much economic activity is unrecorded in official statistics. This “hidden economy” includes legal activities that are not reported to the authorities in order for tax to be evaded and illegal activities. Should official statistics ignore the hidden economy (which is the current practice) or make an effort to incorporate an estimated value in national accounts?
5 Theories of the Public Sector

5.1 Introduction

The statistics of chapter 4 have described the size, growth, and composition of the public sector in a range of developed and developing countries. The data illustrated that the pattern of growth was similar across countries, as was the composition of expenditure. Although there is some divergence in the size of the public sector, it is significant in all the countries. Such observations raise two interrelated questions. First, why is there a public sector at all—would it not be possible for economic activity to function satisfactorily without government intervention? Second, is it possible to provide a theory that explains the increase in size of the public sector and the composition of expenditure? The purpose of this chapter is to consider possible answers to these questions.

The chapter begins with a discussion of the justifications that have been proposed for the public sector. These show how the requirements of efficiency and equity lead to a range of motives for public sector intervention in the economy. Alternative explanations for the growth in the size of the public sector are then assessed. As a by-product, they also provide an explanation for the composition of expenditure. Finally, some economists would argue that the public sector is excessively large. Several arguments for why this may be so are considered.

5.2 Justification for the Public Sector

Two basic lines of argument can be advanced to justify the role of the public sector. These can be grouped under the headings of *efficiency* and *equity*. Efficiency relates to arguments concerning the aggregate level of economic activity, whereas equity refers to the distribution of economic benefits. In considering these arguments, it is natural to begin with efficiency because this is the more fundamental concept.

5.2.1 The Minimal State

The basic motivation for the existence of a public sector follows from the observation that entirely unregulated economic activity cannot operate in a sophisticated way. In short, an economy would not function effectively if there were no *property rights* (the
rules defining the ownership of property) or contract laws (the rules governing the conduct of trade).

Without property rights, satisfactory exchange of commodities could not take place given the lack of trust that would exist between contracting parties. This argument can be traced back to Thomas Hobbes, who viewed the government as a social contract that enables people to escape from the anarchic “state of nature” where their competition in pursuit of self-interest would lead to a destructive “war of all against all.” The institution of property rights is a first step away from this anarchy. In the absence of property rights, it would not be possible to enforce any prohibition against theft. Theft discourages enterprise, since the gains accrued may be appropriated by others. It also results in the use of resources in the unproductive business of theft prevention.

Contract laws determine the rules of exchange. They exist to ensure that the participants in a trade receive what they expect from that trade or, if they do not, have open an avenue to seek compensation. Examples of contract laws include the formalization of weights and measures and the obligation to offer product warranties. These laws encourage trade by removing some of the uncertainty in transactions.

The establishment of property rights and contract laws is not sufficient in itself. Unless they can be policed and upheld in law, they are of limited consequence. Such law enforcement cannot be provided free of cost. Enforcement officers must be employed and courts must be provided in which redress can be sought. In addition an advanced society also faces a need for the enforcement of more general criminal laws. Moving beyond this, once a country develops its economic activity, it will need to defend its gains from being stolen by outsiders. This implies the provision of defense for the nation. As the statistics made clear, national defense has at times been a very costly activity.

Consequently, even if only the minimal requirements of the enforcement of contract and criminal laws and the provision of defense are met, a source of income must be found to pay for them. This need for income requires the collection of revenue, whether these services are provided by the state or by private sector organizations. But they are needed in any economy that wishes to develop beyond the most rudimentary level. Whether it is most efficient for a central government to collect the revenue and provide the services could be debated. Since there are some good reasons for assuming that this is the case, the coordination of the collection of revenue and the provision of services to ensure the attainment of efficient functioning of economic activity provides a natural role for a public sector.

This reasoning illustrates that to achieve even a most minimal level of economic organization, some unavoidable revenue requirements are generated and require financing. From this follows the first role of the public sector, which is to assist with
the attainment of economic efficiency by providing an environment in which trade can flourish. The minimal state provides contract law, polices it, and defends the economy against outsiders. The minimal state does nothing more than this, but without it, organized economic activity could not take place. These arguments provide a justification for at least a minimal state and hence the existence of a public sector and of public expenditure.

Having concluded that the effective organization of economic activity generates a need for public expenditure, one role for public economics is to determine how this revenue should be collected. The collection should be done with as little cost as possible imposed on the economy. Such costs arise from the distortion in choice that arise from taxation. Public economics aims to understand these distortions and to describe the methods of minimizing their impact.

5.2.2 Market versus Government

Moving beyond the basic requirements for organized economic activity, there are other situations where intervention in the economy can potentially increase welfare. However, unlike the minimal provision and revenue requirements, there will always be a degree of contentiousness about additional intervention whatever the ground on which it is motivated. The situations where intervention may be warranted can be divided into two categories: those that involve market failure and those that do not.

When market failure is present, the argument for considering whether intervention would be beneficial is compelling. For example, if economic activity generated externalities (effects that one economic agent imposes on another without their consent), so that there is divergence between private and social valuations and the competitive outcome is not efficient, it may be felt necessary for the state to intervene to limit the inefficiency that results. This latter point can also be extended to other cases of market failure, such as those connected to the existence of public goods and of imperfect competition. Reacting to such market failures is intervention motivated on efficiency grounds.

It must be stressed that this reasoning does not imply that intervention will always be beneficial. In every case it must be demonstrated that the public sector has the ability to improve on what the unregulated economy can achieve. This will not be possible if the choice of policy tools is limited or government information is restricted. It will also be undesirable if the government is not benevolent. These various imperfections in public intervention will be a recurrent theme of this book.

While some useful insights follow from the assumption of an omnipotent, omniscient, and benevolent policy maker, in reality it can give us very misleading ideas about the
possibilities of beneficial policy intervention. It must be recognized that the actions of the state, and the feasible policies that it can choose, are often restricted by the same features of the economy that make the market outcome inefficient. One role for public economics is therefore to determine the desirable extent of the public sector or the boundaries of state intervention. For instance, if we know that markets will fail to be efficient in the presence of imperfect information, then to establish the merit of government intervention, it is crucial to know if a government subject to the same informational limitations can achieve a better outcome.

Furthermore a government managed by nonbenevolent officials and subject to political constraints may fail to correct market failures and may instead introduce new costs of its own creation. It is important to recognize that this potential for government failure is as important as market failure and that both are often rooted in the same informational problems. At a very basic level the force of coercion must underlie every government intervention in the economy. All policy acts take place, and in particular, taxes are collected and industry is regulated, with this force in the background. But the very power to coerce raises the possibility of its misuse. Although the intention in creating this power is that its force should serve the general interest, nothing can guarantee that once public officials are given this monopoly of force, they will not try to abuse this power in their own interest.

5.2.3 Equity

In addition to market failure, government intervention can be motivated by the observation that the economy may have widespread inequality of income, opportunity, or wealth. This can occur even if the economy is efficient in a narrow economic sense. In such circumstances the level of economic welfare as viewed by the government may well be raised by a policy designed to alleviate these inequalities. This is the reasoning through which the provision of state education, social security programs, and compulsory pension schemes are justified. It should be stressed that the gains from these policies are with respect to normative assessments of welfare, unlike the positive criterion lying behind the concept of economic efficiency.

In the cases of both market failure and welfare-motivated policies, policy intervention concerns more than just the efficient collection of revenue. The reasons for the failure of the economy to reach the optimal outcome have to be understood, and a policy that can counteract these has to be designed. Extending the scope of the public economics to address such issues provides the breadth to the subject.
5.2.4 Efficiency and Equity

When determining economic policy, governments are faced with two conflicting aims. All governments are concerned with organizing economic activity so that the best use is made of economic resources. This is the efficiency side of policy design. To varying degrees, governments are also concerned to see that the benefits of economic activity are distributed fairly. This is the equity aspect of policy design.

The difficulty facing the government is that the requirements of equity and efficiency frequently conflict. It is often the case that the efficient policy is highly inequitable, while the equitable policy can introduce significant distortions and disincentives. Given this fact, the challenge for policy design is to reach the correct trade-off between equity and efficiency. Quite where on the trade-off the government should locate is dependent on the relative importance it assigns to equity over efficiency.

In this context it is worth adding one final note concerned with the nature of the arguments often used in this book. A standard simplification is to assume that there is a single consumer or that all consumers are identical. In such a setting there can be no distributional issues, so any policy recommendations derived within it relate only to efficiency and not to equity. The reason for proceeding in this way is that it usually permits a much simpler analysis to be undertaken and for the conclusions to be much more precise. When interpreting such conclusions in terms of practical policy recommendations, their basis should never be overlooked.

5.3 Public Sector Growth

The data of chapter 4 showed quite clearly the substantial growth of the public sector in a range of countries during the past century. There are numerous theories that have been advanced to explain why this has occurred. These differ in their emphasis and perspective and are not mutually exclusive. In fact it is reasonable to argue that a comprehensive explanation would involve elements drawn from all.

5.3.1 Development Models

The basis of the development models of public sector growth is that the economy experiences changes in its structure and needs as it develops. Tracing the nature of the development process from the beginning of industrialization through to the completion of the development process, a story of why public sector expenditure increases can be told.
It is possible to caricature the main features of this story in the following way: The early stage of development is viewed as the period of industrialization during which the population moves from the countryside to the urban areas. To meet the needs that result from this, there is a requirement for significant infrastructural expenditure in the development of cities. The typically rapid growth experienced in this stage of development results in a significant increase in expenditure, and the dominant role of infrastructure determines the nature of expenditure.

In what are called the middle stages of development, the infrastructural expenditure of the public sector becomes increasingly complementary with expenditure from the private sector. Developments by the private sector, such as factory construction, are supported by investments from the public sector, such as the building of connecting roads. As urbanization proceeds and cities increase in size, so does population density. This generates a range of externalities such as pollution and crime. An increasing proportion of public expenditure is then diverted away from spending on infrastructure to the control of these externalities.

Finally, in the developed phase of the economy, there is less need for infrastructural expenditure or for the correction of market failure. Instead, expenditure is driven by the desire to react to issues of equity. This results in transfer payments, such as social security, health, and education, becoming the main items of expenditure. Of course, once such forms of expenditure become established, they are difficult to ever reduce. They also increase with heightened expectations and through the effect of an aging population.

Although this theory of the growth of expenditure concurs broadly with the facts, it has a number of weaknesses. Most important, it is primarily a description rather than an explanation. From an economist’s perspective, the theory is lacking in that it does not have any behavioral basis but is essentially mechanistic. What an economist really would wish to see is an explanation in which expenditure is driven by the choices of the individuals that constitute the economy. In the development model the change is just driven by the exogenous process of economic progress. Changes in expenditure should be related to how choices change as preferences or needs evolve over time.

### 5.3.2 Wagner’s Law

Adolph Wagner was a nineteenth-century economist who analyzed data on public sector expenditure for several European countries, Japan, and the United States. These data revealed the fact that was shown in chapter 4: the share of the public sector in gross domestic product had been increasing over time. The content of Wagner’s law was
Chapter 5: Theories of the Public Sector

an explanation of this trend and a prediction that it would continue. In contrast to the basic developments models, Wagner’s analysis provided a theory rather than just a description and an economic justification for the predictions.

The basis for the theory consists of three distinct components. First, it was observed that the growth of the economy results in an increase in complexity. Economic growth requires continual introduction of new laws and the development of the legal structure. Law and order imply continuing increases in public sector expenditure. Second, there was the process of urbanization and the increased externalities associated with it. These two factors have already been discussed in connection with the development models.

The final component underlying Wagner’s law is the most behavioral of the three and is what distinguishes it from other explanations. Wagner argued that the goods supplied by the public sector have a high income elasticity of demand. This claim appears reasonable, for example, for education, recreation, and health care. Given this fact, as economic growth raises incomes, there will be an increase in demand for these products. In fact from a high elasticity it can be inferred that public sector expenditure does rise as a proportion of income. This conclusion is the substance of Wagner’s law.

There have been many attempts at establishing whether Wagner’s law is empirically valid. The problem that surfaces in all of these tests is how to disentangle the causality between public expenditure and the level of income. Wagner’s law proposes that it is income that explains expenditure. In contrast, there is much macroeconomic theory in favor of the argument that government spending explains the level of income—this was the essential insight of Keynesian economics. Tests to date have not convincingly resolved this issue.

In many ways Wagner’s law provides a good explanation of public sector growth. Its main failing is that it concentrates solely on the demand for public sector services. What must determine the level is some interaction between demand and supply. The supply side is explicitly analyzed in the next model.

5.3.3 Baumol’s Law

Rather than work from the observed data, Baumol’s law starts from an observation about the nature of the production technology in the public sector. The basic hypothesis is that the technology of the public sector is labor-intensive relative to that of the private sector. In addition the type of production undertaken leaves little scope for increases in productivity, and that makes it difficult to substitute capital for labor. As examples, hospitals need minimum numbers of nurses and doctors for each patient, and maximum class sizes place lower limits on teacher numbers in schools.
Competition on the labor market ensures that labor costs in the public sector are linked to those in the private sector. Although there may be some frictions in transferring between the two, wage rates cannot be too far out of line. However, in the private sector it is possible to substitute capital for labor when the relative cost of labor increases. Furthermore technological advances in the private sector lead to increases in productivity. These increases in productivity result in the return to labor rising. The latter claim is simply a consequence of optimal input use in the private sector resulting in the wage rate being equated to the marginal revenue product.

Since the public sector cannot substitute capital for labor, the wage increases in the private sector feed through into cost increases in the public sector. Maintaining a constant level of public sector output must therefore result in public sector expenditure increasing. If public sector output/private sector output remain in the same proportion, public sector expenditure rises as a proportion of total expenditure. This is Baumol’s law, which asserts the increasing proportional size of the public sector.

There are a number of problems with this theory. It is entirely technology driven and does not consider aspects of supply and demand or political processes. There are also reasons for believing that substitution can take place in the public sector. For example, additional equipment can replace nurses, and less qualified staff can take on more mundane tasks. Major productivity improvements have also been witnessed in universities and hospitals. Finally, there is evidence of a steady decline in public sector wages relative to those in the private sector. This reflects lower skilled labor being substituted for more skilled.

5.3.4 A Political Model

A political model of public sector expenditure needs to capture the conflict in public preferences between those who wish to have higher expenditure and those who wish to limit the burden of taxes. It must also incorporate the resolution of this conflict and show how the size and composition of actual public spending reflects the preferences of the majority of citizens as expressed through the political process. The political model we now describe is designed to achieve these aims. The main point that emerges is that the equilibrium level of public spending can be related to the income distribution, and more precisely that the growth of government is closely related to the rise of income inequality.

To illustrate this, consider an economy with $H$ consumers whose incomes fall into a range between a minimum of 0 and a maximum of $\bar{y}$. The government provides a public
good that is financed by the use of a proportional income tax. The utility of consumer $i$ who has income $y_i$ is given by
\[ u_i(t, G) = [1 - t] y_i + b(G), \] (5.1)
where $t$ is the income tax rate and $G$ the level of public good provision. The function $b(\cdot)$ represents the benefit obtained from the public good, and it is assumed to be increasing (so the marginal benefit is positive) and concave (so the marginal benefit is falling) as $G$ increases. We denote by $\mu$ the mean income level in the population of consumers, so the government budget constraint is
\[ G = t \mu. \] (5.2)
Using this budget constraint, a consumer with income $y_i$ will enjoy utility from provision of a quantity $G$ of the public good of
\[ u_i(G) = \left(1 - \frac{G}{\mu} \right) y_i + b(G). \] (5.3)

The ideal level of public good provision for the consumer is given by the first-order condition
\[ \frac{\partial u_i(G)}{\partial G} \equiv - \frac{y_i}{\mu} + b'(G) = 0. \] (5.4)
This condition relates the marginal benefit of an additional unit of the public good, $b'(G)$, to its marginal cost $\frac{y_i}{\mu}$. The quantity of the public good demanded by the consumer depends on their income relative to the mean, since this determines the marginal cost.

The marginal benefit of the public good has been assumed to be a decreasing function of $G$, so it follows that the preferred public good level is decreasing as income rises. The reason for this is that with a proportional income tax the rich pay a higher share of the cost of public good than the poor. Thus public good provision will disproportionately benefit the poor.

The usual way to resolve the disagreement over the desired level of public good is to choose by majority voting. If the level of public good is to be determined by majority voting, which level will be chosen? In the context of this model the answer is clear-cut because all consumers would prefer the level of public good to be as close as possible to their preferred level. Given any pair of alternatives, consumers will vote for that which is closest to their preferred alternative. The alternative that is closest for the largest number of consumers will receive maximal support. There is in fact only one
option that will satisfy this requirement: the option preferred by the consumer with the median income. The reason is that exactly one-half of the electorate, above the median income (the rich), would like less public good and the other half, below the median (the poor), would like more public good. Any alternative that is better for one group would be opposed by the other group with opposite preferences. (We explore the theory of voting in detail in chapter 11.)

The political equilibrium $G^*$, determined by the median voter, is then the solution to

$$b'(G^*) = \frac{y_m}{H\mu},$$

where $\frac{y_m}{\mu}$ is the income of the median voter relative to the mean. Since the marginal benefits decrease as public good provision increases, the political equilibrium level of public good increases with income inequality as measured by the ratio of the median to mean income. Accordingly, more inequality as measured by a lower ratio of the median to mean income would lead the decisive median voter to require more public spending.

Government activities are perceived as redistributive tools. Redistribution can be explicit, such as social security and poverty alleviation programs, or it can take a more disguised form like public employment, which is probably the main channel of redistribution from rich to poor in many countries. Because of its nature, and interaction with the tax system, the demand for redistribution will increase as income inequality increases as demonstrated by this political model.

### 5.3.5 Ratchet Effect

Models of the ratchet effect develop the modeling of political interaction in a different direction. They assume that the preference of the government is to spend money. Explanations of why this should be so can be found in the economics of bureaucracy, which is explored in the next section. For now the fact is just taken as given. In contrast, it is assumed that the public do not want to pay taxes. Higher spending can only come from taxes, so by implication the public partially resists this; they do get some benefit from the expenditure. The two competing objectives are moderated by the fact that governments desire re-election. This makes it necessary for government to take some account of the public’s preferences.

The equilibrium level of public sector expenditure is determined by the balance between these competing forces. In the absence of any exogenous changes or of changes in preferences, the level of expenditure will remain relatively constant. In the historical data on government expenditure, the periods prior to 1914, between 1920 and 1940,
Chapter 5: Theories of the Public Sector

and post-1945 can be interpreted as displaying such constancy. Occasionally, though, economies go through periods of significant upheaval such as occurs during wartime. During these periods normal economic activity is disrupted. Furthermore the equilibrium between the government and the taxpayers becomes suspended. Ratchet models argue that wartime permits the government to raise expenditure with the consent of the taxpayers on the understanding that this is necessary to meet the exceptional needs that have arisen.

The final aspect of the argument is that the level of expenditure does not fall back to its original level after the period of upheaval. Several reasons can be advanced for this. First, the taxpayers become accustomed to the higher level of expenditure and perceive this as the norm. Second, debts incurred during the period of upheaval have to be paid off later. This requires the raising of finance. Third, promises made by the government to the taxpayers during periods of upheaval then have to be met. These can jointly be termed ratchet effects that sustain a higher level of spending. Finally, there may occur an inspection effect after an upheaval whereby the taxpayers and government reconsider their positions and priorities. The discovery of previously unnoticed needs then provides further justification for higher public sector spending.

The prediction of the ratchet-effect model is that spending remains relatively constant unless disturbed by some significant external event. These events can trigger substantial increases in expenditure. The ratchet and inspection effects work together to ensure that expenditure remains at the higher level until the next upheaval.

The description of expenditure growth given by this political model is broadly consistent with the data of chapter 4. Before 1914, between 1918 and 1940, and post-1945 the level of expenditure is fairly constant but steps up between these periods. Whether this provides support for the explanation is debatable because the model was constructed to explain these known facts. In other words, the data cannot be employed as evidence that the model is correct, given that the model was designed to explain that data.

5.4 Excessive Government

The theories of the growth of public sector expenditure described above attempt to explain the facts but do not offer comment on whether the level of expenditure is deficient or excessive. They merely describe processes and do not attempt to evaluate the outcome. There are in fact many economists who argue that public sector expenditure is too large and represents a major burden on the economy. While the evidence on this issue is certainly not conclusive, there are a number of explanations of why this should
be so. Several are now described that reach their conclusions not through a cost–benefit analysis of expenditure but via an analysis of the functioning of government.

### 5.4.1 Bureaucracy

A traditional view of bureaucrats is that they are motivated solely by the desire to serve the common good. They achieve this by conducting the business of government in the most efficient manner possible without political or personal bias. This is the idealistic image of the bureaucrat as a selfless public servant. There is a possibility that such a view may be correct. Having said this, there is no reason why bureaucrats should be any different from other individuals. From this perspective it is difficult to accept that they are not subject to the same motivations of self-serving.

Adopting this latter perspective, the theoretical analysis of bureaucracy starts with the assumption that bureaucrats are indeed motivated by maximization of their private utilities. If they could, they would turn the power and influence that their positions give them into income. But, due to the nature of their role, they face difficulties in achieving this. Unlike similarly positioned individuals in the private sector, they cannot exploit the market to raise income. Instead, they resort to obtaining utility from pursuing nonpecuniary goals. A complex theory of bureaucracy may include many factors that influence utility such as patronage, power, and reputation. However, to construct a basic variant of the theory, it is sufficient to observe that most of these factors can be related to the size of the bureau. The bureaucrat can therefore be modeled as aiming to maximize the size of his bureau in order to obtain the greatest nonpecuniary benefits. It is as a result of this behavior that the size of government becomes excessive.

To demonstrate excessive bureaucracy, let $y$ denote the output of the bureau as observed by the government. In response to an output $y$, the bureau is rewarded by the government with a budget of size $B(y)$. This budget increases as observed output rises ($B'(y) > 0$) but at a falling rate ($B''(y) < 0$). The cost of producing output is given by a cost function $C(y)$. Marginal cost is positive ($C'(y) > 0$) and increasing ($C''(y) > 0$). It is assumed that the government does not know this cost structure—only the bureaucrat fully understands the production process. What restraints the behavior of the bureaucrat is that the budget received from the government is sufficient to cover the costs of running the bureau.

The decision problem of the bureaucrat is then to choose output to maximize the budget subject to the requirement that the budget is sufficient to cover costs. This optimization can be expressed by the Lagrangian

$$L = B(y) + \lambda [B(y) - C(y)],$$  \hspace{1cm} (5.6)
where $\lambda$ is the Lagrange multiplier on the constraint that the budget equals cost. Differentiating the Lagrangian with respect to $y$ and solving characterizes the optimum output from the perspective of the bureaucrat, $y^b$, by

$$B'(y^b) = \frac{\lambda}{\lambda + 1} C'(y^b).$$

(5.7)

Since the Lagrange multiplier, $\lambda$, is positive, this expression implies that $B' < C'$ at the bureaucrats optimum choice of output.

We wish to contrast the bureaucracy outcome with the outcome that occurs when the government has full information. With full information there exists a variety of ways to model efficiency. One way would be to place the bureau within a more general setting and consider its output as one component of overall government intervention. A cost–benefit calculation for government intervention would then determine the efficient level of bureau output. A simpler alternative, and the one we choose to follow, is to determine the efficient output by drawing an analogy between the bureau and a profit-maximizing firm. The firm chooses its output to ensure that the difference between revenue and costs is made as large as possible. By this analogy, the bureau should choose output to maximize its budget less costs, $B(y) - C(y)$. For the bureau this is the equivalent of profit maximization.

Differentiating with respect to $y$, we equate the marginal effect of output on the budget to marginal cost and determine the efficient output $y^*$. The efficient output satisfies $B'(y^*) = C'(y^*)$. The output level chosen by the bureaucrat can easily be shown to be above the efficient level. This argument is illustrated in figure 5.1. The increasing marginal cost curve and declining marginal benefit curve are consequences of the assumptions already made. The efficient output occurs at the intersection of these curves. In contrast, the output chosen by the bureaucrat satisfies $B'(y^b) < C'(y^b)$, so it must lie to the right of $y^*$. In fact the budget covers costs when the area under the marginal budget curve $a$ equals the area under the marginal cost curve $b$. It is clear from this figure that the large size of the bureaucracy is determined by the excesses of an individual bureaucrat.

This simple model shows how the pursuit of personal objectives by bureaucrats can lead to an excessive size of bureaucracy. Adding together the individual bureaus that comprise the public sector makes this excessive in the aggregate. The excessive size is simply an inefficiency, since money is spent on bureaus that are not generating sufficiently valuable results.

The argument just given is enticing in its simplicity, but it is restricted by the fact that it is assumed that the bureaucrats have freedom to set the size of the bureau. There
are various ways this limitation can be addressed. Useful extensions are to have the freedom constrained by political pressure or through a demand function. Although doing either of these would lessen the excess, the basic moral that bureaucrats have incentives to overly enlarge their bureaus would still remain. Whether they do so in practice is dependent on the constraints placed on them.

### 5.4.2 Budget-Setting

An alternative perspective on excessive bureaucracy can be obtained by considering a different process of budget determination. A motivation for this is the fact that each government department is headed by a politician who obtains satisfaction from the size of the budget. Furthermore, in many government systems, budgets for departments are determined annually by a cabinet meeting. This meeting takes the budget bids from the individual departments and allocates a central budget on the basis of these bids. Providing a model incorporating the bids then determines how departments’ budgets evolve over time.

A simple process of this form can be the following: Let the budget for year $t$ be given by $B_t$. The budget claim for year $t + 1$ is then given by

$$B_{t+1}^c = (1 + \alpha)B_t, \quad (5.8)$$
where $\alpha > 0$ is the rate at which departments inflate their budget claims. Such a rule represents a straightforward mechanical method of updating the budget claim—last year’s is taken and a little more added. It is, of course, devoid of any basis in efficiency. The meeting of cabinet then takes these bids and proportionately reduces them to reach the final allocation. The agreed budget is written as

$$B_{t+1} = (1 - \gamma)B_{t+1}^c = (1 - \gamma)(1 + \alpha)B_t,$$

(5.9)

where $0 < \gamma < 1$ is the rate at which the cabinet deflates each budget claim. The expression above gives a description of the change in the budget over time.

It can be seen that if $\alpha > \gamma$, then the budget will grow over time. Its development bears little relationship to needs, so there is every possibility that expenditure will eventually become excessive even if it initially begins at an acceptable level. When $\alpha < \gamma$, the budget will fall over time. Although either case is possible, the observed pattern of growth lends some weight to the former assumption.

This form of model could easily be extended to incorporate more complex dynamics but without enhancing the content of the simple story it tells. The modeling of budget determination as a process entirely independent on what is good for the economy provides an important alternative perspective on how the public sector may actually function. Even if the truth is not quite this stark, reasoning of this kind does put into context models that are based on the assumption that the government is informed and efficient.

### 5.4.3 Monopoly Power

The basis of elementary economics is that market equilibrium is determined via the balance of supply and demand. Those supplying the market are assumed to be distinct from those demanding the product. In the absence of monopoly power, the equilibrium that is achieved will be efficient. If the same reasoning could be applied to the goods supplied by the public sector, then efficiency would also arise there. Unfortunately, there are two reasons why efficiency is not possible. First, the public sector can award itself a monopoly in the supply of its goods and services. Second, this monopoly power may be extended into market capture.

Generally, a profit-maximizing monopolist will always want to restrict its level of output below the competitive level so that monopoly power will provide a tendency for too little government rather than the converse. This would be a powerful argument were it not for the fact that the government can choose not to exercise its monopoly power in this way. If it is attempting to achieve efficiency, then it will certainly not
do so. Furthermore, since the government may not be following a policy of profit maximization, it might actually exploit its monopoly position to oversupply its output. This takes the analysis back in the direction of the bureaucracy model.

The idea of market capture is rather more interesting and arises from the nature of goods supplied by the public sector. Rather than being standard market goods, many of them are complex in nature and not fully understood by those consuming them. Natural examples of such goods would be education and health care. In both cases the consumer may not understand quite what the product is, nor what is best for them. Although this is important, it is also true of many other goods. The additional feature of the public sector commodities is that demand is not determined by the consumers and expressed through a market. Instead, it is delegated to specialists such as teachers and doctors. Furthermore these same specialists are also responsible for setting the level of supply. In this sense they can be said to capture the market.

The consequence of this market capture is that the specialists can set the level of output for the market that most meets their objectives. Naturally, since most would benefit from an expansion of their profession, within limits, this gives a mechanism that leads to supply in excess of the efficient level. The limits arise because they won’t want to go so far that competition reduces the payment received or lowers standards too far. Effectively they are reaching a trade-off between income and power, where the latter arises through the size of the profession. The resulting outcome has no grounds in efficiency and may well be too large.

5.4.4 Corruption

Corruption does not emerge as a moral aberration but as a general consequence of government officials using their power for personal gain. Corruption distorts the allocation of resources away from productive toward rent-seeking occupations. Rent-seeking (studied in chapter 12) is the attempt to obtain a return above what is judged adequate by the market. Monopoly profit is one example, but the concept is much broader. Corruption is not just redistributive (taking wealth from others to give it to some special interests), it can also have enormous efficiency costs. By discouraging the entrepreneurs on whom they prey, corruptible officials may have the effect of stunting economic growth.

Perhaps the most important form of corruption in many countries is predatory regulation. This is the process by which the government intentionally creates regulations that entrepreneurs have to pay bribes to get around. Because it raises the cost of productive activity, this form of corruption reduces efficiency. The damage is particularly large
when several government officials, acting independently, create distinct obstacles to economic activity so that each can collect a separate bribe in return for removing the obstacle (e.g., creating the need for a license and then charging for it). When entrepreneurs face all these independent regulatory obstacles, they eventually cease trying, or else move into the underground economy to escape regulation altogether. Thus corruption is purely harmful from this perspective.

How could we give a positive role for a bribe-based corruption system? One possibility is that bribery is like an auction mechanism that directs resources to their best possible use. For example, corruption in procurement is similar to auctioning off the contract to the most efficient entrepreneur who can afford the highest bribe. However, there are some problems with this bribery-based system. First, we care about the means as well as the ends. Bribery is noxious. Allowing bribery will destroy much of the goodwill that supports the system. Second, people should not be punished for their honesty. Indeed honest government officials can be used to create benchmarks by which to judge the performance of the more opportunistic officials. Third, it is impossible to optimize or even manage underground activities such as bribery.

5.4.5 Government Agency

Another explanation for excessive government is the lack of information available to voters. The imperfect information of voters enables the government to grow larger by increasing the tax burden. From this perspective government growth reflects the abuse of power by greedy bureaucrats. The central question is then how to set incentives that encourage the government to work better and to cost less, subject to the information available.

To illustrate this point, consider a situation where the cost to the government of supplying a public good can vary. The unit cost is either low, at $c_l$, or is high, at $c_h$. The gross benefit to the public from a level $G_i$ of public good is given by the function $b(G_i)$, which is increasing and concave. The net benefit is $b(G_i) - t_i$, where $t_i$ is the tax paid to the government for the public good provision. The chosen quantity of the public good will depend on the unit cost of the government. The benefit to the government of providing the public good is the difference between the tax and the cost. So, when the cost is $c_i$, the benefit is $t_i - c_iG_i$.

When the public is informed about the level of cost of the government, the quantity of public good will be chosen to maximize the net benefit subject to the government breaking even. For cost $c_i$, the public net benefit with the government breaking even is $b(G_i) - c_iG_i$. The public will demand a level of public good such that the marginal
benefit is equal to the marginal cost, namely \( b(G_i) = c_i \), and will pay the government \( t_i = c_i G_i \), for \( i = h, \ell \). This is shown in figure 5.2.

Now assume that the public cannot observe whether the government has cost \( c_\ell \) or \( c_h \). The government can then benefit by misrepresenting the cost to the public: for instance, it can exaggerate the cost by adding expenditures that benefit the government but not the public. When the cost is high, the government cannot exaggerate. When the cost is low, the government is better off pretending the cost is high to get tax \( t_h \) for the amount \( G_h \) of public good instead of getting \( t_\ell \) for producing \( G_\ell \). Misrepresenting in this way leads to the benefit of \( G_h[c_h - c_\ell] \) for the government, which is shown in figure 5.2.

To eliminate this temptation, taxpayers must pay an extra amount \( r > 0 \) to the government in excess of its cost when the government pretends to have the low cost. This is called the informational rent. Since the truly high-cost government cannot further inflate its cost, the public pays \( t_h = c_h G_h \) when the government reports a high cost. If the reported cost is low, the taxpayers demand the amount \( G_\ell \) of public good defined by \( b(G_\ell) = c_\ell \) and pay the government \( t_\ell = c_\ell G_\ell + r \), where \( r \) is exactly the extra revenue the government could have made if it had pretended to have high cost. To give a government with a low cost just enough revenue to offset its temptation to pretend to have higher cost, it is necessary that \( r = [c_h - c_\ell]G_h \). This is the rent required to induce truthful revelation of the cost and have the provision of the public good equal to that when the public is fully informed.

![Figure 5.2](image-url)

**Figure 5.2**
Government agency
It is possible for the taxpayers to reduce this excess payment by demanding that the high-cost government supply less than it would with full information. Assume that cost is low with probability $p_l$ and high with probability $p_h = 1 - p_l$. By maximizing their expected benefit subject to the government telling the truth, it can be shown that revelation can be obtained at the least cost by demanding an amount $G_h$ of public services defined by

$$b'(G_h) = c_h + \frac{p_l}{1 - p_l} (c_h - c_l).$$

This quantity is lower than that with full information. The distortion of the quantity demanded from the high-cost government results from a simple cost–benefit argument. It trades off the benefit of reducing the rent, which is proportional to the cost difference $[c_h - c_l]$, and the probability $p_l$ that the government is of the low-cost type against the cost of imposing the distortion of the quantity on the high-cost government that occurs with probability $1 - p_l$.

Therefore, if the government is truly low cost, it need not be given the high tax. However, to eliminate the temptation for cost inflation, taxpayers have to provide the government just enough of the rent as a reward for reporting truthfully when its cost of public services is low. Because the ability of the government to misrepresent its costs allows it to earn rents and distort the level of provision, eventually the informational rent makes the government bigger than it should be.

### 5.4.6 Cost Diffusion

The last explanation we present for the possibility of excessively large government is the common resource problem. The idea is that spending authorities are dispersed while the treasury has the responsibility of collecting enough revenue to balance the overall budget. Each of the spending authorities has its own spending priorities, with little consideration for others’ priorities, that it can be better met by raiding the overall budget. This is the common resource problem, just like that of several oil companies tapping into a common pool underground or fishermen netting in a single lake. In all cases it leads to excess pressure on the common resource. From this perspective, a single committee with expenditure authority would have a much better sense of the opportunity cost of public funds, and could better compare the merits of alternative proposals, than the actual dispersed spending authorities. The current trend toward federalism and devolution aggravates this common-pool problem. The reason is essentially that each district can impose projects whose cost is shared by all other districts, and so support
higher size projects than they would if they had to cover the full costs. We discuss in more detail the various aspects of federalism in chapter 19.

The problem can also be traced down to the individual level. Consider public services like pensions, health care, and schools and infrastructure work like bridges, roads, and rail tracks. It is clear that for these public services, and many others, the government does not charge the direct users the full marginal cost but subsidizes the activities partly or wholly from tax revenues. There is an obvious equity concern behind this fact. But it is then natural that users who do not bear the full cost will support more public services than they would if they had to cover the full cost. The same argument applies in the opposite direction when contemplating some cut in public spending: contributors who are asked to make concessions are concentrated and possibly organized through a lobby with large per capita benefits from continued provision of specific public services. In contrast, the beneficiaries of downsizing public spending, the taxpayers as a whole, are diffuse with small per capita stakes. This makes it less likely that they can offer organized support for the reform. To sum up, many public services are characterized by the concentration of benefits to a small group of users or recipients and the diffusion of costs to the large group of taxpayers. This results in biases toward continuous demand for more public spending.

5.5 Conclusions

This chapter has provided a number of theories of public sector growth that are designed to explain the data presented in chapter 4. Each theory has some points to commend it, but none is entirely persuasive. It is fair to say that all provide a partial insight and have some element of truth. A more general story drawing together the full set of components, including the ratchet effect, income effect, political process, production technology, and bureaucracy, would have much in its favor. This would be especially so if combined with the voting models of chapter 11.

The bureaucracy models are particularly attractive because they show how economic analysis can be applied to what appears to be a noneconomic problem. In doing so, they generate an interesting conclusion that casts doubt on the efficiency of government. This illustrates how the method of economic reasoning can be applied to understand the outcome of what is at first sight a noneconomic problem.

The perennial question of whether the government has grown too large is difficult to answer. The reason is that the government is both complementary to the market and a competitor of the market. As a major employer the government competes with
businesses looking to hire talented people. The possibility that the best and brightest be-
come public officials and politicians, rather than entrepreneurs, is considered by many
as very costly to society, since they are seen as devoting their talents to taking wealth
from others rather than creating it. When people pay taxes, they have less money to
spend on other goods and services provided by the market. Likewise, when the govern-
ment borrows money, it competes with companies looking to raise capital. In some areas
like health care and education, public and private services compete with each other. But
at the same time the government also serves as a useful complement to every business
activity by providing basic infrastructure and civil order. Every business depends on
the government for things like protection of life and property, a transportation network,
civil courts, and a stable currency. Without these things, people couldn’t do business.

Finally, whether an activity is carried out in the public sector or the private sec-
tor is itself endogenous. As in architecture, the functions suggest the form. Take the
example of education where the goals are multiple (literacy, vocational skills, citizen-
ship, equality of chance, preparation for life) and not precisely measurable and where
several stakeholders are involved (parents, employers, students, teachers, taxpayers)
with possibly conflicting interests. It is not immediately clear that the market with its
single-minded focus can cope adequately with all these aspects, and the risk is that the
market could bias the activity toward dimensions that matter more for profit-making.

**Further Reading**

The concept of the minimal state is explored in:


An account of Wagner’s law can be found in:


Recent empirical tests are reviewed in:

1–17.

The classic study of public sector growth is:

Princeton University Press.

A nontechnical account on corruption and government is:

Cambridge: Cambridge University Press.

The theory of bureaucracy was first developed in:

A fascinating book on bureaucracy from a political scientist is:


The political theory of the size of the government is based on:


The main reference on government agency is:


---

**Exercises**

5.1 Can trade occur in a world with no rules? Is it ever possible to have no rules?

5.2 If it takes four days of labor to produce a week’s food, and one day of labor to steal a week’s food, what will be the equilibrium outcome?

5.3 Would a minimal state finance a fire service?

5.4 Do the data of chapter 4 support the view that governments have expanded beyond the minimal state?

5.5 Discuss whether provision of state education enhances efficiency or equity. What about health care?

5.6 Would a minimal state:
   a. Ensure that wage agreements are enforced?
   b. Limit maximum working hours?
   c. Prevent involuntary overtime?

5.7 Will efficiency be achieved if:
   a. No agent knows what the profit level of a firm will be next year?
   b. One agent does know what the profit level will be?

5.8 Can insider trading occur in the idealized competitive economy?

5.9 All our sulphur emissions are blown into a neighboring country. Can our economy be efficient?

5.10 Are the following policies conducted for efficiency or equity motives:
   a. Provision of unemployment benefits?
   b. Provision of primary education?
   c. Provision of higher education?
   d. Provision of retirement pensions?
   e. Prohibiting smoking in public places?
f. Imposing higher marginal income tax rates on people with higher incomes?

In the case of efficiency motives, discuss the type of market failure involved.

5.11 Should the government intervene with a redistributive policy if income inequality is due to:
   a. Differences in work effort?
   b. Differences in ability?

5.12 Consider two consumers who each have a total of $T$ hours to allocate between production and theft. Assume that production produces output $y_p = \log(t_p)$ for $t_p$ units of time in production. If time $t_f$ is devoted to theft, then a proportion $\alpha t_f/T$ of the other consumer’s output can be stolen. Assuming that each unit of output has price $p$ and both consumers attempt to maximize their wealth, what is the equilibrium? How does the equilibrium depend on the value of $\alpha$? What is the equilibrium if there is no theft? What is the maximum that would be paid to prevent theft?

5.13 Describe the expenditures at each stage of the development process in terms of efficiency and equity.

5.14 a. Provide a graphical two-commodity (one private good and one public good) example of a preference relation generating an income elasticity of the demand for the public good that is greater than one.

b. Show that in this case the fraction of the budget spent on public good increases as income increases. Explain also why the indifference curve in this two-commodity space is negatively sloped and convex (preferences are convex if for any two points on the same indifference curve the line segment between them is in the “weakly preferred” set, which is defined as the set of commodity bundles [weakly] preferred to any bundle that lies on the indifference curve).

5.15 Is a theory of any value if it makes no predictions? Discuss the development model of public sector expenditure growth from this perspective.

5.16 In the same two-commodity economy as in the previous exercise, keeping constant the price of the private good:
   a. Give a graphical illustration of a preference relation generating a price elasticity of demand for public good that is less than one in absolute value.
   b. Show that in this case the fraction of the budget spent on the public good increases as the (relative) price of public good increases.

5.17 Assume that the demand for public output at time $t$, $G_t$, is given by the demand function $G_t = (Y_t)^\alpha$, where $Y_t$ is national income at time $t$.
   a. What is the income elasticity of demand?
   b. For what values $\alpha$ of does Wagner’s law hold? Show that expenditure on public output rises as a fraction of income for these values.
   c. Assume that national income growth is determined by $Y_{t+1} = \beta Y_t + \left[\bar{G} - G_t\right]$. Will an increase in $G_t$ raise $Y_t$ in the cases where Wagner’s law applies? Explain the answer.

5.18 Obtain data on public sector expenditure as a proportion of gross domestic product since 1970. Is expenditure still growing? Assess the answer relative to the arguments of the development model. Do the data describe a relation of demand to income that supports Wagner’s law?
5.19 Sketch a story of learning about preferences that supports the ratchet effect.

5.20 Assume that the rental rate for capital is fixed at $r$. If the private sector has a production function $y = K^{1/2} [tL]^{1/2}$ and sells output at price $p$, what happens to the wage rate as technical progress increases $t$? What would happen if $r$ were not fixed? Relate your answer to Baumol’s law.

5.21 Suppose that the production function is $y = \log(K) + \log(tL)$. If demand is constant and labor productivity doubles, what happens to labor demand? What will happen to the wage rate if the economy has many firms in this position? Does this analysis support Baumol’s law?

5.22 Consider a simplified setting for Baumol’s law where there is no capital. Let the private sector have the production technology $y^p = tL$, where $L$ is labor input and $t$ denotes exogenous technical progress that occurs as time passes.
   a. With output price $p$, use the condition of zero profit at the competitive equilibrium to determine the wage rate.
   b. Calculate the cost function for the firm.
   c. Let the public sector have production function $y^g = L$. Show that the ratio of marginal costs in the two sectors grows at rate $t$.
   d. Find the equilibrium path for the economy if it has a single consumer, with preferences given by $U = \log(y^p) + \log(y^g)$, who can supply one unit of labor in each time period. Comment on the relative size of the public sector.

5.23 Describe the benefits a bureaucrat can obtain from an increase in bureau size. Are there any private costs?

5.24 Assume that utility can be derived from pecuniary and nonpecuniary sources. Consider a bureaucrat whose preferences are described by $U = M^{1/2}(1 + b)^{1/2}$, where $M$ denotes income and $b$ nonpecuniary benefits derived from employment.
   a. Derive an expression for the additional income that must be paid to maintain utility if benefits are withdrawn.
   b. Let $b$ represent the number of staff employed. If each member of staff is paid a wage $w$, what is the cost of generating nonpecuniary benefits of value $M/2$?
   c. Is the provision of nonpecuniary benefits a cost-effective method of raising utility?

5.25 Do regular changes in government assist or hinder bureaucrats in expanding their bureaus?

5.26 Why might it be better to tolerate bureaus of excessive size rather than permit bureaucrats to seek rewards in cash?

5.27 a. In the model of bureaucracy, let $B(y) = y^{1/2}$ and $C(y) = y^2$. Calculate the value $y^*$ that maximizes $B(y) - C(y)$. For what values of $y$ does $B(y) = C(y)$? Use this to find $y^b$. Show that $y^b > y^*$.
   b. Now let the bureaucrat’s income be given by $M = a + by$, and let his utility be given by $U = B(y) + M$. Does this alter the chosen value of $y^b$?
   c. Is there any pay scale relating $y$ to $M$ that can lead the bureaucrat to choose $y^*$?
5.28 How can right-wing and left-wing governments be modeled using the budget-setting framework?

5.29 Consider a profession with \( n \) members and revenue determined by \( r = bn - \left( \frac{1}{4} \right) n^2 \). What value of \( n \) maximizes total revenue? What value maximizes revenue per member of the profession? If the benefit from the profession is \( vn \), what is the efficient membership? Contrast these three membership levels.

5.30 a. For the inverse demand function \( p = a - by \) and cost function \( C(y) = cy \), contrast the output choices of a profit-maximizing monopolist, an output-maximizing monopolist and a revenue-maximizing monopolist. Which is the best description of the public sector?
   b. Now let the number of members in a profession be \( n \). Given a fixed price \( p \) for output and a cost function \( C(y) \), calculate the values of \( y \) and \( n \) that maximize per capita profit. What are the efficient values of \( y \) and \( n \)?

5.31 Consider an economy with two goods (consumption and labor) where individuals differ only in their income-generating ability \( a_i \). Suppose that the distribution of abilities in the population is such that the median ability level, \( a_m \), is strictly less than the average ability level, \( \bar{a} \). Suppose that the income level of each individual \( i \) is \( y_i = (1 - t) a_i \), where \( t > 0 \) is the proportional income tax rate. Suppose also that all tax revenues are redistributed through a uniform lump-sum grant \( g \).
   a. What is the tax rate that maximizes the lump-sum grant?
   b. Using the fact that individual \( i \)’s after-tax income is equal to \( g + (1 - t) y_i \), show that income equality requires \( t = 1 \) and that the poorest \( (a_i = 0) \) can be better off with a lower tax rate (and thus more inequality).
   c. If every individual \( i \)’s preference over \( (t, g) \) is \( v_i = g + \frac{1}{2} [1 - t]^2 a_i \), then what will be the tax rate chosen by majority voting? (Hint: The median ability individual is the decisive voter in this model.)
   d. Show that the majority voting tax rate is increasing with the difference between the average and the median ability levels, \( \bar{a} - a_m > 0 \). Does that mean that increasing inequality raises the relative size of the public sector (as measured by the tax rate)?

5.32 The government orders that all consumers must eat at least 5 units of vegetables per day. Let a consumer have the utility function \( U = v^{0.2} x^{0.8} \), where \( v \) denotes units of vegetables per day and \( x \) other consumption.
   a. Find the optimal consumption plan before the government order when the price of vegetables is 1 per unit and the price of other consumption is 1 per unit and the consumer has income of \( M = 10 \).
   b. What is the consumption plan after the government has ordered that at least 5 units of vegetables must be consumed?
   c. Show that the consumption plan in part b can be obtained by a subsidy to vegetables and a tax on other consumption.
   d. Find the cost for the government of the subsidy and tax calculated in part c.
III DEPARTURES FROM EFFICIENCY
6.1 Introduction

When a government provides a level of national defense sufficient to make a country secure, all inhabitants are simultaneously protected. Equally, when a radio program is broadcast, it can be received simultaneously by all listeners in range of the transmitter. The possibility for many consumers to benefit from a single unit of provision violates the assumption of the private nature of goods underlying the efficiency analysis of chapter 2. The Two Theorems relied on all goods being private in nature, so they can only be consumed by a single consumer. If there are public goods such as national defense in the economy, then market failure occurs and the unregulated competitive equilibrium will fail to be efficient. This inefficiency implies that there is a potential role for government intervention.

The chapter begins by defining a public good, and distinguishing between public goods and private goods. Doing so provides considerable insight into why market failure arises when there are public goods. The inefficiency is demonstrated by analyzing the equilibrium that is achieved when it is left to the market to provide public goods. The Samuelson rule characterizing the efficient provision of a public good is then derived. This permits a comparison of the equilibrium and with an efficient allocation.

The focus of the chapter next turns to the consideration of methods through which efficiency can be achieved. The first of these, the Lindahl equilibrium, is based on the observation that the price each consumer pays for the public good should reflect their valuation of it. The Lindahl equilibrium achieves efficiency, but since the valuations are private information, it generates incentives for consumers to provide false information. Mechanisms designed to elicit the correct statement of these valuations are then considered. The theoretical results are contrasted with the outcomes of experiments designed to test the extent of false statement of valuations and the use of market data to calculate valuations. These results are primarily static in nature. To provide some insight into the dynamic aspects of public good provision, the chapter is completed by the analysis of two different forms of fund-raising campaign that permit sequential contributions.
6.2 Definitions

The pure public good has been the subject of most economic analyses of public goods. In many ways the pure public good is an abstraction that is adopted to provide a benchmark case against which other, more realistic, cases can be assessed. A pure public good has the following two properties:

- **Nonexcludability**  
  If the public good is supplied, no consumer can be excluded from consuming it.

- **Nonrivalry**  
  Consumption of the public good by one consumer does not reduce the quantity available for consumption by any other.

In contrast, a private good is excludable at no cost and is perfectly rivalrous: if it is consumed by one person, then none of it remains for any other. Although they were not made explicit, these properties of a private good have been implicit in how we have analyzed market behavior in earlier chapters. As we will see, the efficiency of the competitive economy is dependent on them.

The two properties that characterize a public good have important implications. Consider a firm that supplies a pure public good. Since the good is nonexcludable, if the firm supplies one consumer, then it has effectively supplied the public good to all. The firm can charge the initial purchaser but cannot charge any of the subsequent consumers. This prevents it from obtaining payment for the total consumption of the public good. The fact that there is no rivalry in consumption implies that the consumers should have no objection to multiple consumption. These features prevent the operation of the market equalizing marginal valuations as it does to achieve efficiency in the allocation of private goods.

In practice, it is difficult to find any good that perfectly satisfies both the conditions of nonexcludability and nonrivalry precisely. For example, the transmission of a television signal will satisfy nonrivalry, but exclusion is possible at finite cost by scrambling the signal. Similar comments apply, for example, to defense spending, which will eventually be rivalrous as a country of fixed size becomes crowded and from which exclusion is possible by deportation. Most public goods eventually suffer from congestion when too many consumers try to use them simultaneously. For example, parks and roads are public goods that can become congested. The effect of congestion is to reduce the benefit the public good yields to each user. Public goods that are excludable, but at a cost, or suffer from congestion beyond some level of use are called impure. The properties
of impure public goods place them between the two extremes of private goods and pure public goods.

A simple diagram summarizing the different types of good and the names given to them is shown in figure 6.1. These goods vary in the properties of excludability and rivalry. In fact, it is helpful to envisage a continuum of goods that gradually vary in nature as they become more rivalrous or more easily excludable.

The pure private good and the pure public good have already been identified. An example of a common property good is a lake that can be used for fishing by anyone who wishes, or a field that can be used for grazing by any farmer. This class of goods (usually called the commons) is studied in chapter 8. The problem with the commons is the tendency of overusing them, and the usual solution is to establish property rights to govern access. This is what happened in the sixteenth century in England where common land was enclosed and became property of the local landlords. The landlords then charged grazing fees, and so cut back the use. In some instances property rights are hard to define and enforce, as is the case of the control over the high seas or air quality. For this reason only voluntary cooperation can solve the international problems of overfishing, acid rain, and the greenhouse effect. Club goods are public goods for which exclusion is possible. The terminology is motivated by sport clubs whose facilities are a public good for members but from which nonmembers can be excluded. Clubs are studied in chapter 7.
Public goods do not conform to the assumptions required for a competitive economy to be efficient. Their characteristics of nonexcludability and nonrivalry lead to the wrong incentives for consumers. Since they can share in consumption, each consumer has an incentive to rely on others to make purchases of the public good. This reliance on others to purchase is called free-riding, and it is this that leads to inefficiency.

To provide a model that can reveal the motive for free-riding and its consequences, consider two consumers who have to allocate their incomes between purchases of a private good and a public good. Assume that the consumers take the prices of the two goods as fixed when they make their decisions. If the goods were both private, we could move immediately to the conclusion that an efficient equilibrium would be attained. What makes the public good different is that each consumer derives a benefit from the purchases of the other. This link between the consumers, which is absent with private goods, introduces strategic interaction into the decision processes. With the strategic interaction the consumers are involved in a game, so equilibrium is found using the concept of a Nash equilibrium.

The consumers have income levels $M^1$ and $M^2$. Income must be divided between purchases of the private good and the public good. Both goods are assumed to have a price of 1. With $x^h$ used to denote purchase of the private good by consumer $h$ and $g^h$ to denote purchase of the public good, the choices must satisfy the budget constraint $M^h = x^h + g^h$. The link between consumers comes from the fact that the consumption of the public good for each consumer is equal to the total quantity purchased, $g^1 + g^2$. Hence, when making the purchase decision, each consumer must take account of the decision of the other.

This interaction is captured in the preferences of consumer $h$ by writing the utility function as

$$U^h(x^h, g^1 + g^2).$$

(6.1)

The standard Nash assumption is now imposed that each consumer takes the purchase of the other as given when they make their decision. By this assumption, consumer 1 chooses $g^1$ to maximize utility given $g^2$, while consumer 2 chooses $g^2$ given $g^1$. This can be expressed by saying that the choice of consumer 1 is the best reaction to $g^2$ and that of consumer 2 the best reaction to $g^1$. The Nash equilibrium occurs when these reactions are mutually compatible, so that the choice of each is the best reaction to the choice of the other.
The Nash equilibrium can be displayed by analyzing the preferences of the two consumers over different combinations of $g^1$ and $g^2$. Consider consumer 1. Using the budget constraint, we can write their utility as $U^1(M^1 - g^1, g^1 + g^2)$. The indifference curves of this utility function are shown in figure 6.2. These can be understood by noting that an increase in $g^2$ will always lead to a higher utility level for any value of $g^1$. For given $g^2$, an increase in $g^1$ will initially increase utility as more preferred combinations of private and public good are achieved. Eventually, further increases in $g^1$ will reduce utility as the level of private good consumption becomes too small relative to that of public good. The income level places an upper limit upon $g^1$.

Consumer 1 takes the provision of $2$ as given when making their choice. Consider consumer 2 having chosen $\bar{g}^2$. The choices open to consumer 1 then lie along the horizontal line drawn at $\bar{g}^2$ in figure 6.2. The choice that maximizes the utility of consumer 1 occurs at the tangency of an indifference curve and the horizontal line—this is the highest indifference curve they can reach. This is shown as the choice $\hat{g}^1$. In the terminology we have chosen, $\hat{g}^1$ is the best reaction to $\bar{g}^2$. Varying the level of $\bar{g}^2$ will lead to another best reaction for consumer 1. Doing this for all possible $\bar{g}^2$ traces out the optimal choices of $g^1$ shown by the locus through the lowest point on each indifference curve. This locus is known as the Nash reaction function (or best-response function) and depicts the value of $g^1$ that will be chosen in response to a value of $g^2$. This construction can be repeated for consumer 2 and leads to figure 6.3. For consumer 2, utility increases with $g^1$, and thus indifference curves further to the right reflect
higher utility levels. The best reaction for consumer 2 is shown by $\hat{g}^2$, which occurs where the indifference curve is tangential to the vertical line at $g^1$. The Nash reaction function links the points where the indifference curves are vertical.

The Nash equilibrium occurs where the choices of the two consumers are the best reactions to each other, so neither has an incentive to change their choice. This can only hold at a point where the Nash reaction functions cross. The equilibrium is illustrated in figure 6.4 in which the reaction functions are simultaneously satisfied at their intersection. By definition, $\hat{g}^1$ is the best reaction to $\hat{g}^2$ and $\hat{g}^2$ is the best reaction to $\hat{g}^1$. The equilibrium is privately optimal: if a consumer were to unilaterally raise or reduce his purchase, then he would move to a lower indifference curve.

Having determined the equilibrium, its welfare properties can now be addressed. From the construction of the reaction functions, it follows that at the equilibrium the indifference curve of consumer 1 is horizontal and that of consumer 2 is vertical. This is shown in figure 6.5. It can be seen that all the points in the shaded area are Pareto-preferred to the equilibrium—moving to one of these points will make both consumers better off. Starting at the equilibrium, these points can be achieved by both consumers simultaneously raising their purchase of the public good. The Nash equilibrium is therefore not Pareto-efficient, although it is privately efficient. No further Pareto improvements can be made when a point is reached where the indifference curves are tangential. The locus of these tangencies, which constitutes the set of Pareto-efficient allocations, is also shown in figure 6.5.
Figure 6.4
Nash equilibrium

Figure 6.5
Inefficiency of equilibrium
The analysis has demonstrated that when individuals privately choose the quantity of the public goods they purchase, the outcome is Pareto-inefficient. A Pareto improvement can be achieved by all consumers increasing the purchases of public goods. Consequently, compared to Pareto-preferred allocations, the total level of the public good consumed is too low. Why is this so? The answer can be attributed to strategic interaction and the free-riding that results. The free-riding emerges from each consumer relying on the other to provide the public good and thus avoiding the need to provide for themselves. Since both consumers are attempting to free-ride in this way, too little of the public good is ultimately purchased. In the absence of government intervention or voluntary cooperation, inefficiency arises.

6.4 Efficient Provision

Efficiency in consumption for private goods is guaranteed by each consumer equating their marginal rate of substitution to the price ratio. The strategic interaction inherent with public goods does not ensure such equality. At a Pareto-efficient allocation with the public good, the indifference curves are tangential. However, this does not imply equality of the marginal rates of substitution because the indifference curves are defined over quantities of the public good purchased by the two consumers. As will soon be shown, the efficiency condition involves the sum of marginal rates of substitution and is termed the Samuelson rule in honor of its discoverer.

The basis for deriving the Samuelson rule can be observed in figure 6.5, where the locus of Pareto-efficient allocations has the property that the indifference curves of the two consumers are tangential. The gradient of these indifference curves is given by the rate at which $g^2$ can be traded for $g^1$ while keeping utility constant. The tangency conditions can then be expressed by requiring that the gradients are equal so that

$$\frac{dg^2}{dg^1}|_{U^1_{const}} = \frac{dg^2}{dg^1}|_{U^2_{const}}. \quad (6.2)$$

Calculating the derivatives using the utility functions (6.1), we write the efficiency condition (6.2) as

$$\frac{U^1_x - U^1_G}{U^1_G} = \frac{U^2_x - U^2_G}{U^2_G}. \quad (6.3)$$
The marginal rate of substitution between the private and the public good for consumer \( h \) is defined by \( \text{MRS}^h_{G,x} \equiv \frac{U^h_G}{U^h_x} \). This can be used to rearrange (6.3) in the form

\[
\left[ \frac{1}{\text{MRS}^1_{G,x}} - 1 \right] \left[ \frac{1}{\text{MRS}^2_{G,x}} - 1 \right] = 1.
\]  

(6.4)

Multiplying across by \( \text{MRS}^1_{G,x} \times \text{MRS}^2_{G,x} \), we solve (6.4) and get the final expression

\[
\text{MRS}^1_{G,x} + \text{MRS}^2_{G,x} = 1.
\]  

(6.5)

This is the two-consumer version of the Samuelson rule.

To interpret this rule, the marginal rate of substitution should be viewed as a measure of the marginal benefit of another unit of the public good. The marginal cost of a unit of public good is one unit of private good. Therefore the rule says that an efficient allocation is achieved when the total marginal benefit of another unit of the public good, which is the sum of the individual benefits, is equal to the marginal cost of another unit. The rule can easily be extended to incorporate additional consumers: the total benefit remains the sum of the individual benefits.

Further insight into the Samuelson rule can be obtained by contrasting it with the corresponding rule for efficient provision of two private goods. For two consumers, 1 and 2, and two private goods, \( i \) and \( j \), this is

\[
\text{MRS}^1_{i,j} = \text{MRS}^2_{i,j} = \text{MRT}_{i,j},
\]  

(6.6)

where \( \text{MRT}_{i,j} \) denotes the marginal rate of transformation, the number of units of one good the economy has to given up to obtain an extra unit of the other good. (The \( \text{MRT}_{G,x} \) between public and private good was assumed to be equal to 1 in the derivation of the Samuelson rule in equation 6.5.) The difference between (6.5) and (6.6) arises because an extra unit of the public good increases the utility of all consumers so that the social benefit of this extra unit is found by summing the marginal benefits. This does not require equalization of the marginal benefit of all consumers. In contrast, an extra unit of private good can only be given to one consumer or another. Efficiency then occurs when it does not matter who the extra unit is given to so that the marginal benefits of all consumers are equalized.

The Samuelson rule provides a very simple description of the efficient outcomes, but this does not mean that efficiency is easily obtained. It was already shown that efficiency will not be achieved if there is no government intervention and agents act noncooperatively (i.e., adopt Nash behavior). But what form should government intervention
take? The most direct solution would be for the government to take total responsibility for provision of the public good and to finance it through lump-sum taxation. Because lump-sum taxes do not cause any distortions, this would ensure satisfaction of the rule. However, there are numerous difficulties in using lump-sum taxation, which will be explored in detail in chapter 13. The same shortcomings apply here, thus ruling out the employment of lump-sum taxes. The use of other forms of taxation would introduce their own distortions, and these would prevent efficiency from being achieved. In addition, to apply the Samuelson rule, the government must know the individual benefits from public good provision. In practice, this information is not readily available, and the government must rely on what individuals choose to reveal.

The consequence of these observations is that efficiency will not be attained through direct public good provision if financed by distortionary taxes. Hence we have the motivation for considering alternative allocation mechanisms that can provide the correct level of public good by eliciting preferences from consumers.

6.5 Voting

The failure of private actions to provide a public good efficiently suggests that alternative allocation mechanisms need to be considered. There are a range of responses that can be adopted to counteract the market failure, ranging from intervention with taxation through to direct provision by the government. In practice, the level of provision for public goods is frequently determined by the political process, with competing parties in electoral systems differing in the level of public good provision they promise. The selection of one of the parties by voting then determines the level of public good provision.

We have already obtained a first insight into the provision of public goods by voting in chapter 5. That analysis focused on voting over the tax rate as a proxy for government size when people had different income levels. What we wish to do here is provide a contrast between the voting outcome and the efficient level of public good provision when people differ in tastes and (possibly) income levels. Consider a population of consumers who determine the quantity of public good to be provided by a majority vote. The cost of the public good is shared equally among the consumers, so, if $G$ units of the public good are supplied, the cost to each consumer is $\frac{G}{H}$. With income $M^h$, a consumer can purchase private goods to the value of $M^h - \frac{G}{H}$ after paying for the public good. This provides an effective price of $\frac{1}{H}$ for each unit of the public good and a level of utility $U^h \left(M^h - \frac{G}{H}, G\right)$. The budget constraint, the highest attainable indifference
curves and the most preferred quantity of public good are shown in the upper part of figure 6.6 (assuming, for convenience, the same income levels for all consumers).

So that the Median Voter Theorem can be applied (see chapter 11 for details), assume that there is an odd number, $H$, of consumers, where $H > 2$, and that each of the consumers has single-peaked preferences for the public good. This second assumption implies that when the level of utility is graphed against the quantity of public good, there will be a single value of $G^h$ that maximizes utility for consumer $h$. Such preferences are illustrated in the lower panel of figure 6.6. The consumers are numbered so that their preferred levels of public good satisfy $G^1 < G^2 < \ldots < G^H$.

By these assumptions, the Median Voter Theorem ensures that the consumer with the median preference for the public good will be decisive in the majority vote. The

![Figure 6.6](image-url)

**Figure 6.6**
Allocation through voting
median preference belongs to the consumer at position $\frac{H+1}{2}$ in the ranking. We label the median consumer as $m$ and denote their chosen quantity of the public good by $G^m$. A remarkable feature of the majority voting outcome is that nobody is able to manipulate the outcome to their advantage by misrepresenting their preference, so sincere voting is the best strategy. The reason is that anyone to the left of the median can only affect the final outcome by voting for a quantity to the right of the median that would move the outcome further away from their preferred position, and vice versa for anyone to the right of the median.

Having demonstrated that voting will reveal preferences and that the voting outcome will be the quantity $G^m$, it now remains to ask whether the voting outcome is efficient. The value $G^m$ is the preferred choice of consumer $m$, so it solves

$$\max_{|G|} U^m \left( M^m - \frac{G}{H}, G \right), \quad (6.7)$$

where $M^m$ denotes the income of the median voter that can differ from the median income with heterogeneous preferences. The first-order condition for the maximization can be expressed in terms of the marginal rate of substitution to show that the voting outcome is described by

$$MRS^m = \frac{1}{H}. \quad (6.8)$$

In contrast, because the marginal rate of transformation is equal to 1, the efficient outcome satisfies the Samuelson rule

$$\sum_{h=1}^{H} MRS^h = 1. \quad (6.9)$$

Contrasting these, the voting outcome is efficient only if

$$MRS^m = \sum_{h=1}^{H} \frac{MRS^h}{H}. \quad (6.10)$$

Therefore majority voting leads to efficient provision of the public good only if the median voter's $MRS$ is equal to the mean $MRS$ of the population of voters. There is no reason to expect that it will, so it must be concluded that majority voting will not generally achieve an efficient outcome. This is because the voting outcome does not take account of preferences other than those of the median voter: changing all
the preferences except those of the median voter does not affect the voting outcome (although it would affect the efficient level of public good provision).

Can any comments be offered on whether majority voting typically leads to too much or too little public good? In general, the answer has to be no, since no natural restrictions can be appealed to and the median voter’s $MRS$ may be lower or higher than the mean. If it is lower, then too little public good will provided. The converse holds if it is higher. The only approach that might give an insight is to note that the distribution of income has a very long right tail. If the $MRS$ is higher for lower income voters, then the nature of the income distribution suggests that the median $MRS$ is higher than the mean. Thus voting will lead to an excess quantity of public good being provided. Alternatively, if the $MRS$ is increasing with income, then voting would lead to underprovision.

### 6.6 Personalized Prices

We have now studied two allocation mechanisms that lead to inefficient outcomes. The private market fails because of free-riding, and voting fails because the choice of the decisive median voter need not match the efficient choice. What these have in common is that the consumers face incorrect incentives. In both cases the decision makers take account only of the private benefit of the public good rather than the broader social benefit (i.e., that public good contribution also benefits others). As a rule, efficiency will only be attained by modifying the incentives to align private and social benefits.

The first method for achieving efficiency involves using an extended pricing mechanism for the public good. This mechanism uses prices that are “personalized,” with each consumer paying a price that is designed to fit their situation. These personalized prices modify the actual price in two ways. First, they adjust the price of the public good in order to align social and private benefits. Second, they further adjust the price to capture each consumer’s individual valuation of the public good.

This latter aspect can be understood by considering the differences between public and private goods. With a private good, consumers face a common price but choose to purchase different quantities according to their preferences. In contrast, with a pure public good, all consumers consume the same quantity. This can only be efficient if the consumers wish to purchase the same given quantity of the public good. They can be induced to do so by correctly choosing the price they face. For instance, a consumer who places a low value on the public good should face a low price, while a consumer with a high valuation should face a high price. This reasoning is illustrated in table 6.1.
Table 6.1
Prices and quantities

<table>
<thead>
<tr>
<th></th>
<th>Private good</th>
<th>Public good</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>Same</td>
<td>Different</td>
</tr>
<tr>
<td>Quantity</td>
<td>Different</td>
<td>Same</td>
</tr>
</tbody>
</table>

The idea of personalized pricing can be captured by assuming that the government announces the share of the cost of the public good that each consumer must bear. For example, it may say that each of two consumers must pay half the cost of the public good. Having heard the announcement of these shares, the consumers then state how much of the public good they wish to have supplied. If they both wish to have the same level, then that level is supplied. If their wishes differ, the shares are adjusted and the process repeated. The adjustment continues until shares are reached at which both wish to have the same quantity. This final point is called a Lindahl equilibrium. It can easily be seen how this mechanism overcomes the two sources of inefficiency. The fact that the consumers only pay a share of the cost reduces the perceived unit price of the public good. Hence the private cost appears lower, and the consumers increase their demands for the public good. Additionally the shares can be tailored to match the individual valuations.

To make this reasoning concrete, let the share of the public good that has to be paid by consumer $h$ be denoted $\tau^h$. The scheme must be self-financing, so, with two consumers, $\tau^1 + \tau^2 = 1$. Now let $G^h$ denote the quantity of the public good that household $h$ would choose to have provided when faced with the budget constraint

$$x^h + \tau^h G^h = M^h.$$  \hspace{1cm} (6.11)

The Lindahl equilibrium shares $\{\tau^1, \tau^2\}$ are found when $G^1 = G^2$. The reason why efficiency is attained can be seen in the illustration of the Lindahl equilibrium in figure 6.7. The indifference curves reflect preferences over levels of the public good and shares in the cost. The shape of these captures the fact that each consumer prefers more of the public good but dislikes an increased share. The highest indifference curve for consumer 1 is to the northwest and the highest for consumer 2 to the northeast. Maximizing utility for a given share (which gives a vertical line in the figure) achieves the highest level of utility where the indifference curve is vertical. Below this point the consumer is willing to pay a higher share for more public good, and above it is just the other way around. Hence the indifference curves are backward-bending. The Lindahl
reaction functions are then formed as the loci of the vertical points of the indifference curve. The equilibrium requires that both consumers demand the same level of the public good; this occurs at the intersection of the reactions functions. At this point the indifference curves of the two consumers are tangential and the equilibrium is Pareto-efficient.

To derive the efficiency result formally, note that utility is given by the function $U^h(M^h - \tau^h G^h, G^h)$. The first-order condition for the choice of the quantity of public good is

$$\frac{U^h_G}{U^h_x} = \tau^h, \quad h = 1, 2. \quad (6.12)$$

Summing these conditions for the two consumers yields

$$\frac{U^1_G}{U^1_x} + \frac{U^2_G}{U^2_x} = \text{MRS}^1_{G,x} + \text{MRS}^2_{G,x} = \tau^1 + \tau^2 = 1. \quad (6.13)$$

This is the Samuelson rule for the economy, and it establishes that the equilibrium is efficient. The personalized prices equate the individual valuations of the supply of public goods to the cost of production in a way that uniform pricing cannot. They also correct for the divergence between private and social benefits.
Although personalized prices seem a very simple way of resolving the public good problem, when considered more closely a number of difficulties arise in actually applying them. First, there is the very practical problem of determining the prices in an economy with many consumers. The practical difficulties involved in announcing and adjusting the individual shares are essentially insurmountable. Second, there are issues raised concerning the incentives for consumers to reveal their true demands.

The analysis assumed that the consumers were honest in revealing their reactions to the announcement of cost shares, meaning they simply maximize utility by taking the share of cost as given. However, there will be a gain to any consumer who attempts to cheat, or *manipulate*, the allocation mechanism. By announcing preferences that do not coincide with their true preferences, it is possible for a consumer to shift the outcome in their favor, provided that the other does not do likewise. To see this, assume that consumer 1 acts honestly and that consumer 2 knows this and knows the reaction function of 1. In figure 6.8 an honest announcement on the part of consumer 2 would lead to the equilibrium $e_L$ where the two Lindahl reaction functions cross. However, by claiming their preferences to be given by the dashed Lindahl reaction function rather than the true function, the equilibrium can be driven to point $e_M$, which represents the maximization of 2’s utility given the Lindahl reaction function of 1. This improvement for consumer 2 reveals the incentive for dishonest behavior.
The use of personalized prices can achieve efficiency only if the consumers act honestly. A consumer who acts strategically is able to manipulate the outcome to her advantage. Indeed it is clear that the search for a means of attaining an allocation that satisfies the Samuelson rule should be restricted to allocation mechanisms that cannot be manipulated in this way. This is the focus of the next section.

6.7 Mechanism Design

The previous section showed how consumers have an incentive to reveal false information on demand when personalized prices are being determined. From the consistent application of the assumption of utility maximization we observed that a consumer will behave dishonestly if it is in their interests to do so. This fact has led to the search for allocation mechanisms that are immune from attempted manipulation. As will be shown, the design of some of these mechanisms leads households to reveal their true preferences. Because of this property these mechanisms are called preference revelation mechanisms.

6.7.1 Examples of Preference Revelation

The general problem of preference revelation is now illustrated by considering two simple examples. In both examples people are shown to gain by making false statements of their preferences. If they act rationally, then they will choose to make false statements. Since these situations have the nature of strategic games, we call the participants players.

Example 1: False Understatement

The decision that has to be made is whether to produce or not produce a fixed quantity of a public good. If the public good is not produced, then \( G = 0 \). If it is produced, \( G = 1 \). The cost of the public good is given by \( C = 1 \). The gross benefit of the public good for players 1 and 2 is given by \( v^1 = v^2 = 1 \). Since the social benefit of providing the good is \( v^1 + v^2 = 2 \), which is greater than the cost, it is socially beneficial to provide the public good.

Each player makes a report, \( r^h \), of the benefit they receive from the public good. This report can either be false, in which case \( r^h = 0 \), or truthful so that \( r^h = v^h = 1 \). Based on the reports, the public good is provided if the sum of announced valuations is at least as high as the cost. This gives the collective decision rule to choose \( G = 1 \) if \( r^1 + r^2 \geq C = 1 \), and to choose \( G = 0 \) otherwise. The cost of the public good is shared
between the two players, with the shares proportional to the announced valuations. In detail,

\[ c^h = 1 \text{ if } r^h = 1 \text{ and } r^{h'} = 0, \]  
(6.14)

\[ c^h = \frac{1}{2} \text{ if } r^h = 1 \text{ and } r^{h'} = 1, \]  
(6.15)

\[ c^h = 0 \text{ if } r^h = 0 \text{ and } r^{h'} = 0 \text{ or } 1. \]  
(6.16)

The net benefit, the difference between true benefit and cost, which is termed the payoff from the mechanism, is then given by

\[ U^h = v^h - c^h \text{ if } r^1 + r^2 \geq 1, \]  
(6.17)

\[ = 0 \text{ otherwise.} \]  
(6.18)

This information is summarized in the payoff matrix in figure 6.9.

From the payoff matrix it can be seen that the announcement \( r^h = 0 \) is a weakly dominant strategy for both players. For instance, if player 2 chooses \( r^2 = 1 \), then player 1 will choose \( r^1 = 0 \). Alternatively, if player 2 chooses \( r^2 = 0 \), then player 1 is indifferent between the two strategies of \( r^1 = 0 \) and \( r^1 = 1 \). The Nash equilibrium of the game is therefore \( \hat{r}^1 = 0, \hat{r}^2 = 0 \).

In equilibrium both players will understate their valuation of the public good. As a result the public good is not provided, despite provision being socially beneficial. The reason is that the proportional cost-sharing rule gives an incentive to underreport preferences for public good. With both players underreporting, the public good is not provided. To circumvent this problem, we can make contributions independent of the reports. This is our next example.

![Figure 6.9](image-url)

Announcements and payoffs
Example 2: False Overstatement

The second example is distinguished from the first by considering a public good that is socially nondesirable with a cost greater than the social benefit. The possible announcements and the charging scheme are also changed.

It is assumed that the gross payoffs when the public good is provided are

\[ v_1 = 0 < v_2 = \frac{3}{4}, \]  

(6.19)

With the cost of the public good remaining at 1, these payoffs imply that

\[ v_1 + v_2 = \frac{3}{4} < C = 1, \]  

(6.20)

so the social benefit from the public good is less than its cost.

The possible announcements of the two players are given by \( r_1 = 0 \) or 1 and \( r_2 = \frac{3}{4} \) or 1. These announcements permit the players to either tell the truth or overstate the benefit so as to induce public good provision. Assume that there is also a uniform charge for the public good if it is provided, so \( c^h = \frac{1}{2} \) if \( r_1 + r_2 \geq c = 1 \), and \( c^h = 0 \) otherwise. These valuations and charges imply the net benefits

\[ U^h = v^h - c^h \quad \text{if} \quad r_1 + r_2 \geq 1, \]  

(6.21)

\[ U^h = 0 \quad \text{otherwise}. \]  

(6.22)

These can be used to construct the payoff matrix in figure 6.10.

The weakly dominant strategy for player 1 is to play \( r_1 = 0 \) and the best response of player 2 is to select \( r_2 = 1 \) (which is also a dominant strategy). Therefore the Nash equilibrium is \( \hat{r}_1 = 0, \hat{r}_2 = 1 \), which results in the provision of a socially
nondesirable public good. The combination of payoffs and charging scheme has resulted in overstatement and unnecessary provision. The explanation for this is that the player 2 is able to guarantee the good is provided by announcing \( r^2 = 1 \). Their private gain is \( \frac{1}{4} \), but this is more than offset by the loss of \( -\frac{1}{2} \) for player 1.

### 6.7.2 Clarke–Groves Mechanism

The preceding examples showed that true valuations may not be revealed for some mechanisms linking announcement to contribution. Even worse, it is possible for the wrong social decision to be made. The question then arises as to whether there is a mechanism that will always ensure that true values are revealed (as for voting), and at the same time that the level of public good provided is efficient (which voting cannot do).

The potential for constructing such a mechanism, and the difficulties in doing so, can be understood by retaining the simple allocation problem of the examples that involves the decision on whether to provide a single public good of fixed size. The construction of a length of road or the erection of a public monument both fit this scenario. It is assumed that the cost of the project is known, and it is also known how the cost is allocated among the consumers that make up the population. What needs to be found from the consumers is how much their valuation of the public good exceeds, or falls short of, their contribution to the cost. Each consumer knows the benefit they will gain if the public good is provided, and they know the cost they will have to pay. The difference between the benefit and the cost is called the net benefit. This can be positive or negative. The decision rule is that the public good is provided if the sum of reported net benefits is (weakly) positive.

Consider two consumers with true net benefits \( v^1 \) and \( v^2 \). The mechanism we consider is the following. Each consumer makes an announcement of his or her net benefit. Denote by \( r^h \) the report of \( h \), \( h = 1, 2 \). The public good is provided if the sum of announced net benefits satisfies \( r^1 + r^2 \geq 0 \). If the public good is not provided, each consumer receives a payoff of 0. If the good is provided, then each consumer receives a side payment equal to the reported net benefit of the other consumer; hence, if the public good is provided, consumer 1 receives a total payoff of \( v^1 + r^2 \) and consumer 2 receives \( v^2 + r^1 \). It is these additional side payments that will lead to the truth being told by inducing each consumer to “internalize” the net benefit of the public good for the other. If the public good is not provided, no side payments are made.

To see how this mechanism works, assume that the true net benefits and the reports can take the values of either \(-1\) or \(+1\). The public good will not be provided if both
Chapter 6: Public Goods

report a value of $-1$, but if at least one reports $+1$, it will be provided. The payoffs to the mechanism are summarized in the payoff matrix in figure 6.11. The claim we now wish to demonstrate is that this mechanism provides no incentive to make a false announcement of the net benefit. To do this, it is enough to focus on player 1 and show that the player will report truthfully when $v^1 = -1$ and when $v^1 = +1$. The payoffs relating to the true values are in the two payoff matrices in figure 6.12.

Take the case of $v^1 = -1$. Then consumer 1 finds the true announcement to be weakly dominant—the payoff from being truthful (the top row) is greater if $r^2 = -1$ and equal if $r^2 = +1$. Next take the case of $v^1 = +1$. Consumer 1 is indifferent between truth and nontruth. But the point is that there is now no incentive to provide a false announcement. Hence truth should be expected.

The problem with this mechanism is the side payments that have to be made. If the public good is provided and $v^1 = v^2 = +1$, then the total side payments are equal to

\[
\begin{array}{c|cc}
   \text{Announcement} & 0 & v^2 + 1 \\
   \text{of player 2} & v^1 + 1 & v^2 + 1 \\
   \hline
   -1 & v^2 - 1 & v^1 - 1 \\
   +1 & v^1 + 1 & v^1 + 1 \\
\end{array}
\]

Figure 6.11
Clarke–Groves mechanism

\[
\begin{array}{c|cc}
   \text{Announcement} & 0 & 2 \\
   \text{of player 2} & v^1 = +1 \\
   \hline
   -1 & 0 & 2 \\
   +1 & 0 & 2 \\
\end{array}
\]

Figure 6.12
Payoffs for player 1
which is equal to the total net benefit of the public good. These side payments are money that has to be put into the system to support the telling of truth. Obtaining the truth is possible, but it is costly.

6.7.3 Clarke Tax

The problem caused by the existence of the side payments can be reduced but can never be eliminated. The reason it cannot be eliminated entirely is simply that the mechanism is extracting information, and this can never be done for free. The way in which the side payments can be reduced is to modify the structure of the mechanism.

One way to do this is for side payments to be made only if the announcement of a player changes the social decision. To see what this implies, consider calculating the sum of the announced benefits of all players but one. Whether this is positive or negative will determine a social decision for those players. Now add the announcement of the final player. Does this change the social decision? If it does, then the final player is said to be pivotal, and a set of side payments is implemented that requires taxing the pivotal agent for the cost inflicted on the other agent through the changed social decision. This process is repeated for each player in turn. These side payments are the Clarke taxes that ensure that the correct decision is made so that the public good is produced if it is socially desirable and not otherwise. The use of Clarke taxes reduces the number of circumstances in which the side payments are made.

In a game with only two players, the payoffs for player 1 when the Clarke taxes are used are given by

$$v^1 \text{ if } r^1 + r^2 \geq 0 \text{ and } r^2 \geq 0,$$

$$v^1 - t^1 \text{ if } r^1 + r^2 \geq 0 \text{ and } r^2 < 0, \text{ with } t^1 = -r^2 > 0,$$

$$-t^1 \text{ if } r^1 + r^2 < 0 \text{ and } r^2 \geq 0, \text{ with } t^1 = r^2 \geq 0,$$

$$0 \text{ if } r^1 + r^2 < 0 \text{ and } r^2 < 0.$$

Only in the second and third cases is player 1 pivotal (respectively, by causing provision and stopping provision of the public good), and for these cases a tax is levied on player 1 reflecting the cost to the other agent of changing public good provision ($t^1 = -r^2 > 0$ for the cost of imposing provision, and $t^1 = r^2 \geq 0$ for the cost of stopping provision).

The Clarke taxes induce truth-telling and guarantee that the public good is provided if and only if it is socially desirable. The explanation is that any misreport that changes the decision about the public good would induce the payment of a tax in excess of the
benefit from the change in decision. Indeed, suppose that the public good is socially desirable, so $v^1 + v^2 \geq 0$, but that player 1 dislikes it, so $v^1 < 0$. Then, given an honest announcement from player 2 with $r^2 = v^2$, by underreporting sufficiently to prevent provision of the public good (so $r^1 < -r^2$), player 1 becomes pivotal. Player 1 will have to pay a tax of $t^1 = r^2 = v^2$, which is in excess of the gain from nonprovision, $-v^1$ (since $v^1 + v^2 \geq 0 \Rightarrow v^2 \geq -v^1$). Hence player 1 is better off telling the truth, and given this truth-telling, player 2 is also better off telling the truth (although in this case he is the pivotal agent, inducing provision and paying a tax equal to the damage of public good provision for player 1, $t^2 = -r^1 = -v^1$).

The conclusion is that the Clarke tax induces preference revelation, and by restricting side payments to pivotal agents only, it lowers the cost of information revelation.

### 6.7.4 Further Comments

The theory of mechanism design shows that it is possible to construct schemes that ensure that the truth will be revealed and correct the social decision made. These mechanisms may work, but they are undoubtedly complex to implement. Putting this objection aside, it can still be argued that such revelation mechanisms are not actually needed in practice. Two major reasons can be provided to support this contention.

First, the mechanisms are built on the basis that the players will be rational and precise in their strategic calculations. In practice, many people may not act as strategically as the theory suggests. As in the theory of tax evasion, which we discuss in chapter 17, nonmonetary benefits may be derived simply from acting honestly. These benefits may provide a sufficient incentive that the true valuation is reported. In such circumstances the revelation mechanism will not be needed.

Second, the market activities of consumers often indirectly reveal the valuation of public goods. To give an example of what is meant by this, consider the case of housing. A house is a collection of characteristics, such as the number of rooms, size of garden, and access to amenities. The price that a house purchaser is willing to pay is determined by their assessment of the total value of these characteristics. Equally the cost of supplying a house is also dependent on the characteristics supplied. By observing the equilibrium prices of houses with different characteristics, it is possible to determine the value assigned to each characteristic separately. If one of the characteristics relates to a public good, for example, the closeness to a public park, the value of this public good can then be inferred. Such implicit valuation methods can be applied to a broad range of public goods by carefully choosing the related private good. Since consumers
have no incentive to act strategically in purchasing private goods, the true valuations should be revealed.

The fact that consumers have an incentive to falsely reveal their valuations can also be exploited to obtain an approximation of the true value. This can be done by running two preference revelation mechanisms simultaneously. If one is designed to lead to an underreporting of the true valuation and the other one to overreporting, then the true value of the public good can be taken as lying somewhere between the over- and underreports. The Swedish economist Peter Bohm has conducted an experimental implementation of this procedure. In the experiment 200 people from Stockholm had to evaluate the benefit of seeing a previously unshown television program. The participants were divided into four groups each of which faced the following payment mechanisms: (1) pay stated valuation, (2) pay a fraction of stated valuation such that costs are covered from all payments, (3) pay a low flat fee, and (4) no payment. Although the first two provide an incentive to underreport and the latter two to overreport, the experiment found that there was no significant difference in the stated valuations, suggesting that misrevelation may not be as important as suggested by the theory.

6.8 More on Private Provision

The analysis of the private purchase of a public good in section 6.3 focused on the issue of efficiency. The analysis showed that a Pareto improvement can be made from the equilibrium point if both consumers simultaneously raise their contributions, so the equilibrium cannot be efficient. This finding was sufficient to develop the contrast with efficient provision and to act as a basis for investigating mechanism design.

Although useful, these are not the only results that emerge from the private purchase model. The model also generates several remarkably precise predictions about the effect of income transfers and increases in the number of purchasers. These results are now described and then contrasted with empirical and experimental evidence.

6.8.1 Neutrality and Population Size

The first result concerns the effect of redistributing income. Consider transferring an amount of income $\Delta$ from consumer 1 to consumer 2 so that the income of consumer 1 falls to $M_1 - \Delta$ and that of consumer 2 rises to $M_2 + \Delta$. The objective is to calculate the effect that this transfer has on the equilibrium level of public good purchases. To do this, notice that the equilibrium in figure 6.5 is identified by the fact that it occurs
where an indifference curve for consumer 1 crosses an indifference curve for consumer 2 at right angles. Hence the effect of the transfer on the equilibrium can be found by determining how it affects the indifference curves.

Consider consumer 1 who has their income reduced by $\Delta$. If we reduce their public good purchase by $\Delta$ and raise that of consumer 2 by $\Delta$, the utility of consumer 1 is unchanged because

$$U_1^1(M^1 - g^1, g^1 + g^2) = U_1^1\left(M^1 - \Delta, g^1 - \Delta, g^1 - \Delta\right)$$

This transfer of income causes the indifference curves and the best-reaction function of consumer 1 to move as illustrated in figure 6.13. The indifference curve through any point $g^1, g^2$ before the income transfer shifts to pass through the point $g^1 - \Delta, g^2 + \Delta$ after the income transfer.

The transfer of income has the same effect on the indifference curves and best-reaction function of consumer 2. From the reduction in purchase of consumer 1 and the increase by consumer 2, it follows that

$$U_2^2(M^2 - g^2, g^1 + g^2) = U_2^2\left(M^2 + \Delta, g^1 + \Delta, g^2 + \Delta\right)$$

Figure 6.13
Effect of a transfer
For consumer 2 the indifference curve through $g^1, g^2$ before the income transfer becomes that through $g^1 - \Delta, g^2 + \Delta$ after the transfer.

These shifts in the indifference curves result in the equilibrium moving as in figure 6.14. The point where the indifference curves cross at right angles shifts in the same way as the individual indifference curves. If the equilibrium was initially at $\hat{g}^1, \hat{g}^2$ before the income transfer, it is located at $\hat{g}^1 - \Delta, \hat{g}^2 + \Delta$ after the transfer.

The important result now comes from noticing that in the move from the original to the new equilibrium, consumer 1 reduces his purchase of the public good by $\Delta$, but consumer 2 increases her purchase by the same amount $\Delta$. These changes in the value of purchases exactly match the change in income levels. The net outcome is that the levels of private consumption remain unchanged for the two consumers, and the total supply of the public good is also unchanged. As a consequence the income transfer does not affect the levels of consumption in equilibrium—all it does is to redistribute the burden of purchase. Income redistribution is entirely offset by an opposite redistribution of the responsibility for purchases of the public good. This result, known as *income distribution invariance*, is a consequence of the fact that the utility levels of the consumers are linked via the quantity of a public good.

The second interesting result is that the transfer of income leaves the utility levels of the two consumers unchanged. This has to be so because, as we have just seen,
the consumption levels do not change. Therefore the redistribution of income has not affected the distribution of welfare; the transfer is simply offset by the change in public good purchases. If the income redistribution was due to government policy, this becomes an example of policy neutrality: by changing their behavior, the individuals in the economy are able to undo what the government is trying to do. Income redistribution will always be neutral until the point is reached at which one of the consumers no longer purchases the public good. Only then will further income transfers affect the distribution of utility.

A third result follows easily from income invariance. Let both consumers have the same utility function but possibly different income levels. Since the quantity of public good consumed by both must be the same, the first-order conditions require that both must also consume the same quantity of private good; hence $x^1 = x^2$. Further these common levels of consumption imply that the consumers must have the same utility levels even if there is an initial income disparity. The private-purchase model therefore implies that when the consumers have identical utilities, the choices made by the consumers will equalize utilities even in the face of income differentials. The poor set their purchases sufficiently lower than the purchases of the rich to make them equally well off.

This model can also be used to consider the consequence of variations in the number of households. Maintaining the assumption that all the consumers are identical in terms of both preferences and income, for an economy with $H$ consumers the total provision of the public good is $G = \sum_{h=1}^{H} g^h$ and the utility of $h$ is

$$U^h = U(M - g^h, G) = U(M - g^h, \tilde{G}^h + g^h).$$

Here $\tilde{G}^h$ is the total contributions of all consumers other than $h$. Since all consumers are identical, it makes sense to focus on symmetric equilibria where all consumers make the same contribution. Hence let $g^h = g$ for all $h$. It follows that at a symmetric equilibrium

$$g = \frac{\tilde{G}}{H - 1}. \quad (6.29)$$

In a graph of $g$ against $\tilde{G}$ an allocation satisfying (6.29) must lie on a ray through the origin with gradient $H - 1$. For each level of $H$, the equilibrium is given by the intersection of the appropriate ray with the best-reaction function. This is shown in figure 6.15.

The important point is what happens to the equilibrium level of provision as the number of consumer tends to infinity (the idealization of a “large” population). What
happens can be seen by considering the consequence of the ray in figure 6.15 becoming vertical: the equilibrium will be at the point where the reaction function crosses the vertical axis. As this point is reached, the provision of each consumer will tend to zero, but aggregate provision will not, since it is the sum of infinitely many zeros. This result can be summarized by saying that in a large population each consumer will effectively contribute nothing.

6.8.2 Experimental Evidence

The analysis of private purchase demonstrated that the equilibrium will not be Pareto-efficient and that, compared to Pareto-preferred allocations, too little of the public good will be purchased. A simple explanation of this result can be given in terms of each consumer relying on others to purchase and hence deciding to purchase too little. Each consumer is free-riding on others' purchases, and since all attempt to free-ride, the total value of purchases fails to reach the efficient level. This conclusion has been subjected to close experimental scrutiny.

The basic form of the experiment is to give participants a number of tokens that can be invested in either a private good or a public good. Each participant makes a single purchase decision. The private good provides a benefit only to its purchaser while purchase of public good provides a benefit to all participants. The values are set so that the private benefit is less than the social benefit. The benefits are known to the
participants and the total benefit from purchases is the payoff to the participant at the end of the experiment. It is therefore in the interests of each participant to maximize his or her payoff.

To see how this works in detail, assume that there are 10 participants in the game. Allow each participant to have 10 tokens to spend. A unit of the private good costs 1 token and provides a benefit of 5 units (private benefit = social benefit = 5). A unit of the public good also costs 1 token but provides a benefit of 1 unit to all the participants in the game (private benefit = 1 < social benefit = 10). The returns are summarized in figure 6.16.

If the game is played once (a “one-shot” game), the Nash equilibrium strategy is to purchase only the private good, since each token spent on the private good yields a return five times higher than for the public good. In equilibrium, the total return to each player is 50. In contrast, the socially efficient outcome is for all players to purchase only the public good and to generate a payoff of 100 to each player. The fact that the Nash equilibrium differs from the efficient outcome is because the private benefits diverge from the social benefits. Thus, in the one-shot game, all tokens should be spent on the private good.

In experimental implementations of this game the average value of purchases of the public good has been approximately 30 to 90 percent of tokens, with most observations falling in the 40 to 50 percent range. Interestingly, among student participants contributions have been lowest for those studying economics, and fall with the number of years of economics taken. Since the purchase of the public good is significantly different from 0, these results clearly do not support the predictions of the private-purchase model.

Some experiments have repeated the purchase decision over several rounds, with the view that this should allow time for the participants to learn about free-riding and develop the optimal strategy. The results from such experiments are not as clear and a wider range of purchases occurs. Free-riding is not completely supported, but instances

<table>
<thead>
<tr>
<th></th>
<th>Private good</th>
<th>Public good</th>
</tr>
</thead>
<tbody>
<tr>
<td>Private benefit</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Social benefit</td>
<td>5</td>
<td>10</td>
</tr>
</tbody>
</table>

Figure 6.16
Public good experiment
have been reported in which it does occur. However, this finding should be treated with caution, since having several rounds of the game introduces aspects of repeated game theory. While it remains true that the only credible equilibrium of the repeated game is the private-purchase equilibrium of the corresponding single-period game, it is possible that in the experiments some participants were attempting to establish cooperative equilibria by playing in a fashion that invited cooperation. Additionally those not trained in game theory would have been unable to derive the optimal strategy even though they could solve the single-period game.

Other results show that increasing group size leads to increased divergence from the efficient outcome when accompanied by a decrease in marginal return from the public good, but the results do not support a pure numbers-in-group effect. This finding is compatible with the theoretical finding that the effect of group size on the divergence from optimality is, in general, indeterminate.

These results indicate that there is little evidence of free-riding in single-period, or one-shot, games but in the repeated games the purchases fall toward the private-purchase level as the game is repeated. In total, these experiments do not provide great support for the equilibrium based on the private-purchase model with Nash behavior. In the single-period games free-riding is unambiguously rejected. Although it appears after several rounds in repeated games, the explanation for the strategies involved is not entirely apparent. Neither a strategic nor a learning hypothesis is confirmed. What seems to be occurring is that the participants are initially guided more by a sense of fairness than by Nash behavior. When this fairness is not rewarded, the tendency is then to move toward the Nash equilibrium. The failure of experimentation to support free-riding lends some encouragement to the views that although such behavior may be individually optimal, it is not actually observed in practice.

6.8.3 Modifications

The experimental evidence has produced a number of conflicts with the predictions of the theoretical model. The analysis of private purchase was based on two fundamental assumptions. The utility of consumers was assumed to depend only on the consumption of the private good and the total supply of the public good. This ensures that consumers do not care directly about the size of their own contribution nor do they care about the behavior of other consumers, except for how it affects the total level of the public good. The second assumption was that the consumers acted noncooperatively and played according to the assumptions of Nash equilibrium.
The simplest modification that can be made to the model is to consider the game being played in a different way. The foundation of the Nash equilibrium is that each player takes the behavior of the others as given when optimizing. One way to change this is to consider “conjectural variations” so that each player forms an opinion as to how their choice will affect that of others. If the conjectural variation is positive, each player predicts that the others will respond to an increase in purchase by also making additional purchases. Such a positive conjecture can be interpreted as being more cooperative than the zero conjecture that arises in the Nash equilibrium and leads to the equilibrium having greater total public good supply than the Nash equilibrium.

Moving to non-Nash conjectures may alter the equilibrium level of the public good, but it does not eliminate the neutrality properties. Furthermore the major objection to this approach is that it is entirely arbitrary. There are sensible reasons founded in game theory for focusing on the Nash equilibrium, and no other set of conjectures can appeal to similar justification. If the Nash equilibrium of the private-purchase model does not agree with observations, it would seem that the objectives of the consumers and the social rules they observe should be reconsidered, not the conjectures they hold when maximizing.

One approach to modified preferences is to assume that consumers derive utility directly from the contributions they make. For instance, making a donation to charity can make a consumer feel good about herself; she is acting as a “good citizen.” This is often referred to as the warm glow effect. With a warm glow, a purchase of the public good provides a return from direct consumption of the public good and a further return from the warm glow. The private warm glow effect increases the value of the purchase and so raises the equilibrium level of total purchases. The equilibrium also no longer has the same invariance properties. This would seem a significant advance were it not that the specification of the warm glow is entirely arbitrary.

A final modification is to remove the individualism and allow for social interaction by modifying the rules of social behavior. In the same way that social effects can arise with tax evasion, they can also occur with public goods. One way to do this is to introduce reciprocity, by which each consumer considers the contributions of others and contrasts them to what she feels she should have made. If the contributions of others match, or exceed, what is expected, then the consumer is assumed to feel under an obligation to make a similar contribution. This again raises the equilibrium level of contribution.
6.9 Fund-Raising Campaigns

The model of voluntary purchase that we have considered so far has involved a single one-off contribution decision. It is easy to appreciate that once these contributions have been made, the consumers may look again at the situation and realize that it is inefficient. This could give them an incentive to conduct a second round of contribution that will move the equilibrium closer to efficiency. Repeatedly applying this argument suggests that it may be possible to eventually reach efficiency. We now assess this claim by addressing it within a simple fund-raising game.

The basis of the fund-raising game is that a target level of funds must be achieved before a public good can be provided. For example, consider the target as the minimum cost of construction for a public library. Subscribers to the campaign take it in turn to make either a contribution or a pledge to contribute. Only when the target is met does the process cease. The basic question is whether such a fund-raising campaign can be successful given the possibility of free-riding.

We model a campaign as a game with an infinite horizon, meaning that solicitation for donations can continue until the goal is met. There is one public good (or joint project) whose production cost is \( C \) and two identical players \( X \) and \( Y \). These players derive the same benefit, \( B \), from the public good, so the total benefit is \( 2B \). Both also have the same discount rate \( \delta \), \( 0 < \delta < 1 \), for delaying completion of the project by one period.

The players alternate in making contributions. The sequential (marginal) contributions are denoted \((\ldots, x_{t-1}, y_t, x_{t+1}, \ldots)\), where \( x_{t-1} \) denotes the contribution of player \( X \) at time \( t-1 \) and \( y_t \) denotes the contribution of player \( Y \) at time \( t \). The game ends, and the public good is provided, only when the total contributions cover the cost of the public good. Individuals derive no benefits from the public good before completion of the fund-raising, so the marginal contributions yield no return until the cost is met. It follows that the incentive of each player to wait for the other one to contribute (free-riding) must be balanced against the cost of delaying completion of the project. We suppose that the public good is “socially desirable” \( (C < 2B) \) but that no single player values the public good enough to bear the full cost \( (B < C) \). We now contrast two different forms of fund-raising campaigns. In the first, the contribution campaign, the contributions are paid at the time they are made. In the second, the subscription campaign, players are asked in sequence to make donation pledges that are not paid until the cost is met.
6.9.1 The Contribution Campaign

In the contribution campaign, contributions are sunk at the time they are made because a credible commitment cannot be made to make contributions later. The lack of commitment leads each player to back his contribution to ensure that the other players contribute their share. This is because past contributions are sunk and cannot influence the division of the remaining cost between the players. As a result we show that it is never possible to raise the money, even though the project is worthwhile.

The two players are asked in sequence to make a contribution. While there is no natural end period, there is a total contribution level that is close enough to the cost \( C \) that the contributor whose turn it is should complete the fund-raising rather than waiting for the other one to make up the difference. Suppose that it is player \( X \)’s turn to make a contribution offer at that final round \( T \). There exists a deficit sufficiently small that player \( X \) is indifferent between making up the difference and getting a payoff of \( B - x_T \) or between waiting in the expectation (at best) that player \( Y \) will make up the difference in the next round and producing a payoff with delayed completion of \( \delta B \). Hence the maximal contribution of player 1 in the final round \( T \) is

\[
x_T = (1 - \delta)B,
\]

so the contribution is equal to the benefit of speeding up completion of the project.

We suppose that \([1 - \delta] B < C \) so that such a contribution cannot cover the full cost and a donation from player \( Y \) must be solicited. Working backward, it is now player \( Y \)’s turn to make a contribution at time \( T - 1 \). Player \( Y \) anticipates that in bringing (total) contributions up to \( C - x_T \) at date \( T - 1 \), player \( X \) will complete the project the next period. So there exists a sufficiently small deficit such that player \( Y \) is indifferent between bringing total contributions up to that level, giving a payoff \( \delta B - y_{T-1} \), and waiting for the other player to make such contribution while himself making the final contribution \( x_T \), which produces a payoff \( \delta^2 [B - x_T] \) (i.e., two periods later you get the completed project benefit \( B \) and pay the last contribution \( x_T \)). Hence, substituted for \( x_T \), the contribution at time \( T - 1 \) that makes player \( Y \) indifferent is

\[
y_{T-1} = \delta \left(1 - \delta^2 \right) B.
\]

Proceeding backward to date \( T - 2 \), it is now the turn of player \( X \) to make a contribution. Using the same line of argument, there exists a total contribution level at date \( T - 2 \) such that player \( X \) is indifferent between bringing total contribution up to that level to get a payoff \( \delta^2 [B - x_T] - x_{T-2} \) from completion in two periods or waiting and delaying
completion to get a payoff $\delta^3 B - \delta^2 y_{T-1}$ (in which from the switching position it becomes worthwhile to contribute $y_{T-1}$). Substituting for $x_T$ and $y_{T-1}$ gives

$$x_{T-2} = \delta^3 \left[ 1 - \delta^2 \right] B.$$  \hfill (6.32)

Moving back to round $T-3$ and following the same reasoning, the potential contribution at time $T - 3$ from player $Y$ is

$$y_{T-3} = \delta^5 \left[ 1 - \delta^2 \right] B,$$  \hfill (6.33)

and the potential contribution at time $T - 4$ is

$$x_{T-4} = \delta^7 \left[ 1 - \delta^2 \right] B.$$  \hfill (6.34)

Going back further, it is possible to calculate how much each player is willing to contribute at each stage. This is illustrated in figure 6.17.

Summing these contributions by starting from the end of the campaign, we have the total potential for contributions as

$$[1 - \delta]B + \delta \left[ 1 - \delta^2 \right] B + \delta^3 \left[ 1 - \delta^2 \right] B + \delta^5 \left[ 1 - \delta^2 \right] B + \delta^7 \left[ 1 - \delta^2 \right] B + \ldots = B.$$  \hfill (6.35)

---

**Figure 6.17**

Two-player contribution campaign
In (6.35) we used the geometric progression fact that \(1 + \delta^2 + \delta^4 + \delta^6 + \ldots = \frac{1}{1-\delta^2}\). The remarkable feature is that the total potential for contributions never exceeds the individual benefit from the project, and because \(B < C\), it is not possible to raise sufficient contributions for a successful campaign.

6.9.2 The Subscription Campaign

In the subscription game, agents alternate in making donation pledges and bear the cost of their contribution only when and if enough contributions are pledged to complete the project. In a sense, agents are able to make certain conditional commitments to contribute in the future. This possibility to commit modifies the strategic structure of the game and alters the total amount that can be raised. As we now show, in this case it becomes possible to raise an amount equal to the total valuation of all the contributors.

Once again, we start when the fund-raising operation is over and work backward. Fix an arbitrary end point \(T\) with player \(X\)’s turn to make a donation pledge at date \(T\). There must exist a contribution deficit sufficiently small to make player \(X\) indifferent between financing the deficit himself to obtain a payoff \(B - x_T\) and waiting for player \(Y\) to make up the difference in the next period, with a delayed completion payoff of \(\delta B\). So the potential pledge of player \(X\) at date \(T\) is

\[x_T = [1 - \delta]B.\] (6.36)

We continue to assume that \([1 - \delta]B < C\) so that we can solicit player \(Y\)’s donation. Working backward, it is then up to player \(Y\) to pledge at date \(T - 1\). Player \(Y\) anticipates that in bringing the total amount pledged up to \(C - x_T\) at date \(T - 1\), player \(X\) will complete the project in the next period. So there exists a sufficiently small deficit such that player \(Y\) is indifferent between making up the difference to get a payoff \(\delta [B - y_{T-1}]\) and leaving player \(X\) to make up the difference, thereby delaying completion to get a payoff of \(\delta^2 [B - x_T]\) (in which case it becomes worthwhile for \(Y\) to pledge himself \(x_T\) at date \(T\)). Hence, substituting for \(x_T\), we obtain

\[y_{T-1} = [1 - \delta^2]B.\] (6.37)

Going back to date \(T - 2\), it is up to player \(X\) to pledge. Again, there exists a total amount pledged close enough to \(C - x_T - y_{T-1}\) such that player \(X\) is indifferent between bringing the total contribution up to that level, anticipating completion in two rounds with a payoff \(\delta^3 [B - x_T - x_{T-2}]\), and waiting for \(Y\) to pledge instead with a payoff from switching position of \(\delta^3 [B - y_{T-1}]\). Substituting for \(x_T\) and \(y_{T-1}\) gives
Proceeding likewise, we can go back further and calculate how much player Y will pledge at date $T - 3$ as

$$y_{T-3} = \delta^2 [1 - \delta^2] B,$$

and player X will pledge at date $T - 4$ the amount

$$x_{T-4} = \delta^3 [1 - \delta^2] B.$$  

Going back to calculate how much each player is willing to pledge at each stage and summing up potential pledges, we get

$$[1 - \delta] B + [1 - \delta^2] B + \delta [1 - \delta^2] B + \delta^2 [1 - \delta^2] B + \delta^3 [1 - \delta^2] B + \ldots = 2B.\quad (6.41)$$

This is the maximum amount that can be raised and is equal to the total valuations of all the contributors. Hence it is always possible to raise enough money for any worthwhile project because $C < 2B$.

These results have shown how allowing contributions to be repeated may lead to efficient private provision of the public good. But this conclusion is sensitive to the assumptions made on the ability of contributors to make binding commitments.

### 6.10 Conclusions

This chapter has reviewed the standard analysis of the efficient level of provision of a public good leading to the Samuelson rule. The analysis of private purchase emphasized the fact that this outcome will not be achieved without government intervention. The efficiency rule describes an allocation that can only be achieved if the government is unrestricted in its policy tools or, as the Lindahl equilibrium demonstrates, using prices that are personalized for each consumer.

One aspect of public goods that prevents the government from making efficient decisions is the government’s lack of knowledge of consumers’ preferences and their willingness to pay for public goods. Mechanisms were constructed that provide the right incentives for consumers to correctly reveal their true valuation of the public good. Experimental evidence suggests that consumer behavior when confronted with
decision problems involving public goods does not fully conform with the theoretical prediction and that the private-purchase equilibrium may not be as inefficient as theory suggests. Furthermore misrevelation has not been confirmed as the inevitable outcome.

Further Reading

The classic paper on the efficient provision of public goods is:

The private provision model is developed fully in:

The independence between income distribution and public good allocation is in:

Further developments of the model are in:


The effect of group size on private provision is in:


The effect of altruism on private provision is in:

Preference revelation for public goods was first described as a dominant strategy mechanism in:

A simple mechanism for preference revelation as a Nash equilibrium is the “round table” scheme in:

There is also a mechanism that induces truth-telling as a Bayesian–Nash equilibrium in:

A very good survey of the preference revelation mechanisms is in:

The fund-raising campaign is based on private provision of discrete public good in:

More on private provision of discrete public goods (such as the volunteer dilemma) is in:

Experimental results are surveyed in:


### Exercises

**6.1** Which of the following are public goods? Explain why.

- a. Snowplowing services during the winter.
- b. A bicycle race around France during the summer.
- c. Foreign aid to Africa to feed its famine-stricken people.
- d. Cable television programs.
- e. Radio programs.
- f. Back roads in the country.
- g. Waste collection services.
- h. Public schools.

**6.2** What are their features with respect to the properties of rivalry and excludability?

**6.3** How does a nonrival good differ from a nonexcludable good?
6.4 In the United Kingdom the lifeboat service is funded by charitable donations. How can this work? How are the rescue services funded in other countries?

6.5 Discuss how television technology can turn a public good into a private good.

6.6 What is a public good? How can one determine the efficient level of provision of a public good?

6.7 Let each dollar spent on a private good give you 10 units of utility but each dollar spent on a public good give you and your two neighbors 5 units each. If you have a fixed income of $10, how much would you spend on the public good? What is the value of the total purchases at the Nash equilibrium if your neighbors also have $10 each? What level of expenditure on the public good maximizes the total level of utility?

6.8 How many allocations satisfy the Samuelson rule?

6.9 How do prices ensure that the efficiency condition is satisfied for private goods? Why is the same not true when there is a public good?

6.10 Consider two consumers with the following demand functions for a public good:

\[ p_1 = 10 - \frac{1}{10} G, \]
\[ p_2 = 20 - \frac{1}{10} G, \]

where \( p_i \) is the price that \( i \) is willing to pay for quantity \( G \).

a. What is the optimal level of the public good if the marginal cost of the public good is $25?  
b. Suppose that the marginal cost of the public good is $5. What is the optimal level?  
c. Suppose that the marginal cost of the public good is $40. What is the optimal level? Should the consumers make an honest statement of their demand functions?

6.11 There are three consumers of a public good. The demands for consumers are as follows:

\[ p_1 = 50 - G, \]
\[ p_2 = 110 - G, \]
\[ p_3 = 150 - G, \]

where \( G \) measures the number of units of the good and \( p_i \) the price in dollars. The marginal cost of the public good is $190.

a. What is the optimal level of provision of the public good? Illustrate your answer with a graph.  
b. Explain why the public good may not be supplied at all because of the free-rider problem.  
c. If the public good is not supplied at all, what is the size of the deadweight loss arising from this market failure?

6.12 Take an economy with 2 consumers, 1 private good, and 1 public good. Let each consumer have an income of \( M \). The prices of public and private good are both 1. Let the consumers have utility functions
Part III: Departures from Efficiency

\[ U^A = \log(x^A) + \log(G), \quad U^B = \log(x^B) + \log(G). \]

a. Assume that the public good is privately provided so that \( G = g^A + g^B \). Eliminating \( x^A \) from the utility function using the budget constraint, show that along an indifference curve

\[ dg^A \left[ \frac{1}{g^A + g^B} - \frac{1}{M - g^A} \right] + dg^B \left[ \frac{1}{g^A + g^B} \right] = 0, \]

and hence that\( pt \)

\[ \frac{dg^B}{dg^A} = \frac{g^A + g^B}{M - g^A} - 1. \]

Solve the last equation to find the locus of points along which the indifference curve of A is horizontal, and use this to sketch the indifference curves of A.

b. Consider A choosing \( g^A \) to maximize utility. Show that the optimal choice satisfiesp\( pt \)

\[ g^A = \frac{M}{2} - \frac{g^B}{2}. \]

c. Repeat part b for B, and calculate the level of private provision of the public good.

d. Calculate the optimal level of provision for the welfare function

\[ W = U^A + U^B. \]

Contrast this with the private provision level.

6.13 Let there be \( H \) consumers all with the utility function \( U^h = \log(x^h) + \log(G) \) and an income of 1. Noting that the utility with private purchase can be written

\[ U^h = \log(x^h) + \log(g^h + \sum_{h' 
eq h} g^{h'}), \]

and that the equilibrium must be symmetric, calculate the private purchase equilibrium and the social optimum for the welfare function

\[ W = \sum_{h=1}^{H} U^h. \]

Comment on the effect of changing \( H \) on the contrast between the equilibrium and the optimum.

6.14 Assume there are two consumers with incomes \( M^1 \) and \( M^2 \). The preferences of consumer \( h, h = 1, 2 \), are described by the utility function \( U = \log(x^h) + \log(r^1 + g^2) \), where \( x^h \) is consumption of the private good and \( g^h \) contribution to the public good. The consumers must determine how to allocate income between the private good and contribution to the public good. Each consumer takes the other consumer’s decision as given.

a. Assuming that the solution to the contribution game is interior \( (x^h > 0, g^h > 0, h = 1, 2) \), find the individual contributions.
b. Show that the total contribution $g_1 + g_2$ depends only on the sum, $M_1 + M_2$, of income.

c. Contrast the outcome of the contribution game to outcome that maximizes a utilitarian social welfare function.

Now let $M_2 = \bar{M} - M_1$, for given $\bar{M}$.

d. Find the limits on $M_1$ as function of $\bar{M}$ for which the solution is interior.

e. Find the solution inside and outside of these limits. Plot the amount of public good as a function of $M_2$.

f. Derive the level of welfare as a function of $M_2$. Show that the function is kinked, and increases (locally) on one side of each kink point.

6.15 \text{(Pareto-improving transfers; Cornes and Sandler 2000). Let there be one consumer with income $M$ and $H$ consumers with income $m$, where $m < M$. All consumers have preferences described by $U = \log(x^h) + \log(G)$, where $x^h$ is consumption of a private good and $G = \sum_{h=1}^{H+1} g_h$ is total contribution to a public good.}

a. Find the condition for the consumers with income $m$ not to contribute.

b. Derive the condition for a transfer of income $dm$ from each consumer with income $m$ to the consumer with income $M$ to raise welfare.

c. Explain how a Pareto improvement can arise when $H$ consumers have their incomes reduced.

6.16 Consider two consumers (1, 2), each with income $M$ to allocate between two goods. Good 1 provides 1 unit of consumption to its purchaser and $\alpha$, $0 \leq \alpha \leq 1$, units of consumption to the other consumer. Each consumer $i$, $i = 1, 2$, has the utility function $U_i = \log(x_i^1) + \log(x_i^2)$, where $x_i^1$ is consumption of good 1 and $x_i^2$ is consumption of good 2.

a. Provide an interpretation of $\alpha$.

b. Suppose that good 2 is a private good. Find the Nash equilibrium levels of consumption when both goods have a price of 1.

c. By maximizing the sum of utilities, show that the equilibrium is Pareto-efficient if $\alpha = 0$ but inefficient for all other values of $\alpha$.

d. Now suppose that good 2 also provides 1 unit of consumption to its purchaser and $\alpha$, $0 \leq \alpha \leq 1$, units of consumption to the other consumer. For the same preferences, find the Nash equilibrium and show that it is efficient for all values of $\alpha$.

e. Explain the conclusion in part d.

6.17 Consider four students deciding to jointly share a textbook. Describe a practical method for using the Lindahl equilibrium to determine how much each should pay.

6.18 Let there be two identical consumers. What would be the share of the cost each should pay for a public good at the Lindahl equilibrium? Use this result to argue that there must be a subsidy to the price of the public good that makes the private purchase equilibrium efficient.

6.19 Assume there are two consumers with preferences described by $U = \log(x) + \log(G)$. Both consumers have income $M$. The government asks each consumer to announce their demand for the public good as a function of the share of the cost they pay. The cost shares are chosen so that both consumers demand the quantity of public good in equilibrium. Consumer 1 pays
share $\tau_1$ and consumer 2 pays share $\tau_2$, with $\tau_1 + \tau_2 = 1$. The government insists that the demand function is linear, and consumers announce the intercept they think is best. Hence consumer $i$ announces the value of $a_i$ in the demand function $G = a_i - b\tau_i$.

a. Determine the efficient level of public good provision.

b. Given a pair of announcements $G = a_1 - b\tau_1$ and $G = a_2 - b\tau_2$, find the resulting values of $G$ and of $\tau_1$ and $\tau_2$ as functions of $a_1$, $a_2$, and $b$.

c. Express the utility of each consumer as a function of $a_1$ and $a_2$. Solve for the Nash equilibrium in announcements and derive the equilibrium value of $G$.

d. How is the equilibrium value of $G$ affected by changes in $b$?

6.20 What would be the equilibrium outcome if both consumers tried to manipulate the Lindahl equilibrium?

6.21 Discuss the effect that an increase in the number of consumers involved in a mechanism has on the consequences of manipulation.

6.22 Consider a two-good economy (one private good and one public good) and a large number $H$ of individuals with single-peaked preferences for the public good. Suppose that the provision of the public good is decided by majority voting, and that it costs one unit of private good to produce one unit of public good. The cost is equally divided among the $H$ individuals. Show that the majority voting outcome is Pareto-efficient if the median marginal rate of substitution is equal to the average marginal rate of substitution.

6.23 Consider a collective decision by three individuals to produce, or not, one public good that costs $150. Suppose that if the public good is produced, the cost is equally shared among the three individuals, namely $50 each. Assume that the gross benefits from the public good differ among individuals and are respectively $20, $40, and $100 for individuals 1, 2, and 3. Each individual is asked to announce his own benefit for the public good, and the public good is produced only if the sum of reported benefits exceeds the total cost.

a. Show that the Groves–Clarke tax induces truth-telling as a dominant strategy if each individual reports independently his own benefit.

b. Show that the resulting provision of public good is optimal.

c. Show that the Groves–Clarke tax is not robust to collusion in the sense that two individuals could be better off by jointly misreporting their benefit from the public good.

d. What would be the provision of public good if the decision were taken by a majority vote, assuming that the cost is equally shared in the event of public good provision? Compare your answer with part b, and interpret the difference.

6.24 Consider three consumers ($i = 1, 2, 3$) who care about their consumption of a private good and their consumption of a public good. Their utility functions are respectively $u_1 = x_1 G$, $u_2 = x_2 G$, and $u_3 = x_3 G$, where $x_i$ is consumer $i$’s consumption of private good and $G$ is the amount of public good jointly consumed by all of them. The unit cost of the private good is $1$ and the unit cost of the public good is $10$. Individual wealth levels in are $w_1 = 30$, $w_2 = 50$, and $w_3 = 20$. What is the efficient amount of public good for them to consume?

6.25 Albert and Beth are thinking of buying a sofa. Albert’s utility function is $u_a(s, m_a) = [1 + x] m_a$ and Beth’s utility function is $u_b(s, m_b) = [2 + x] m_b$, where $s = 0$ if they don’t
get the sofa and \( s = 1 \) if they do, and \( m_a \) and \( m_b \) are the amounts of money they have respectively to spend on private consumption. Albert and Beth each have a total of \( w = 100 \) (in $) to spend. What is the maximum amount that they could pay for the sofa and still both be better off than without it?

6.26 Are the following statements true or false? Explain why.

a. If the supply of public good is determined by majority vote, then the outcome must be Pareto-efficient.

b. If preferences are single-peaked, then everyone will agree about the right amount of public goods to be supplied.

c. Public goods are those goods that are supplied by the government.

d. The source of the free-rider problem is the absence of rivalry in the consumption of public goods.

e. The source of the preference revelation problem is the nonexcludability of public goods.

f. If a public good is provided by voluntary contributions, too little will be supplied relative to the efficient level.

6.27 Why does the free-rider problem make it difficult for markets to provide the efficient level of public goods?

6.28 Four people are considering whether to hire a boat for a day out. Describe questions that will elicit over- and undervaluations of the boat hire.

6.29 People are observed traveling a long distance to visit a scenic countryside. How can this fact be used to place a lower bound on their valuation of the countryside?
7 Club Goods and Local Public Goods

7.1 Introduction

One of the defining features of the public goods of chapter 6 was nonrivalry: once the good is provided, its use by one consumer does not affect the quantity available for any other consumer. This is clearly an extreme assumption. Many commodities, such as parks, roads, and sports facilities, satisfy nonrivalry to a point but are eventually subject to congestion. Although not pure public goods, these goods cannot be classed as private goods either.

A good that has some degree of nonrivalry but for which excludability is possible is called a **club good**. The name is intended to reflect the fact that there are benefits to groups of consumers forming a club to coordinate provision and that the group size may be less than the total population. The name also captures the fact that the clubs we observe in practice are formed by groups of consumers to coordinate the provision of such goods. For instance, a tennis club provides courts that are excludable and nonrival for users at different times. International bodies, such as NATO, can also be interpreted as clubs: NATO provides defense for its members, which is again partly nonrivalrous and partly excludable (only partly because if the existence of NATO deters aggression generally, nonmembers will also benefit).

In our description of economic activity in the previous chapters we did not pay any attention to the geography of trade. In effect we assumed that there is either a single marketplace with consumers located close to it or that travel to markets is costless. It is a fact of actual economic activity that consumers and markets are dispersed, and that travel costs can be significant. As a consequence public goods provided in a particular geographical location need not be available except for those in the close vicinity. For instance, radio and television signals can only be received within range of the transmitter and a police service may only patrol a limited jurisdiction. Provided that a consumer is located within the relevant area, he can benefit from the public good; otherwise, the public good is unavailable to that consumer because the cost of traveling to enjoy it exceeds the benefit. Such goods are again not pure public goods as defined in chapter 6 and are termed **local public goods**, with the name capturing the idea of geographical restriction. The geographical restriction on availability can also be accompanied by congestion within the region.
The issues that the chapter addresses are similar to those involved with pure public goods. It begins by defining club goods and local public goods and investigating the relationships between them. The efficiency question is addressed next for single-product clubs and is related to the charging scheme required to support efficiency. The clubs are then placed within an economy to consider whether efficiency is achieved at this level. Local public goods are introduced, and the efficiency question is again addressed. The extension is subsequently made to consider heterogeneous consumers, which leads into a discussion of the influential Tiebout hypothesis of preference matching for local public goods. The chapter is completed by a review of the empirical evidence on this hypothesis.

7.2 Definitions

The purpose of this section is to provide precise definitions of the classes of goods under discussion. Once this is done, it is possible to describe how these classes are related.

The essential aspect of a club good is that it is possible for those who pay for its provision to exclude those who do not. This is in contrast to the pure public good, which was defined by the impossibility of exclusion. In addition, club goods are often assumed to suffer from congestion, but this is not strictly necessary. However, congestion provides a motive for exclusion and for the forming of a club to supply the good.

A formal definition can be given as follows:

**Definition 7.1** (Club good) A club good is a good that is either nonrivalrous or partly rivalrous but for which exclusion by the providers is possible.

The exclusion aspect of a club good can be taken literally, such as a check on membership credentials at the door to the club, or taken as representing some more general legal authority to bar nonmembers. Its consequence is that issues of preference revelation are not important for club goods. The benefits of the club can only be obtained by voluntarily choosing to become a member, and doing so immediately reveals preferences. This observation is clearly important for the potential attainment of efficiency by the market.

The defining feature of a local public good is one of geography and the need to locate within a specific geographical area in order to benefit from the good. Once outside this
area, the benefit of the good is no longer obtained. This geographical constraint may also be linked with congestion, which causes partial rivalry.

**Definition 7.2** (Local public good) A local public good can only benefit those within a given geographical area. It may be nonrivalrous within that area or it may be partially rivalrous.

This definition of a local public good makes clear that the unique feature is the geographic restriction. It leaves open the question of whether or not a local public good is excludable. This is important for the following reason: as will be seen, the focus of local public good theory is the analysis of local government and decisions on taxation and expenditure. Whether the local public goods provided are excludable then becomes a matter of policy rather than an inherent feature of the good. By this, it is meant that local governments can use a variety of regulations to control access to the public goods they offer. As examples, registration at schools can be restricted by policy choice to pupils in the local area and the size of the local population can be controlled by prohibition on new building. Another example is immigration policy that aims to limit access to national public goods to native residents.

Consequently there are large overlaps between clubs and local public goods, and the terms have often been used interchangeably. What has mostly distinguished the two in the literature has been the issues that have been addressed using each concept. The discussion of club goods has focused more on issues of efficiency with homogeneous populations. In contrast, local public goods have found their most prominent use in the analysis of heterogeneous populations and preference revelation. Furthermore local public goods have been used to understand the role and structure of local government whereas club goods have been more about the market. Even these distinctions are not always binding.

### 7.3 Single-Product Clubs

The analysis of efficiency for a pure public good involves determining how much of it should be provided. With a club good, it is not just the quantity of the good that needs to be decided but also the size of the club membership. The latter is important because of the effect of congestion. Adding a new member allows the cost of providing a given quantity of public good to be spread among more members but reduces the benefit obtained by each existing member. With a club good that suffers from
congestion, there is a second efficiency condition involved concerning the correct level of membership.

7.3.1 Fixed Utilization

Consider now the simplest model of a club. There is a homogeneous population of consumers who are identical in terms of tastes and income. One private good is available and one club good. The club good can potentially suffer from congestion. The focus of attention is on the decision of a single club. It is assumed that a club has formed with the intention of supplying the club good (imagine a small committee of founder members setting out its constitution) and is now in the process of deciding how much of the good to supply and how many members to admit.

To complete the description of the decision problem, it is necessary to consider the financing of the club. Since the club has the ability to exclude nonmembers, it is able to charge members for the privilege of membership. Unlike a pure public good, there is then no barrier to financing provision of the club good provided that enough potential members are willing to pay for membership. The most natural assumption to make on the method of charging is that the cost of the club be divided equally among the members. This charging policy will ensure that the club just breaks even.

Let each consumer have the utility function $U(x, G, n)$, where $x$ is the consumption of a private good, $G$ provision of the club good, and $n$ the number of club members. Utility increases in $x$ and $G$, and decreases in $n$ if there is congestion. If the cost of providing $G$ units of the club good is $C(G)$, then the budget constraint of a member with income $M$ when the cost of the club is shared equally between members will be

$$M = x + \frac{C(G)}{n}. \quad (7.1)$$

The decision problem for those in charge of the club involves choosing $G$ and $n$ to maximize the welfare of a typical member. Putting together the budget constraint and the utility function, this can be expressed as

$$\max_{G, n} U\left(M - \frac{C(G)}{n}, G, n\right). \quad (7.2)$$

The first-order conditions for this optimization produce the following pair of equations that characterize efficiency:
The first of these conditions, \((7.3)\), is a version of the Samuelson rule (6.5) and describes the level of public good, \(G\), that the club should supply. It states that the sum of marginal rates of substitution between the public good and the private good for the \(n\) identical members of the club should be equated to the marginal rate of transformation (or the marginal cost), \(C_G\), of another unit of the club good. What it is most important to observe from this condition is that the process of decision-making within the club ensures that this efficiency condition is satisfied. The ability to exclude nonmembers from consuming the club good permits the club to achieve the correct level of provision. A club therefore achieves efficient public good provision for its members.

To interpret \((7.4)\), it should first be noted that \(U_n \leq 0\). If there is congestion, \(U_n < 0\) and an increase in the number of club members for a given level of provision will reduce the utility of each through congestion effects. We can treat \(\frac{U_n}{U_x}\) as the marginal utility cost of another member of the club. This marginal utility cost is equated to the extent to which another club member reduces the share of the cost for each existing member.

With \(U_n < 0\), \((7.4)\) will determine an efficient level of membership for the club that is positive and finite. Again, the club will achieve efficiency through its internal decision-making. In the absence of congestion \(U_n = 0\), so the efficient club membership will be infinite. In practice, this can be interpreted as the club encompassing the entire population. However, in contrast to the pure public good, the ability to exclude permits the levy of a membership fee that can finance the cost of the club. The club therefore achieves an efficient level of membership.

The arguments to this point can be summarized as follows. A club is able to exclude nonmembers from consumption of the public good and can levy a charge on members. If all consumers are identical, then the club will achieve an efficient level of the club good and an efficient level of membership. If the club good suffers from congestion, then the membership will be restricted. Without congestion, the entire population will be members of the club. The collection of membership fees by the club ensures that it breaks even in its financing of the provision of the club good. This fundamental insight that clubs can attain efficiency in the provision of public goods is attributed to the seminal work of Buchanan (1965) who was the first to develop the theory of clubs.
In terms of the earlier discussion, Buchanan observed that joining a club constitutes an act of preference revelation that permits the attainment of efficiency.

### 7.3.2 Variable Utilization

The model of the club used above does not probe too deeply into the nature of the good that the club supplies. When this is considered further, it becomes apparent that it is not the number of club members that matters for congestion but how frequently the facilities of the club are used. Retaining the assumption that all club members are identical, the total use of the club is equal to the product of the number of members and the number of visits that each member makes to the club. In determining its provision, a club will wish to optimize the number of visits in addition to the size of facility and the membership.

The model can be easily extended to incorporate a variable rate of visitation into the analysis. Let \( v \) be the number of visits that each member makes to the club. An increase in the number of visits raises the utility of the member making those visits but causes congestion through the total number of visits of all members. Letting the total number of visits be \( V = nv \), the utility function is written \( U = U(x, G, v, V) \), with the marginal utility to a visit, \( U_v \), positive and the marginal congestion effect, \( U_V \), negative. The cost function for providing the club is also modified to make it dependent on the total number of visits, \( nv \).

With this extension the optimization problem for the club becomes

\[
\max_{\{x, G, v, n\}} U(x, G, v, V) \quad \text{subject to} \quad M = x + \frac{C(G, nv)}{n}. \tag{7.5}
\]

The necessary condition for efficient provision of the public good by the club is

\[
n \frac{U_G}{U_x} = C_G. \tag{7.6}
\]

Condition (7.6) is again the Samuelson rule for the club equating the sum of marginal rates of substitution to the marginal cost of provision. The necessary condition for efficient club membership is

\[
v \frac{U_V}{U_x} = -\frac{C}{n^2} + \frac{vC_V}{n}. \tag{7.7}
\]

In this condition, \( v \frac{U_V}{U_x} \) is the marginal loss of utility through the congestion caused by an additional club member. This is equated to the reduction in cost through increased
Chapter 7: Club Goods and Local Public Goods

membership, \(-\frac{C}{n^*}\), offset by the increased cost of servicing additional visits, \(\frac{\nu C_V}{n^*}\). The third efficiency condition determines the number of visits to the club that each member should make. This is given by

\[
\frac{U_v}{U_x} = C_V - n \frac{U_V}{U_x},
\]

which equates the marginal benefit of an additional visit to the marginal maintenance cost plus the marginal congestion cost an extra visit imposes on all members of the club.

As with the case of fixed visits, if the decision-making of the club is guided by these three efficiency conditions, then it will ensure an efficient allocation of resources for its members. It will accept the correct number of members, provide the correct quantity of public good, and set visit levels correctly. Therefore introducing a variable visitation rate does not affect the basic conclusion that clubs will supply excludable public goods efficiently.

However, there is a very important distinction between the cases of variable and fixed utilization. This analysis of variable utilization retained the assumption that there is a fixed charge for membership but no further charges for visits. Consequently, once someone has become a member of the club, the price for each additional visit is zero. In choosing visits, each member will only take account of the private cost of the increase in congestion and not the cost they impose on other members. Therefore they will make an excessive number of visits to the club. In brief, the fixed charge does not impose the correct incentives on members to decentralize the efficient outcome. To implement the optimum defined, it is therefore necessary for a club charging a fixed fee to directly regulate the number of visits. This is rather strong restriction on the behavior of the club and motivates the study of an alternative pricing scheme.

7.3.3 Two-Part Tariff

To provide a starting point for the study of a more sophisticated pricing scheme, it is worth formalizing the final comments of the previous subsection. Assume that the club has chosen its efficient provision, \(G^*\), membership, \(n^*\), and visits, \(v^*\), and that its membership fee, which is based on all members abiding by the number of visits, is given by \(F^* = \frac{C(G^*, n^*v^*)}{n^*}\). Now consider the incentives facing a member of the club who believes all other members will make \(v^*\) visits. Putting together the budget constraint, \(M = x + F^*\), and the utility function, the club member faces the optimization
max \( U(M - F^*, G^*, v, [n^* - 1] v^* + v) \). \hfill (7.9)

The choice of \( v \), taking the choices of \( G^* \), \( [n^* - 1] v^* \), and \( F^* \) as given, then satisfies the necessary condition

\[
U_v + U_V = 0. \hfill (7.10)
\]

Consequently the member will choose to make visits to the point at which the marginal utility of visits is completely offset by the marginal disutility of congestion. This is not the efficiency condition as given by (7.8); it in fact leads to a number of visits in excess of the optimum because the member disregards the congestion cost imposed on others. This condition demonstrates how the membership fee fails to place the correct incentives in place, so it can only be efficient if visits are directly regulated.

Assume that instead of a membership fee, the club charges a price per visit (or user fee). If the price is denoted \( p \) and the membership fee is set at \( F = 0 \), then the number of visits is chosen to solve

\[
\max_{x,v} U(x, G^*, v, [n^* - 1] v^* + v) \quad \text{subject to} \quad M = x + pv. \hfill (7.11)
\]

The necessary conditions for this optimization can be combined to give

\[
p = \frac{U_v}{U_x} + \frac{U_V}{U_x}. \hfill (7.12)
\]

Given the price, visits will be made up to the point at which the price is equal to the marginal benefit of another visit less the additional congestion cost it causes. Contrasting this to (7.8) shows that the efficient number of visits will be sustained if the price is set so that

\[
p = C_V - (n - 1) \frac{U_V}{U_x}. \hfill (7.13)
\]

However, it follows from efficient membership condition (7.7) that at this price the total revenue raised falls short of the cost of the club, since

\[
nvp = C + n v \frac{U_V}{U_x} < C \hfill (7.14)
\]

by the fact that \( U_V < 0 \). This inequality shows the important result that a membership fee alone cannot both generate the correct number of visits and allow the club to break even.
The charging scheme that is required to finance the club and control visits is a two-part tariff consisting of a membership fee and a user fee. Let the fixed part of the two-part tariff be given by $F$ and the user fee be $p$. With this tariff the club solves

$$\max_{\{x, v, G, n\}} U(x, G, v, nv) \quad \text{subject to} \quad M = x + F + pv,$$

and

$$nF + pnv = C(G, nv), \quad (7.15)$$

where both the individual budget constraint and the break-even constraint for the club have been imposed. The necessary conditions for this optimization readily yield the efficiency conditions (7.6) and (7.8), while the charging condition becomes

$$F + pv = vC_V - nv \frac{U_V}{U_x}, \quad (7.16)$$

which is the analogue of (7.7). Taken together, these observations show that the two-part tariff allows the club to break even and attain efficiency.

This section has addressed the issue of charging when the number of visits to the club cannot be controlled directly. It has shown that with variable utilization a two-part tariff is required. The cost per visit is used to control the number of visits while the fixed fee covers any residual payment needed for the club to break even.

### 7.4 Clubs and the Economy

The analysis of the decision process of an individual club demonstrated that the club will ensure efficiency of provision for its membership. It is tempting to conclude from this observation that the argument can be extended to the economy as a whole, with efficiency in public good provision attained by the population of consumers separating themselves into a series of efficient clubs. This was the conclusion reached by Buchanan (1965). We now argue that although this may sometimes be so, it is by no means guaranteed.

There are two settings in which the issue of economywide efficiency can be considered. The first setting, and the analytically simpler of the two, is to consider an economy where the efficient size of club is small relative to the total population. This situation applies when the club suffers from significant congestion, so its efficient size is relatively small, and population size is large. The second setting is when the efficient size of the club is large relative to the total population. This can arise either through limited congestion or through there being a small population. Either of these settings
can potentially occur, and they give very different perspectives on the efficiency of clubs at the level of the economy.

### 7.4.1 Small Clubs

Consider first an economy in which the size of the efficient membership of a club is small relative to the size of the total population. This allows some very clear conclusions to be obtained.

To understand the effect of this assumption, consider what happens as population size increases. Initially, with a small population, there will either be some of the population who are not in an efficiently sized club, or else every club will differ slightly in size from the efficient level. In the first case, as the size of the population increases, the number of those not in an efficiently sized club becomes trivial compared to the total population, so the deviation from efficiency tends to zero. In the second case, as the population increases, the deviation of each club from the optimum size becomes smaller and smaller, so again the inefficiency tends to zero. Therefore increases in population size eventually wipe out the deviation from efficiency.

The limiting interpretation of a large population is one which is infinite in number. Assuming an infinite population allows the standard “tricks” that can be played with infinity. In particular, if the population size is infinite, then it can be divided exactly into an infinite number of optimal efficiently sized clubs. The provision of public goods by clubs is then efficient for each club and for the economy as a whole.

The conclusion of this analysis is that if the efficient membership of each club is small relative to the total population, then the outcome for the economy will be that a very large number of clubs will form each with the correct number of members and each providing the efficient level of service. Hence efficiency will be attained for the economy as a whole. In this case the efficiency of each individual club is reflected at the aggregate level.

### 7.4.2 Large Clubs

The second case, which is more interesting from both a practical and an analytical perspective, arises when the efficient membership of each club is relatively large compared to the total population. In this case the population size can support only a limited number of efficiently sized clubs.

Two outcomes are then possible. It may be that the total population size is an integer multiple of the number of clubs. This allows the population to be divided neatly among
the clubs and efficiency is achieved. However, such a neat match between club size and population is very unlikely. The more likely outcome is that there will be some remainder when the total population is divided by optimum club size. The outcome in this situation requires some careful analysis.

To focus the argument, assume that the total population is more than the optimum size of a club but less than twice the optimum. With the total population denoted by \( N \) and the optimum club size by \( n^* \), utility as a function of the size of the club is graphed in figure 7.1. The assumptions imply that membership of a club of size \( \frac{N}{2} \) produces less utility than that of a club of size \( n^* \).

To determine the equilibrium, it is necessary to be clear about what is possible and what is not in terms of membership fees. For reasons that will become clear, a distinction must be made between cases where all members of a club must pay the same fee and cases where fees can be different among members. The latter case can be interpreted as all club members paying the same fee but making transfers, or “compensation” payments, among themselves. If this occurs, the fees net of transfers will differ.

**Equal Fees**

Let all members of each club pay the same fee. In this case it is easy to see that two clubs of size \( \frac{N}{2} \) cannot be an equilibrium. Start from such a position and consider the decision problem of an individual. Assume that all other club members remain in their
initial clubs, which is an application of Nash equilibrium. The individual can stay in the same club or move to the other club. If individuals change clubs, the club they move to increases in membership from $\frac{N}{2}$ to $\frac{N+1}{2}$. This larger club provides greater utility, so moving is the preferred choice. Existing members of the club also benefit from an increase in membership and will welcome a new member. Because there is an incentive to change clubs, the initial position could not have been an equilibrium. For there to be equilibrium, one club must be of efficient size $n^*$ and the other club of size $N - n^*$. In equilibrium the members of the larger club have a higher level of utility. Thus they have no motive to move to the smaller club, nor to accept members of the smaller club, because they will be made worse off by doing so.

The next question is whether this outcome is optimal from the viewpoint of society. The answer is dependent on the precise situation, but for the example in figure 7.1, the possibilities can be grouped into four categories. To see this, note that the social decision must be to choose whether to have one club or two clubs. When there is a single club, it may be beneficial to exclude some individuals from the club altogether. Alternatively, the single club may contain the entire population. If it is optimal to have two clubs, these may be equally sized or may be dissimilar. Summarizing this discussion gives the following breakdown of potential optimum configurations:

1. a single club, some of the population excluded;
2. a single club containing the entire population;
3. two equally sized clubs; or
4. two unequal clubs.

Outcome 1 will occur if it is too costly to form a new club for a small number of members and additional membership of the single club reduces the benefit of existing members significantly. The contrast with outcomes 2, 3, and 4 depends on the costs of congestion relative to the gains from being closer to the efficient level of membership. For instance, in outcome 3, with two equally sized clubs, both must have less than the efficient membership. The question then has to be asked whether it is better to take one closer to the efficient level (moving to outcome 4). Those in the larger club will gain while those left in the smaller club will lose. Contrasting outcome 4 to outcome 2, the question has to be asked whether the smaller club in outcome 4 should be closed completely and the population all placed in a single club. This will cause a congestion cost for those initially in the larger club but may benefit those who were in the smaller club, since the per capita cost will be lower and public good provision higher.
At this level of generality it is not possible to proceed to identify the nature of the optimal allocation without being completely specific about the relationships (the utility function and congestion function) that underlie the model. What can be concluded is that there is no necessity for the equilibrium position with two dissimilar sized clubs to be the optimal outcome. So, from the perspective of the entire economy, the actions of the clubs though individually efficient do not guarantee social optimality. The reason is that both clubs are competing to become larger, so when one club attracts new members in order to grow, it does not take into account the cost inflicted on the members of the other club that is becoming smaller.

To illustrate this point consider the following example: The total population, $N$, is normalized to have size one ($N = 1$), and this population has to be allocated between two clubs in proportions $n$ and $1 - n$ (with $0 \leq n \leq 1$). The utility of being in a club of size $n$ is given by

$$U(n) = n^3(1 - n)$$

(7.17)

so the utility-maximizing club size is three-quarters of the population (which is greater than half the population, giving the situation illustrated in figure 7.1). Clubs with either the entire population as members or with no members provide zero utility.

We graph the utility of each club member for each partition of the population between the two clubs in figure 7.2. The figure measures the membership $n$ in club $A$ from the left corner and the membership $1 - n$ of club $B$ from the right corner. The width of the

**Figure 7.2**
Optimum with unequally sized clubs
figure is the total population, which is normalized to one. Utility in club $A$ begins at 0 when $n = 0$ and rises to a maximum when $n = \frac{3}{4}$. Reading from the right corner, we see that utility in club $B$ begins at 0 when $n = 1$ and rises to a maximum at $n = \frac{1}{4}$. The equilibrium outcome occurs when one club has an efficient membership, with $\frac{3}{4}$ of the population, and the other is inefficient, with just $\frac{1}{4}$.

The key feature of this example is that population is too small to allow both clubs to reach their utility maximizing size of $\frac{3}{4}$. The optimal outcome is obtained by maximizing total welfare

$$ W(n) = nU(n) + (1-n)U(1-n). \quad (7.18) $$

The necessary condition for optimality is then

$$ U(n) + nU'(n) = U(1-n) + (1-n)U'(1-n), \quad (7.19) $$

which requires that the marginal gains of another member be the same for both clubs. The average level of welfare, $\frac{W(n)}{N}$, is depicted by twin-peaked curve in figure 7.2. It is then readily seen that optimality is achieved at one of the two peaks where there are two unequally sized clubs. Furthermore the membership allocation at either of these peaks has one club that exceeds the efficient membership and another club that falls below. The attainment of optimality requires that the size of the larger club be pushed beyond the size that maximizes the utility of each member. The reason for this is that welfare is concerned with the product of $n$ and $U(n)$, so there is always an incentive to raise the membership of the club generating the higher utility for its members.

Although this incentive always exists, it is never the dominant effect. Changing the utility function can affect the optimal outcome, but it will preserve the fundamental inefficiency of the equilibrium outcome. Figure 7.3 depicts the situation for the utility function

$$ U(n) = n^2(1-n). \quad (7.20) $$

The equilibrium involves two unequal clubs, one with $\frac{2}{3}$ of the population and the other with $\frac{1}{3}$. The resulting average welfare function is single-peaked, with its maximum occurring with two equally sized clubs of size $n = 1 - n = \frac{1}{2}$. Hence optimality is attained when both clubs are below the efficient membership.

### Unequal Fees

The case where equal fees were paid by all members of a club was complex in terms of possible outcomes, but that where fees can be unequal is much more so. To gain some
As a starting point, let the allocation of members be the equilibrium found for the no-transfer case where there is one club of size \( n^* \) and one of \( N - n^* \). It was previously argued that there was no incentive for those in the optimal club to move and no possibility of those in the smaller club being allowed to move. When unequal fees are allowable, neither of these claims need be true. Consider first a member of the smaller club. If they were to move to the larger club, they would obtain a utility gain of

\[ U(n^* + 1) - U(N - n^*) \]

Their presence makes the previously efficient club too large, so the welfare of its existing members will fall. However, the gain of the new member could be sufficiently great that it would more than compensate the existing members for their losses and yet still be better off. In other words, the new member pays a fee greater than that of the existing members and the fee of existing members is reduced more than sufficiently to compensate them for the additional crowding. If this compensation is possible, then the move between clubs will be allowed and the initial position cannot be an equilibrium.

Now consider reversing the argument and considering the incentive for a member of the efficient club to move to the smaller club. With equal fees this would never happen. Now let unequal fees be allowed. If the move did occur, the club member moving would lose utility of value \( U(n^*) - U(N - n^* + 1) \), but the existing members of the smaller club would each gain \( U(N - n^* + 1) - U(N - n^*) \). If they could collectively agree to pay compensation to the new member (meaning let them pay a lower membership
fee), then it is possible that the existing members could more than compensate the new member for the loss incurred in their move while still remaining better off themselves.

These arguments reveal that members of the efficient club may be enticed to the smaller club and that members of the smaller club may be able to “buy” themselves into the efficient club. Both of these mechanisms may even be functioning simultaneously. The outcome of this reasoning is that it may not be possible to find any equilibrium, and even if an equilibrium exists, it is not easy to characterize. Furthermore there is even less reason to expect any equilibrium that is achieved to be efficient. All of this occurs because the population cannot be allocated to a set of clubs each with efficient membership except in the unlikely case of population size being a integer multiple of the efficient membership level. This problem does not diminish even when population size increases.

There is one situation in which this argument does not apply. Consider again the graph of utility as a function of club size drawn in figure 7.1. The problems of dividing the population into efficient clubs resulted from the fact that there was a unique value for efficient club membership. If the graph were instead like figure 7.4, with a flat section at its peak, then there would be a range of efficient sizes. To see the effect of this, let the efficient club size range from 2 to 3. Then a population of 11 consumers could be divided into three clubs of size 3 and one of size 2 and the economy would achieve efficiency. Of course, with a population of size 11, this could not be done if efficient

![Figure 7.4](image)

Non-unique club size
club size was unique (unless it was 11). Furthermore any population size greater than 11 can be divided into efficiently sized clubs.

The general version of this argument is illustrated in figure 7.5. For a single club the range of efficient memberships is between \( n' \) and \( n'' \). When there are two clubs, efficiency can occur for the range \( 2n' \) to \( 2n'' \). This extension of the range continues as additional clubs are introduced. Eventually, if the total population is large enough, the ranges of values of total population for which efficiency cannot be achieved shrink to zero (alternatively, the ranges of optimal size overlap) and all consumers can be placed in efficient clubs.

7.4.3 Conclusion

The conclusion of this section has to be that the efficiency of the individual club does not always translate into efficiency for the economy. In a large population approximate efficiency will be achieved, and individual utility will be virtually equal to maximal attainable utility. However, when there is a small-number problem, efficiency will not be achieved by the equilibrium allocation of members between clubs. This should not be surprising since small numbers introduces problems akin to those found in oligopoly markets. What occurs is that small groups of consumers are able to affect their own utility levels by choosing to form efficiently sized clubs. Therefore they possess market power, and this is reflected in the inefficiency. These problems are eliminated if there are a range of efficient club sizes.
7.5 Local Public Goods

The concept of a local public good was introduced in section 7.2. A local public good has the feature that its benefits are restricted to a particular geographical area and it cannot be enjoyed outside of that area. In relating this idea to the analysis of club goods, one could think of local communities as clubs that are formed to provide local public goods. To become a member of a local community, a consumer must move into the area (i.e., join the club) and pay whatever local taxes are levied in that community (i.e., pay the membership fee). Once they have done this, they can then enjoy the local public goods that are provided.

An important feature of the club good was that exclusion was possible, and it is interesting to discuss whether this is the case with local public goods. There are two points at which exclusion may be possible. First, a consumer must become resident in an area in order to benefit from the local public good. Although few (if any) local authorities have the right to prevent the resale of houses or to forcibly evict existing occupants, they do have the power to prevent additional new building. Consequently, although reductions in population may be hard to achieve (unlike expulsions in an ordinary club), the exclusion of additional members is possible. Second, there is the payment of taxes. Any resident who refuses to pay local taxes can be either forced to pay or excluded from the club, since local authorities have legal authority to collect taxation. If we impose the possibility of exclusion, then the analysis of local public goods becomes exactly that of the clubs we have already considered. However, the analysis of nonexclusion is also of interest with local public goods, since this captures the idea of a freely operating market in which individuals have the freedom to select their preferred residential location. We now focus on nonexclusion.

The concept of local public goods can be applied to the provision of public services by local regions in order to understand the allocation of a population between different localities. Intuitively we can think of localities competing for population by setting the package of public good provision and taxation they offer. Members of the population look at what is offered in different localities and select the one that offers the highest utility level. This will cause population flows until no one can gain by moving locality. This is similar to the adjustment process for club goods except for the fact that there is free access (i.e., no possibility of barring access to new migrants even if the existing population would lose from the immigration).

In this framework it is natural to question whether an efficient equilibrium will be attained. The localities are competing for population and no restrictions are imposed
on the freedom of the population to move between regions. With clubs, efficiency was achieved at least within the clubs. To see whether the same is true for local public goods, it is necessary to construct a model of location choice.

Consider a total population of \( H \) consumers that is to be divided between two localities, with \( h \) being the population of a locality (we use different notation to avoid confusion with club membership). Each locality provides a local public good financed through a charge on the population. As the population increases, the unit cost of the public good per resident is reduced. This is the benefit from increased population. There is also a cost to increasing population. This can be motivated by assuming that there is a fixed resource in each region so that income per person falls as the population rises and this resource has to be shared among a greater number.

These assumptions imply that income can be written as a decreasing function, \( M(h) \), of the population of a locality. Think of wages or welfare benefits reducing with increased immigration. If the locality provides \( G \) units of the public good, the charge per resident is \( \frac{G}{h} \). Combining these obtains the income left to spend on private goods, \( M(h) - \frac{G}{h} \), and the resulting level of utility, \( U(M(h) - \frac{G}{h}, G) \).

It is assumed that localities choose the level of public good optimally given their population. This eliminates the possibility of inefficiency through a level of provision that does not satisfy the Samuelson rule. Given a population \( h \), the level of public good provision satisfies the Samuelson rule

\[
h \frac{U_G}{U_x} = 1. \tag{7.21}
\]

This condition can be solved to find the level of public good, \( G(h) \), which depends on the population of the locality. Substituting the level of the public good into the utility function determines the level of utility as a function of population. This relationship is written in brief as \( U(h) \). The implications of the model follow from the fact that an increase in \( h \) can increase or decrease utility. Differentiating \( U(h) \) with respect to \( h \) shows that

\[
U' = U_x M' + U_x \left( \frac{G}{h^2} \right) \tag{7.22}
\]

The first term on the right-hand side is negative, since an increase in population reduces income \( M \), while the second term is positive because the cost of the public good is reduced. It is therefore unclear what the net effect will be. To analyze the model further, assume that utility initially increases with the population until it reaches a maximum and then decreases. In addition let \( U(H) > U(0) \) so that having all the population in
a single locality leads to higher utility than having no population. This can be motivated by the fact that a small number of people find it very expensive to provide the public good, but the income is not reduced too far when the entire population is in one locality.

The dynamics of migration are that the population always flows from the locality with the lower utility to the locality with the higher utility. An equilibrium is reached when both localities offer the same utility level or else all the population is in one region. Consequently, if $U(H) \geq U(0)$, an equilibrium can have all the population locating in one region or have the population divided between the two localities with utilities equalized. In the latter equilibrium $U(h^1) = U(h^2)$. The outcomes that can arise in this model can be illustrated by graphing the utility against the population in the two regions.

A possible structure of the utility function is shown in figure 7.6. This figure measures the population in locality 1 from the left corner and the population in locality 2 from the right corner. The width of the figure is the total population. The essential feature of this figure is that the population level that maximizes utility is less than half the total population. There are five potential equilibria at a, b, c, d, and e. The equilibrium at c is symmetric with both regions having a population of $\frac{H}{2}$. This equilibrium is also stable and will arise from any starting point between b and d. The two asymmetric equilibria at b and d are unstable. For instance, starting just above b, the population

---

**Figure 7.6**

Stability of the symmetric equilibrium
will adjust to $c$. Starting just below $b$, the population will adjust to $a$. The two extremes points, $a$ and $e$, where all the population are located within one of the two localities are stable but inefficient.

An alternative structure of utility is shown in figure 7.7. The change made is that the utility-maximizing population of a locality is now greater than one-half of the total population. There is still a symmetric and efficient equilibrium at $b$. But this equilibrium is now unstable: starting with a population below $b$, the flow of population will lead to the extreme outcome at $a$, whereas starting above $b$ will lead to $c$. The two extreme equilibria are stable but inefficient. All consumers would prefer the symmetric equilibrium to either of the extreme equilibria.

What this simple model shows is that there is no reason why flows of population between localities will achieve efficiency. It is possible for the economy to get trapped in an inefficient equilibrium. In this case the market economy does not function efficiently. The reason for this is that the movement between localities of one consumer affects both the population left behind and the population the consumer joins. These nonmarket linkages lead to the inefficiency.

![Figure 7.7](image.png)

**Figure 7.7**
Inefficient stable equilibrium
7.6 The Tiebout Hypothesis

The previous section has shown that inefficiency can arise when the population divides between two regions on the basis of their provision of local public goods. From this result it would be natural to infer that inefficiency will always be an issue with local public goods. It is therefore surprising that the Tiebout hypothesis asserts instead that efficiency will always be obtained with local public goods.

Tiebout observed that pure public goods lead to market failure because of the difficulties connected with information transmission. Since the true valuation by a consumer of a public good cannot be observed and a pure public good is nonexcludable, free-riding occurs and private provision is inefficient. This point was explored in the previous chapter. Now assume that there are a number of alternative communities where a consumer can choose to live and that these differ in their provision of local public goods. In contrast to the pure public good case, a consumer’s choice of which location to live in provides a very clear signal of preferences. The chosen location is obviously the one offering the provision of local public goods closest to the consumer’s ideal. Hence, through community choice, preference revelation takes place. Misrepresenting preference cannot help a consumer here, since the choice of a nonoptimal location merely reduces the consumer’s welfare level. The only rational choice is to act honestly.

The final step in the argument can now be constructed. When preference revelation is taking place, it follows that if there are enough different types of community and enough consumers with each kind of preference, then all consumers will allocate themselves to a community that is optimal for them and each community will be optimally sized. Thus the market outcome will be fully efficient, and the inefficiencies discussed in connection with pure public goods will not arise. Phrased more prosaically, consumers reveal their preferences by voting with their feet, and this ensures the construction of optimal communities. This also shows why the analysis of the previous section failed to find efficiency. The existence of at most two localities violated the large-number assumption employed in this argument.

The significance of this efficiency result, which is commonly called the Tiebout hypothesis, has been much debated. Supporters view it as another demonstration of the power of the market in allocating resources. Critics denounce it as simply another empty demonstration of what is possible under unrealistic assumptions. Certainly the Tiebout hypothesis has much the same foundations as the Two Theorems of Welfare Economics, since both concern economies with no rigidities and large numbers of
participants. But there is one important difference between the two: formalizing the Tiebout hypothesis is a more difficult task.

To obtain an insight into this difficulty, some of the steps in the previous argument need to be retraced. It was assumed that consumers could move between communities or at least choose between them with no restrictions on their choice. If housing markets function efficiently, there should not be a problem in finding accommodation. Where problems do arise is in the link between income and location. An assumption that can justify the previous analysis is that consumers obtain all their income from “rents” such as from the ownership of land, property, or shares. In this case it does not matter where the consumers choose to reside, since the rents will accrue regardless of location. Once some income is earned from employment, the Tiebout hypothesis only holds if all employment opportunities are replicated in all communities. Otherwise, communities with better employment prospects will appear more attractive even if they offer a slightly less appealing set of local public goods. If the two issues become entangled in this way, then the Tiebout hypothesis will naturally fail.

Further difficulties with the hypothesis arise when the numbers of communities and individuals is considered. When these are both finite, the problems already discussed above with achieving efficiency through market behavior arise again. These are compounded when individuals of different types are needed to make communities work. For example, assume that community A needs 10 doctors and 20 teachers to provide the optimal combination of local public goods while community B requires 10 police officers and 20 teachers. If doctors, teachers, and police officers are not found in the proportions 1:4:1, then efficiency in allocation between the communities cannot be achieved. Furthermore, if all teachers have different tastes from doctors and from police officers, then neither community can supply the ideal local public good combination to meet all tastes.

The efficiency of the allocation can nevertheless be recovered in two steps. First, if we appeal again to the large population assumption, the issue of achieving the precise mix of different types is eliminated—there will always be enough people of each type to populate the localities in the correct proportions. Second, even if tastes are different, it is still possible to obtain agreement on the level of public good through the use of personalized prices. This issue has already been discussed for public goods in connection with the Lindahl equilibrium. The same idea can be applied to local public goods, in which case it would be the local taxes that are differentiated among residents to equalize the level of public good demand and to attain efficiency with a heterogeneous population.
The Tiebout hypothesis depends on the freedom of consumers to move to preferred locations. This is only possible if there are no transactions costs involved in changing location. In practice, such transactions costs arise in the commission that has to be paid to estate agents, in legal fees, and in the physical costs of shipping furniture and belongings. These can be significant and cause friction in the movement of consumers to the extent that suboptimal levels of provision will be tolerated to avoid paying these costs.

To sum up, the Tiebout hypothesis provides support for allowing the market, by which is meant the free movement of consumers, to determine the provision of local public goods. By choosing communities, consumers reveal their tastes. They also have to abide by local tax law, so free-riding is ruled out. Hence efficiency is achieved. Although apparently simple, there are a number of difficulties when the practical implementation of this hypothesis is considered. The population may not partition neatly into the communities envisaged, and employment ties may bind consumers to localities whose local public good supply is not to their liking. Transactions costs in housing markets may be significant, and therefore limit the freedom of movement that is key to the hypothesis. The hypothesis provides an interesting insight into the forces at work in the formation of communities, but it does not guarantee efficiency.

7.7 Empirical Tests

The Tiebout hypothesis provides the reassuring conclusion that efficiency is attained by local communities providing public goods efficiently. If correct, the forces of economics and local politics can be left to work unrestricted by government intervention. Given the strength of this conclusion, and some of the doubts cast on whether the Tiebout argument really works, it is natural to conduct empirical tests of the hypothesis.

In testing any hypothesis, it is first necessary to determine what the observational implications of the hypothesis will be. For Tiebout, this means isolating what may be different between an economy in which the Tiebout hypothesis applies and one in which it does not. Empirical testing has been handicapped by the difficulty of establishing quite what this difference is.

The earlier empirical studies focused on property taxes, public good provision, and house prices. The reason for this was made clear by Oates, who initiated this line of research in 1969: local governments fund their activities primarily through property taxes and the manner in which these taxes are reflected in house prices provides evidence on the Tiebout hypothesis. Assume that all local governments provide the same level of
Chapter 7: Club Goods and Local Public Goods

public goods. Then the jurisdictions with higher property tax rates will be less attractive and have lower house prices. Now let the provision of public goods vary. With tax rates held constant, house prices should be higher in areas with more public good provision. These effects offset each other, and if the public good effect is sufficiently strong, jurisdictions with higher tax rates will actually have higher property prices. Oates considered evidence on house prices, property tax rates, and educational provision for 53 primarily residential municipalities in New Jersey. These municipalities were chosen because the majority of residents commuted to work and hence were not tied by employment to a particular location. The analysis showed that house prices were reduced by high property taxes but increased by greater public good provision.

Whether these results were evidence in favor of the Tiebout hypothesis became the subject of a debate that focused on the implications of the theory. Whereas Oates took differences in property prices as an indication of the Tiebout hypothesis at work (on the ground that more attractive locations would witness increased competition for the housing stock), an alternative argument suggests that a given quality of house would have the same price in all jurisdictions if Tiebout applied. The argument for uniform prices is based on the view that property taxes are the price paid for the bundle of public goods provided by the local government. If this price reflects the benefit enjoyed from the public goods, as it should if the Tiebout hypothesis is functioning, then it should not affect property prices. Uniform property prices should therefore be expected if the Tiebout hypothesis applies—an observation that lead to a series of studies looking for uniform house prices across jurisdictions with different levels of public good provision. Unfortunately, as Epple, Zelenitz, and Visscher (1978) show, the same conclusion is true even when the Tiebout hypothesis does not hold so that net-of-tax property prices should be uniform in all jurisdictions in all circumstances. Instead, they argue that when the Tiebout hypothesis applies, housing demand is not affected by the property tax rate, but when Tiebout does not apply, it is affected. Looking at prices, which are equilibrium conditions, cannot then provide a test of Tiebout. Rather, a test has to be based on the structural equations of housing demand and location demand and their dependence, or otherwise, on tax rates. This conclusion undermines the earlier work on property values but does not provide an easily implementable test.

As a response to these difficulties, alternative tests of the hypothesis have been constructed. One approach to determining whether the Tiebout hypothesis applies is to consider the level of demand for public goods from the residents of each locality. If the Tiebout hypothesis applies, residents should have selected a residential location that provides a level of public goods in line with their preferences. Hence within each locality there should be a degree of homogeneity in the level of demand for public goods. Note
Part III: Departures from Efficiency

carefully that this does not assert that all residents have the same preferences but only that, given the taxes and other local charges they pay, their demands are equalized. The test of the hypothesis is then to consider the variance in demand within regions relative to the variance in demand across regions. Such a test was conducted by Gramlich and Rubinfeld who studied households in Michigan suburbs and provided compelling evidence that there was less variation within regions than across regions.

It is necessary to note that these results do not confirm that the Tiebout hypothesis is completely operating but only that some sorting of residents is occurring. It is supportive evidence for the hypothesis but not complete confirmation. This conclusion is only to be expected since, given the extent of frictions in the housing market, the freedom of movement necessary for the hypothesis to hold exactly is lacking.

Overall, the empirical work is suggestive that the right forces are at work to push the economy toward the efficient outcome of Tiebout but that there are residual frictions that prevent the complete sorting required for the efficiency. Having said this, the tests have been limited to data from suburban areas that have the highest chance of producing the right outcome. In other locations, where the separation between work and location is not so simple, the hypothesis would have less chance of applying.

7.8 Conclusions

The chapter has discussed the nature of club goods and local public goods, and drawn the distinction between these and pure public goods. For a club good, the essential feature is the possibility of exclusion, and it has been shown how exclusion allows an individual club to attain efficiency. Although it is tempting to extend this argument to the economy as a whole, a series of new issues arise when the allocation of a population between clubs is analyzed. Efficiency may be attained, but it is not guaranteed.

Many of the same issues arise with local public goods whose benefits are restricted to a given geographical area. We have treated local public goods as a model of provision by localities where each locality is described by the package of public good and taxation that it offers. When there is no exclusion from membership, there is no implication that efficiency will be attained when residential choice can be made from only a small number of localities.

In contrast to this, the Tiebout hypothesis evokes a large-number assumption to argue that the population will be able to sort itself into a set of localities, each of which is optimal for its residents. At the heart of this argument is that choice of locality reveals preferences for public goods, so efficiency becomes attainable. The Tiebout hypothesis
has been subjected to empirical testing, but the evidence is at best inconclusive. While it shows some degree of sorting and is certainly not a rejection of Tiebout, it does not go as far as confirming that the promised efficiency is delivered.

Further Reading

The potential for clubs to achieve efficiency in the provision of public goods was first identified in:

A more extensive discussion of many of these issues can be found in:


A study of public goods with exclusion and user fees is in:

The problems of attaining efficiency in a club economy are explored by:

The problem of attaining efficiency in a local public goods economy with mobility is in:


The influential Tiebout hypothesis was first stated in:

A strong critique of the hypothesis is in:

Tests of the Tiebout hypothesis can be found in:


Part III: Departures from Efficiency


Exercises

7.1 If a tennis club does not limit membership, what will be the consequence?

7.2 Is education a local public good?

7.3 Can club theory be applied to analyze immigration policy?

7.4 Two clubs are engaged in competition for members. They compete by choosing the level of provision. Consumers are free to move between clubs, so equilibrium membership levels must equalize the utility levels offered by the two clubs. Assume that there is a fixed population of size \(N\), that the public good has unit cost, and that clubs share the cost equally over members. Show that a club that maximizes the utility of a representative member will not be efficient.

7.5 Consider a population of consumers. When a consumer is a member of a club providing a level of provision \(G\) and having \(n\) members, they obtain utility

\[
U = M - \frac{G}{n} + \log(G) - \frac{n}{k},
\]

where \(k\) is a positive constant and \(\frac{G}{n}\) is the charge for club membership.

a. Derive the optimal membership for the club if it maximizes the utility of each member.

b. Assuming that the club chooses \(G\) optimally given its membership, calculate the loss due to membership of a club with suboptimal size.

c. Assume that the total population is of size \(m\), with \(k < m < 2k\). Show that there is a continuum of Pareto-efficient allocations of population to clubs.

d. What club size maximizes total utility produced by the club? Contrast to the answer for part a.

7.6 Will a club be efficient if it does not exercise exclusion?

7.7 What will be the efficient membership level of a club if there is no congestion? Is it still appropriate to call it a club good if there is no congestion?

7.8 Do all members of a club agree with the club’s choices? What about nonmembers?

7.9 Assume that a consumer receives a utility of \([a + bG - \beta n] + M - p\) when paying a price \(p\) to be in a club with \(n\) members and provision \(G\) of the public good and a utility of \(M\) if the consumer is not in a club. \(a\), \(b\), and \(\beta\) are positive constants.

a. Show that the willingness-to-pay of a consumer for club membership satisfies \(p \leq [a + bG - \beta n]\).

b. Assume that the club is provided by a monopolist who chooses membership and provision to maximize profit. If the cost of running the club is \(G + n\), what are the profit-maximizing choices \(G\) and \(n\)?
Chapter 7: Club Goods and Local Public Goods

c. What choice of \( G \) and \( n \) maximizes the welfare of a typical member if costs of the club are shared equally?
d. Compare the monopolistic and welfare-maximizing equilibrium values and discuss the contrasts.

7.10 Theme parks do not use two-part tariffs. What is the consequence? Why do they choose not to use two-part tariffs?

7.11 How can a monopolist employ a two-part tariff to extract consumer surplus? Is the outcome efficient?

7.12 How does the design of two-part tariffs have to be modified when consumers are heterogeneous?

7.13 Assume an economy with 100 identical consumers. Assume that if consumers belong to a club with \( n \) members and the cost of the club is shared equally, they obtain utility

\[
U = \begin{cases} 
  n & \text{for } n \leq 5, \\
  5 & \text{for } 5 < n < 6, \\
  (11 - n) & \text{for } n \geq 6.
\end{cases}
\]

a. Sketch this utility function and comment on the optimal club size.
b. Show that a population of size 14 cannot be allocated among optimal membership clubs. Beyond what population size is it possible to guarantee optimality?
c. How would your answers change if the utility function were instead

\[
U = \begin{cases} 
  n & \text{for } n \leq 5.5, \\
  (11 - n) & \text{for } n \geq 5.5.
\end{cases}
\]
d. Discuss which of the two specifications you find most compelling. Does this lead you to believe clubs will attain efficiency for the economy?

7.14 “A club will always seek to achieve the best outcome for its members. Therefore an economy with clubs achieves efficiency.” Explain and critically appraise this statement.

7.15 Consider a club where the utility function (incorporating the charge) of a member is \( U = a + bn - cn^2 \). Find the optimal membership of the club. What is the membership level that maximizes the total utility of the club? Contrast the two levels and explain the difference.

7.16 Explain why the economy will be closer to an efficient equilibrium when congestion occurs with a small membership level.

7.17 If the optimal club size is between 4 and 5, what is the smallest population beyond which efficiency is always achieved? What if the optimal size is between 3 and 4?

7.18 Let \( U = 40n - 2n^2 \). Find the optimal club membership \( n^* \). Graph the value of \( U \) against population size \( N \) when the population is divided among

a. 1 club,
b. 2 clubs,
c. 3 clubs.

7.19 What does the Tiebout hypothesis suggest for the organization of a city’s structure?
7.20 Should local communities be restricted in tax powers?

7.21 (Bewley 1981) Imagine a world with 2 consumers and 2 potential jurisdictions. Each of the consumers has one unit of labor to supply and preferences described by

\[ U = U(G^i) \]

where \( G^i \) is the quantity of public good provided in the jurisdiction \( i \) of residence. Denote labor supply in jurisdiction \( i \) by \( L_i \); the public good is produced from labor with production function

\[ G^i = L^i. \]

The regions both levy a tax on labor income to finance provision of their public good supply.

a. Assuming that consumers take taxes and public good provision as given when choosing their location, construct an inefficient equilibrium for this economy.

b. Discuss the inconsistency of consumer beliefs in this equilibrium.

c. How is the equilibrium modified if there is a continuum of consumers, each of whom is “small” relative to the economy?

7.22 (Scotchmer 1985) Suppose that consumers have income \( M \), preferences represented by

\[ U = x + 5 \log(G) - n \]

and the public good produced with cost function

\[ C(G) = G. \]

a. Show that the utility-maximizing membership is \( n = 5 \) with provision level \( G = 25 \).

b. Prove that if \( G \) is chosen optimally given \( n \), utility as a function of \( n \) is

\[ U = M + 10 \log(n) - 2n. \]

Hence, for a total population of 18, calculate the efficient (integer) number of clubs and their (possibly noninteger) membership level. What price for membership will give zero profit with these values of \( n \) and \( G \)?

c. Given the utility achieved at the solution to part b, show that the willingness to pay is given by

\[ p = 9.5 - 5 \log(22.5) + 5 \log(G) - n. \]

From this, find the profit maximizing choice of \( G \) and \( n \) and show that profit is positive. Comment on the possibility of an efficient, zero-profit equilibrium.

d. Discuss the integer issues in this analysis.

7.23 Explain the reasons why a consumer will relocate from one jurisdiction to another. What condition must be satisfied for several jurisdictions to have positive populations in equilibrium? Would such an equilibrium always be stable? If the entire population chooses to locate in a single jurisdiction must the equilibrium be efficient?

7.24 Consider two consumers with preferences

\[ U^h = 1 - T^i + \alpha^h \log(G^i), \quad h = 1, 2, \]
Chapter 7: Club Goods and Local Public Goods

where \( T_i \) is the tax levied in jurisdiction \( i \) and \( G_i \) is public good provision. Assume \( \alpha^2 > \alpha^1 \). The level of public good in each region is decided by majority voting of its residents. If there are two residents, assume that the supply is the average of the preferred quantities of the residents.

a. Show that the preferred quantity of public good for consumer \( h \) if they locate in jurisdiction \( i \) is given by
\[
G^h = n^i \alpha^h,
\]
where \( n^i \) is the jurisdiction population.

b. Assuming consumers correctly predict the consequences of location choice, show that there is no equilibrium if \( \alpha^2 > 1 + 2 \log \left( 1 + \alpha^2 \right) \).

c. Show that there is an equilibrium if consumers take provision levels as given.

7.25 (Lockwood 2002) There are three regions \( N = \{1, 2, 3\} \). In each region \( i \in N \) there is a local public good \( g_i \) to be produced or not (so \( g_i = \{0, 1\} \)). If \( g_i = 1 \), the cost is \( c_i \) units of private good, the benefit for resident in \( i \) is \( b_i \) and the external benefit of project \( i \) for resident in region \( j \neq i \) is \( e_{ij} \) (which can be either positive or negative) and \( e_{ii} = 0 \). Regions are ranked by increasing project cost \( c_1 < c_2 < c_3 \). For \( g_i = 0 \) there is no cost, benefit, or externality involved. Define the set of implemented projects \( G = \{i \in N | g_i = 1\} \). The preferences of residents in region \( i \) are given by
\[
u_i = \begin{cases} x_i + b_i + \sum_{j \neq i} e_{ij} g_j & \text{if } i \in G, \\ x_i + \sum_{j \neq i} e_{ij} g_j & \text{if } i \notin G, \end{cases}
\]
where \( x_i \) is private consumption. A resident in \( i \) is endowed with one unit of the private good and pays income tax \( t_i \) so private consumption is \( x_i = 1 - t_i \).

a. Define the efficient production of local public goods. What is the set of public goods that is produced?

b. With decentralization, the regional projects are funded locally and the regional budget constraint is \( t_i = c_i g_i \). What is the production outcome under decentralization?

c. What is the set, \( G^D \), of public goods that is produced under decentralization? \( (G^D = \{i \in N | g_i^D = 1\} \), where \( g_i^D \) is the (local) production decision) Compare with efficient outcome in part a.

c. Discuss why the decentralization outcome is unlikely to be efficient.

7.26 Consider the exercise above under centralization. With centralization, the regional projects are funded centrally with equal cost sharing among the three regions.

a. Derive the tax to be paid by any resident in each region if the set of funded projects is \( G^C = \{i \in N | g_i^C = 1\} \), where \( g_i^C \) is the funding decision in region \( i \) under centralization.

b. What is the payoff to any resident in region \( i \) from the set of funded projects \( G^C \)?

c. Suppose that project externalities are uniform \( e_{ij} = e > 0 \) for all \( i \neq j \). Suppose that funding decisions are made by majority voting so that for any region to get its project funded it must receive the support of at least one other region. What is the majority voting outcome?
d. Compare the majority voting outcome with the efficient outcome. Discuss the reasons why the centralized outcome will not be efficient in general.

7.27 Assume that there are three types of consumer with preferences \( U_1 = \alpha_1 \log(G) + x \), \( U_2 = \alpha_2 \log(G) + x \) and \( U_3 = \alpha_3 \log(G) + x \). There is an equal number of each type and all consumers have the same income level. If there are two jurisdictions that levy a tax and provide the public good, what is the equilibrium allocation? What is the efficient allocation?
8 Externalities

8.1 Introduction

An externality is a link among economic agents that lies outside the price system of the economy. Everyday examples include the pollution from a factory that harms a local fishery and the envy that is felt when a neighbor proudly displays a new car. Such externalties are not controlled directly by the choices of those affected—the fishery cannot choose to buy less pollution nor can you choose to buy your neighbor a worse car. This prevents the efficiency theorems described in chapter 2 from applying. Indeed the demonstration of market efficiency was based on the following two presumptions:

• The welfare of each consumer depended solely on her own consumption decision.
• The production of each firm depended only on its own input and output choices.

In reality these presumptions may not be met. A consumer or a firm may be directly affected by the actions of other agents in the economy; that is, there may be external effects from the actions of other consumers or firms. In the presence of such externalities the outcome of a competitive market is unlikely to be Pareto-efficient because agents will not take account of the external effects of their (consumption/production) decisions. Typically the economy will generate too great a quantity of “bad” externalities and too small a quantity of “good” externalities.

The control of externalities is an issue of increasing practical importance. Global warming and the destruction of the ozone layer are two of the most significant examples, but there are numerous others, from local to global environmental issues. Some of these externalities may not appear immediately to be economic problems, but economic analysis can expose why they occur and investigate the effectiveness of alternative policies. Economic analysis can generate surprising conclusions and challenge standard policy prescriptions. In particular, it shows how government intervention that induces agents to internalize the external effects of their decisions can achieve a Pareto improvement.

The starting point for the chapter is to provide a working definition of an externality. Using this, it is shown why market failure arises and the nature of the resulting inefficiency. The design of the optimal set of corrective, or Pigouvian, taxes is then addressed and related to missing markets for externalities. The use of taxes is contrasted with direct control through tradable licenses. Internalization as a solution to externalities is
considered. Finally these methods of solving the externality problem are set against the claim of the Coase theorem that efficiency will be attained by trade even when there are externalities.

8.2 Externalities Defined

An externality has already been described as an effect on one agent caused by another. This section provides a formal statement of this description, which is then used to classify the various forms of externalities. The way of representing these forms of externalities in economic models is introduced.

There have been several attempts at defining externalities and of providing classifications of various types of externalities. From among these the following definition is the most commonly adopted. Its advantages are that it places the emphasis on recognizing externalities through their effects and it leads to a natural system of classification.

**Definition 8.1** (Externality) An externality is present whenever some economic agent’s welfare (utility or profit) is “directly” affected by the action of another agent (consumer or producer) in the economy.

By “directly” we exclude any effects that are mediated by prices. That is, an externality is present if a fishery’s productivity is affected by the river pollution of an upstream oil refinery but not if the fishery’s profitability is affected by the price of oil (which may depend on the oil refinery’s output of oil). The latter type of effect (often called a *pecuniary externality*) is present in any competitive market but creates no inefficiency (since price mediation through competitive markets leads to a Pareto-efficient outcome). We will present later an illustration of a pecuniary externality.

This definition of an externality implicitly distinguishes between two broad categories. A *production externality* occurs when the effect of the externality is on a profit relationship and a *consumption externality* whenever a utility level is affected. Clearly, an externality can be simultaneously both a consumption and a production externality. For example, pollution from a factory may affect the profit of a commercial fishery and the utility of leisure anglers.

Using this definition of an externality, it is possible to move on to how they can be incorporated into the analysis of behavior. Denote, as in chapter 2, the consumption levels of the households by $x = \{x^1, \ldots, x^H\}$ and the production plans of the firms by $y = \{y^1, \ldots, y^m\}$. It is assumed that consumption externalities enter the utility
Chapter 8: Externalities

functions of the households and that production externalities enter the production sets of the firms. At the most general level, this assumption implies that the utility functions take the form

\[ U^h = U^h(x, y), \quad h = 1, \ldots, H, \]  

(8.1)

and the production sets are described by

\[ Y^j = Y^j(x, y), \quad j = 1, \ldots, m. \]  

(8.2)

In this formulation the utility functions and the production sets are potentially dependent on the entire arrays of consumption and production levels. The expressions in (8.1) and (8.2) represent the general form of the externality problem, and in some of the discussion below a number of further restrictions will be employed.

It is immediately apparent from (8.1) and (8.2) that the actions of the agents in the economy will no longer be independent or determined solely by prices. The linkages via the externality result in the optimal choice of each agent being dependent on the actions of others. Viewed in this light, it becomes apparent why competition will generally not achieve efficiency in an economy with externalities.

8.3 Market Inefficiency

It has been accepted throughout the discussion above that the presence of externalities will result in the competitive equilibrium failing to be Pareto-efficient. The immediate implication of this fact is that incorrect quantities of goods, and hence externalities, will be produced. It is also clear that a non–Pareto-efficient outcome will never maximize welfare. This provides scope for economic policy to improve the outcome. The purpose of this section is to demonstrate how inefficiency can arise in a competitive economy. The results are developed in the context of a simple two-consumer model, since this is sufficient for the purpose and also makes the relevant points as clear as possible.

Consider a two-consumer, two-good economy where the consumers have utility functions

\[ U^1 = x^1 + u_1(z^2) + v_1(z^1) \]  

(8.3)

and

\[ U^2 = x^2 + u_2(z^2) + v_2(z^1). \]  

(8.4)
The externality effect in (8.3) and (8.4) is generated by consumption of good $z$ by the consumers. The externality will be positive if $v_h(\cdot)$ is increasing in the consumption level of the other consumer and negative if it is decreasing.

To complete the description of the economy, it is assumed that the supply of good $x$ comes from an endowment $\omega_h$ to consumer $h$, whereas good $z$ is produced from good $x$ by a competitive industry that uses one unit of good $x$ to produce one unit of good $z$. Normalizing the price of good $x$ at 1, the structure of production ensures that the equilibrium price of good $z$ must also be 1. Given this, all that needs to be determined for this economy is the division of the initial endowment into quantities of the two goods.

Incorporating this assumption into the maximization decision of the consumers, the competitive equilibrium of the economy is described by the equations

$$u_h'(z^h) = 1, \quad h = 1, 2, \quad (8.5)$$

$$x^h + z^h = \omega^h, \quad h = 1, 2, \quad (8.6)$$

and

$$x^1 + z^1 + x^2 + z^2 = \omega^1 + \omega^2. \quad (8.7)$$

It is equation (8.5) that is of primary importance at this point. For consumer $h$ these state that the private marginal benefit from each good, determined by the marginal utility, is equated to the private marginal cost. The external effect does not appear directly in the determination of the equilibrium. The question we now address is whether this competitive market equilibrium is efficient.

The Pareto-efficient allocations are found by maximizing the total utility of consumers 1 and 2, subject to the production possibilities. The equations that result from this will then be contrasted to (8.5). In detail, a Pareto-efficient allocation solves

$$\max_{\{x^h, z^h\}} U^1 + U^2 = \left[ x^1 + u_1(z^1) + v_1(z^2) \right] + \left[ x^2 + u_2(z^2) + v_2(z^1) \right], \quad (8.8)$$

subject to

$$\omega^1 + \omega^2 - x^1 - z^1 - x^2 - z^2 \geq 0. \quad (8.9)$$

The solution is characterized by the conditions

$$u_1'(z^1) + v_2'(z^1) = 1 \quad (8.10)$$

and
Chapter 8: Externalities

\[ u'_2(z^2) + v'_1(z^2) = 1. \]  \hspace{1cm} (8.11)

In (8.10) and (8.11) the externality effect can be seen to affect the optimal allocation between the two goods via the derivatives of utility with respect to the externality. If the externality is positive, then \( v'_h > 0 \) and the externality effect will raise the value of the left-hand terms. It will decrease their value if there is a negative externality, so \( v'_h < 0 \). It can then be concluded that at the optimum with a positive externality the marginal utilities of both consumers are below their value in the market outcome. The converse is true with a negative externality. The externality leads to a divergence between the private valuations of consumption given by (8.5) and the corresponding social valuations in (8.10) and (8.11). This observation has the implication that the market outcome is not Pareto-efficient.

In general, it can also be concluded that if the externality is positive then more of good \( z \) will be consumed at the optimum than under the market outcome. The converse holds for a negative externality. This situation is illustrated in figure 8.1.

The market outcome is represented by equality between the private marginal benefit of the good \( (PMB) \) and its marginal cost \( (MC) \). The social marginal benefit \( (SMB) \) of the good is the sum of the private marginal benefit, \( u'_h(z^h) \), and the marginal external effect, \( v'_h(z^h) \). When \( v'_h(z^h) \) is positive, \( SMB \) is above \( PMB \). The converse holds when \( v'_h(z^h) \) is negative. The Pareto-efficient outcome equates the social marginal benefit to marginal cost. The market failure is characterized by too much consumption of a good

\[ \text{Marginal benefit and cost} \]

\[ \text{SMB}(v'_h > 0) \]

\[ \text{PMB} \]

\[ \text{MC} \]

\[ \text{SMB}(v'_h < 0) \]

\[ z^h \]

\[ \text{Figure 8.1} \]

Deviation of private from social benefits
causing a negative externality and too little consumption of a good generating a positive externality.

8.4 Externality Examples

The previous section has discussed externalities at a somewhat abstract level. We now consider more concrete examples of externalities. Some of the examples are very simple because of the binary nature of the choice and the assumption of identical individuals. This modeling choice was widely used by Schelling to achieve an extremely simple exposition that brings out the line of the argument very clearly. In addition it will illustrate the range of situations that fall under the general heading of externalities.

8.4.1 River Pollution

This example, from Louis Gevers, is one of the simplest examples that can be described using only two agents. Assume that two firms are located along the same river. The upstream firm \( u \) pollutes the river, which reduces the production (e.g., the output of fish) of the downstream firm \( d \). Both firms produce the same output, which they sell at a constant unit price of 1 so that total revenue coincides with production.

Labor and water are used as inputs. Water is free, but the equilibrium wage \( w \) on the competitive labor market is paid for each unit of labor. The production technologies of the firms are given by \( F^u(L^u) \) and \( F^d(L^d, L^u) \), with \( \frac{\partial F^d}{\partial L^u} < 0 \) to reflect that the pollution reduces downstream output. Decreasing returns to scale are assumed with respect to own labor input. Each firm acts independently and seeks to maximize its own profit \( \pi^i = F^i(\cdot) - wL^i \), taking prices as given.

The equilibrium is illustrated in figure 8.2. The total stock of labor is allocated between the two firms. The labor input of the upstream firm is measured from the left, that of the downstream from the right. Each point on the horizontal axis represents a different allocation between the firms. The upstream firm’s profit maximization process is represented in the upper part of the diagram and the downstream firm’s in the lower part. As the input of the upstream firm increases, the production function of the downstream firm moves progressively toward the horizontal axis. Given the profit-maximizing input level of the upstream firm, denoted \( L^u^* \), the downstream firm can do no better than choose \( L^d^* \). At these choices the firms earn profits \( \pi^u \) and \( \pi^d \) respectively. This is the competitive equilibrium. We now show that this is inefficient and that reallocating labor between the firms can increase total profit and reduce pollution.
Consider starting at the competitive equilibrium and make a small reduction in the labor input to the upstream firm. Since the choice was optimal for the upstream firm, the change has no effect on profit for the upstream firm (recall that $\frac{\partial \pi^u}{\partial L^u} = 0$). However, it leads to an outward shift of the downstream firm’s production function. This raises its profits. Hence the change raises aggregate profit. This demonstrates that the competitive equilibrium is not efficient and that the externality results in the upstream firm using too much labor and the downstream too little. Shifting labor to the downstream firm raises total production and reduces pollution.

### 8.4.2 Traffic Jams

The next example considers the externalities imposed by drivers on each other. Let there be $N$ commuters who have the choice of commuting by train or by car. Commuting by train always takes 40 minutes regardless of the number of travelers. The commuting
time by car increases as the number of car users increases. This congestion effect, which raises the commuting time, is the externality for travelers. Individuals must each make decisions to minimize their own transportation time.

The equilibrium in the choice of commuting mode is depicted in figure 8.3. The number of car users will adjust until the travel time by car is exactly equal to the travel time by train. For the travel time depicted in the figure, the equilibrium occurs when 40 percent of commuters travel by car. The optimum occurs when the aggregate time saving is maximized. This occurs when only 20 percent of commuters use a car.

The externality in this situation is that the car drivers take into account only their own travel time but not the fact that they will increase the travel time for all other drivers. As a consequence too many commuters choose to drive.

### 8.4.3 Pecuniary Externality

Consider a set of students each of whom must decide whether to be an economist or a lawyer. Being an economist is great when there are few economists, and not so great when the labor market becomes crowded with economists (due to price competition). If the number of economists grows high enough, they will eventually earn less than their lawyer counterparts. Suppose that each person chooses the profession with the best earnings prospects. The externality (a pecuniary one!) comes from the fact that when one more person decides to become an economist, he lowers all other economists’
incomes (through competition), imposing a cost on the existing economists. When making his decision, he ignores this external effect imposed on others. The question is whether the invisible hand will lead to the correct allocation of students across different jobs.

The equilibrium depicted in figure 8.4 determines the allocation of students between jobs. The number of economists will adjust until the earnings of an economist are exactly equal to the earnings of a lawyer. The equilibrium is given by the percentage of economists at point $E$. To the right of point $E$, lawyers would earn more and the number of economists would decrease. Alternatively, to the left of point $E$ economists are relatively few in number and will earn more than lawyers, attracting more economists into the profession.

The laissez-faire equilibrium is efficient because the external effect is a change in price. The cost to an economists of a lower income is a benefit to employers. Since employers’ benefits equals employees’ costs, there is zero net effect. The policy implication is that there is no need for government intervention to regulate the access to professions. It follows that any public policy that aims to limit the access to some profession, like the *numerus clausus*, is not justified. Market forces will correctly allocate the right number of people to each of the different professions.

![Figure 8.4](image-url)

*Figure 8.4*

*Job choice*
8.4.4 The Rat Race Problem

The rat race problem is a contest for relative position as pointed out by George Akerlof. It can help explain why students work too hard when final marking takes the form of a ranking. It can also explain the intense competition for a promotion in the workplace when candidates compete with each other and only the best is promoted. We take the classroom example here. Assume that performance is judged not in absolute terms but in relative terms so that what matters is not how much is known but how much is known compared to what other students know.

In this situation an advantage over other students can only be gained by working harder than they do. Since this applies to all students, all must work harder. But since performance is judged in relative terms, all the extra effort cancels out. The result of this is an inefficient rat race in which each student works too hard to no ultimate advantage. If all could agree to work less hard, the same grades would be obtained with less work. Such an agreement to work less hard cannot be self-supporting, since each student would then have an incentive to cheat on the agreement and work harder.

A simple variant of the rat race with two possible effort levels is shown in figure 8.5. In this figure, \( c, 0 < c < \frac{1}{2} \), denotes the cost of effort. For both students high effort is a dominant strategy. In contrast, the Pareto-efficient outcome is low effort. This game is an example of the prisoners’ dilemma whereby a Pareto improvement could be made if the players could make a commitment to the low-effort strategy.

Another example of rat race is the use of performance-enhancing drugs by athletes. In the absence of effective drug regulations, many athletes will feel compelled to enhance their performance by using anabolic steroids, and the failure to use steroids might seriously reduce their success in competition. Since the rewards in athletics are

\[
\begin{array}{c|cc}
\text{Player 1} & \text{Low} & \text{High} \\
\hline
\text{Low} & 1/2 & 1 - c \\
\text{High} & 0 & 1/2 - c \\
\end{array}
\]

Figure 8.5
Rat race
Chapter 8: Externalities

determined by performance relative to others, anyone that uses such drugs to increase their chance of winning must necessarily reduce the chances of others (an externality effect). The result is that when the stakes are high in the competition, unregulated contests almost always lead to a race for using more and more performance-enhancing drugs. However, when everyone does so, the use of such drugs yields no real benefits for the contestants as a whole: the performance-enhancing actions cancel each other. At the same time the race imposes substantial risks. Anabolic steroids have been shown to cause cancer of the liver and other serious health problems. Given what is at stake, voluntary restraint is unlikely to be an effective solution, and public intervention now requires strict drug testing of all competing athletes.

The rat race problem is present in almost every contest where something important is at stake and rewards are determined by relative position. In an electoral competition race, contestants spend millions on advertising, and governing bodies have now put strict limits on the amount of campaign advertising. Similarly a ban on cigarette advertising has been introduced in many countries. Surprisingly enough, this ban turned out to be beneficial to cigarette companies. The reason is that the ban helped them out of the costly rat race in defensive advertising where a company had to advertise because the others did.

8.4.5 The Tragedy of the Commons

The tragedy of the commons arises from the common right of access to a resource. The inefficiency to which it leads results again from the divergence between the individual and social incentives that characterizes all externality problems.

Consider a lake that can be used by fishermen from a village located on its banks. The fishermen do not own boats but instead can rent them for daily use at a cost $c$. If $B$ boats are hired on a particular day, the number of fish caught by each boat will be $F(B)$, which is decreasing in $B$. A fisherman will hire a boat to fish if they can make a positive profit. Let $w$ be the wage if they choose to undertake paid employment rather than fish, and let $p = 1$ be the price of fish so that total revenue coincide with fish catch $F(B)$. Then the number of boats that fish will be such as to ensure that profit from fishing activity is equal to the opportunity cost of fishing, which is the forgone wage $w$ from the alternative job (if profit were greater, more boats would be hired and the converse if it were smaller). The equilibrium number of boats, $B^*$, then satisfies

$$\pi = F(B^*) - c = w.$$  
(8.12)
The optimal number of boats for the community, $B^*$, must be that which maximizes the total profit for the village, net of the opportunity cost from fishing. Hence $B^*$ satisfies

$$\max_{B} \{ F(B) - c - w \}. \quad (8.13)$$

This gives the necessary condition

$$F\left( B^* \right) - c - w + BF'\left( B^* \right) = 0. \quad (8.14)$$

Since an increase in the number of boats reduces the quantity of fish caught by each, $F'(B^*) < 0$. Therefore contrasting (8.12) and (8.14) shows that $B^* < B^*$, so the equilibrium number of boats is higher than the optimal number. This situation is illustrated in figure 8.6.

The externality at work in this example is that each fisherman is concerned only with their own profit. When deciding whether to hire a boat, they do not take account of the fact that they will reduce the quantity of fish caught by every other fisherman. This negative externality ensures that in equilibrium too many boats are operating on the lake. Public intervention can take two forms. There is the price-based solution consisting of a tax per boat so as to internalize the external effect of sending a boat on the lake. As indicated in the figure, a correctly chosen tax will reduce the number of boats so as to restore the optimal outcome. Alternatively, the quantity-based solution consists of setting a quota of fishing equal to the optimal outcome.
8.4.6 Bandwagon Effect

The bandwagon effect studies the question of how standards are adopted and, in particular, how it is possible for the wrong standard to be adopted. The standard application of this is the choice of arrangement for the keys on a keyboard.

The current standard, Qwerty, was designed in 1873 by Christopher Scholes in order to deliberately slow down the typist by maximizing the distance between the most used letters. The motivation for this was the reduction of key-jamming problems (remember this would be for mechanical typewriters in which metal keys would have to strike the ink ribbon). By 1904 the Qwerty keyboard was mass produced and became the accepted standard. The key-jamming problem is now irrelevant, and a simplified alternative keyboard (Dvorak’s keyboard) has been devised that reduces typing time by 5 to 10 percent.

Why has this alternative keyboard not been adopted? The answer is that there is a switching cost. All users are reluctant to switch and bear the cost of retraining, and manufacturers see no advantage in introducing the alternative. It has therefore proved impossible to switch to the better technology.

This problem is called a bandwagon effect and is due to a network externality. The decision of a typist to use the Qwerty keyboard makes it more attractive for manufacturers to produce Qwerty keyboards, and hence for others to learn Qwerty. No individual

![Diagram](image_url)

**Figure 8.7**
Equilibrium keyboard choice
has any incentive to switch to Dvorak. The nature of the equilibrium is displayed in figure 8.7. This shows the intertemporal link between the percentage using Qwerty at time $t$ and the percentage at time $t + 1$. The natural advantage of Dvorak is captured in the diagram by the fact that the number of Qwerty users will decline over time starting from a position where 50 percent use Qwerty at time $t$. There are three equilibria. Either all will use Qwerty or Dvorak or else a proportion $p^*$, $p^* > 50$ percent, will use Qwerty and $1 - p^*$ Dvorak. However, this equilibrium is unstable, and any deviation from it will lead to one of the corner equilibria. The inefficient technology, Qwerty, can dominate in equilibrium if the initial starting point is to the right of $p^*$.

### 8.5 Pigouvian Taxation

The description of market inefficiency has shown that its basic source is the divergence between social and private benefits (or between social and private costs). This fact has been reinforced by the examples. A natural means of eliminating such divergence is to employ appropriate taxes or subsidies. By modifying the decision problems of the firms and consumers these can move the economy closer to an efficient position.

To see how a tax can enhance efficiency, consider the case of a negative consumption externality. With a negative externality the private marginal benefit of consumption is always in excess of the social marginal benefit. These benefits are depicted by the $PMB$ and $SMB$ curves respectively in figure 8.8. In the absence of intervention, the equilibrium occurs where the $PMB$ intersects the private marginal cost ($PMC$). This gives a level of consumption $x^m$. The efficient consumption level equates the $PMC$ with the $SMB$; this is at point $x^o$. As already noted, with a negative externality the market outcome involves more consumption of the good than is efficient. The market outcome can be improved by placing a tax on consumption. What it is necessary to do is to raise the $PMC$ so that it intersects the $SMB$ vertically above $x^o$. This is what happens for the curve $PMC'$, which has been raised above $PMC$ by a tax of value $t$. This process, often termed Pigouvian taxation, allows the market to attain efficiency for the situation shown in figure 8.8.

Based on arguments like that exhibited above, Pigouvian taxation has been proposed as a simple solution to the externality problem. The logic is that the consumer or firm causing the externality should pay a tax equal to the marginal damage the externality causes (or a subsidy if there is a marginal benefit). Doing so makes them take account of the damage (or benefit) when deciding how much to produce or consume. In many ways this is a compellingly simple conclusion.
The previous discussion is informative but leaves a number of issues to be resolved. Foremost among these is the fact that the figure implicitly assumes there is a single agent generating the externality whose marginal benefit and marginal cost are exhibited and that there is a single externality. The single tax works in this case, but will it still do so with additional externalities and agents? This is an important question to be answered if Pigouvian taxation is to be proposed as a serious practical policy.

To address these issues, we use our example from the market failure section again. This example involved two consumers and two goods with the consumption of one of the goods, \( z \), causing an externality. The optimal structure of Pigouvian taxes is determined by characterizing the social optimum and inferring from that what the taxes must be. Recall from (8.10) and (8.11) that the social optimum is characterized by the conditions

\[
\frac{d}{dz}u_1(z_1) + \frac{d}{dz}v_2(z_1) = 1 \quad (8.15)
\]

and

\[
\frac{d}{dz}u_2(z_2) + \frac{d}{dz}v_1(z_2) = 1. \quad (8.16)
\]

It is from contrasting these conditions to those for individual choice that the optimal taxes can be derived.
Utility maximization by consumer 1 will equate their private marginal benefit, $u'_1(z^1)$, to the consumer price $q_1$. Given that the producer price is equal to 1 in this example, (8.15) shows that efficiency will be achieved if the price, $q_1$, facing consumer 1 satisfies

$$q_1 = 1 - v'_2(z^1).$$  \hspace{1cm} (8.17)

Similarly from (8.16) efficiency will be achieved if the price facing consumer 2 satisfies

$$q_2 = 1 - v'_1(z^2).$$  \hspace{1cm} (8.18)

These identities reveal that the taxes that ensure the correct difference between consumer and producer prices are given by

$$t_1 = -v'_2(z^1)$$  \hspace{1cm} (8.19)

and

$$t_2 = -v'_1(z^2).$$  \hspace{1cm} (8.20)

Therefore the tax on consumer 1 is the negative of the externality effect their consumption of good $z$ inflicts on consumer 2. Hence, if the good causes a negative externality ($v'_2(z^1) < 0$), the tax is positive. The converse holds if it causes a positive externality. The same construction and reasoning can be applied to the tax facing consumer 2, $t_2$, to show that this is the negative of the externality effect caused by the consumption of good $z$ by consumer 2. The argument is now completed by noting that these externality effects will generally be different, and so the two taxes will generally not be equal. Another way of saying this is that efficiency can only be achieved if the consumers face personalized prices that fully capture the externalities that they generate.

So what does this say for Pigouvian taxation? Put simply, the earlier conclusion that a single tax rate could achieve efficiency was misleading. In fact the general outcome is that there must be a different tax rate for each externality-generating good for each consumer. Achieving efficiency needs taxes to be differentiated across consumers and goods. Naturally this finding immediately shows the practical difficulties involved in implementing Pigouvian taxation. The same arguments concerning information that were placed against the Lindahl equilibrium for public good provision with personalized pricing are all relevant again here. In conclusion, Pigouvian taxation can achieve efficiency but needs an unachievable degree of differentiation.

If the required degree of differentiation is not available, for instance, information limitations require that all consumers must pay the same tax rate, then efficiency will
not be achieved. In such cases the chosen taxes will have to achieve a compromise. They cannot entirely correct for the externality but can go some way toward doing so. Since the taxes do not completely offset the externality, there is also a role for intervening in the market for goods related to that causing the externality. For instance, pollution from car use may be lessened by subsidizing alternative mode of transports. These observations are meant to indicate that once the move is made from full efficiency, many new factors become relevant, and there is no clean and general answer as to how taxes should be set.

A final comment is that the effect of the tax or subsidy is to put a price (respectively positive or negative) on the externality. This leads to the conclusion, which will be discussed in detail below, that if there are competitive markets for the externalities, efficiency will be achieved. In other words, efficiency does not require intervention but only the creation of the necessary markets.

### 8.6 Licenses

The reason why Pigouvian taxation can raise welfare is that the unregulated market will produce incorrect quantities of externalities. The taxes alter the cost of generating an externality and, if correctly set, will ensure that the optimal quantity of externality is produced. An apparently simpler alternative is to control externalities directly by the use of licenses. This can be done by legislating that externalities can only be generated up to the quantity permitted by licenses held. The optimal quantity of externality can then be calculated and licenses totaling this quantity distributed. Permitting these licenses to be traded will ensure that they are eventually used by those who obtain the greatest benefit.

Administratively, the use of licenses has much to recommend it. As was argued in the previous section, the calculation of optimal Pigouvian taxes requires considerable information. The tax rates will also need to be continually changed as the economic environment evolves. The use of licenses only requires information on the aggregate quantity of externality that is optimal. Licenses to this value are released and trade is permitted. Despite these apparently compelling arguments in favor of licenses, when the properties of licenses and taxes are considered in detail, the advantage of the former is not quite so clear.

The fundamental issue involved in choosing between taxes and licenses revolves around information. There are two sides to this. The first is what must be known to calculate the taxes or determine the number of licenses. The second is what is known
Part III: Departures from Efficiency

when decisions have to be taken. For example, does the government know costs and benefits for sure when it sets taxes or issues licenses?

Taking the first of these, although licenses may appear to have an informational advantage this is not really the case. Consider what must be known to calculate the Pigouvian taxes. The construction of section 8.5 showed that taxation required the knowledge of the preferences of consumers and, if the model had included production, the production technologies of firms. Such extensive information is necessary to achieve the personalization of the taxes. But what of licenses? The essential feature of licenses is that they must total to the optimal level of externality. To determine the optimal level requires precisely, the same information as is necessary for the tax rates. Consequently taxes and licenses are equivalent in their informational demands.

Now consider the issue of the information that is known when decisions must be made. When all costs and benefits are known with certainty by both the government and individual agents, licenses and taxation are equivalent in their effects. This result is easily seen by reconsidering figure 8.8. The optimal level of externality is $x^o$, which was shown to be achievable with tax $t$. The same outcome can also be achieved by issuing $x^o$ licenses. This simple and direct argument shows there is equivalence with certainty.

In practice, it is more likely that the government must take decisions before the actual costs and benefits of an externality are known for sure. Such uncertainty brings with it the question of timing: Who chooses what and when? The natural sequence of events is the following. The government must make its policy decision (the quantity of licenses or the tax rate) before costs and benefits are known. In contrast, the economic agents can act after the costs and benefits are known. For example, in the case of pollution by a firm, the government may not know the cost of reducing pollution for sure when it sets the tax rate but the firm makes its abatement decision with full knowledge of the cost.

The effect of this difference in timing is to break the equivalence between the two policies. This can be seen by considering figure 8.9, which illustrates the pollution abatement problem for an uncertain level of cost. In this case the level of private marginal cost takes one of two values, $PMC_L$ and $PMC_H$, with equal probability. Benefits are known for sure. When the government chooses its policy, it is not known whether private marginal cost is high or low, so it must act on the expected value, $PMC_E$. This leads to pollution abatement $z^*$ being required (which can be supported by licenses equal in quantity to present pollution less $z^*$) or a tax rate $t^*$.

Under the license scheme, the level of pollution abatement will be $z^*$ for sure—there is no uncertainty about the outcome. With the tax, the level of abatement will
depend on the realized level of cost since the firm chooses abatement after this is known. Therefore, if the cost turns out to be $PMC_L$, so that the cost of abatement is low, the firm will be willing to undertake abatement up to level $z_L$. If the realized cost is $PMC_H$, so abatement cost is high, the firm will choose to undertake the reduced level of abatement $z_H$. This is shown in figure 8.9. Two observations emerge from this. First, the claim that licenses and taxation will not be equivalent when there is uncertainty is confirmed. Second, when cost is realized to be low, taxation leads to abatement in excess of $z^*$. The converse holds when cost is high.

The analysis of figure 8.9 may be taken as suggesting that licenses are better, since they do not lead to the variation in abatement that is inherent in taxation. However, it should also be realized that the choices made by the firm in the tax case are responding to the actual cost of abatement, so there is some justification for what the firm is doing. In general, there is no simple answer to the question of which of the two policies is better.

### 8.7 Internalization

Consider the example of a beekeeper located next door to an orchard. The bees pollinate the trees and the trees provide food for the bees, so a positive production externality runs in both directions between the two producers. According to the theory developed above, the producers acting independently will not take account of this externality. This leads to too few bees being kept and too few trees being planted.
The externality problem could be resolved by using taxation or insisting that both producers raise their quantities. Although both these would work, there is another simpler solution. Imagine the two producers merging and forming a single firm. If they were to do so, profit maximization for the combined enterprise would naturally take into account the externality. By so doing, the inefficiency is eliminated. The method of controlling externalities by forming single units out of the parties affected is called internalization, and it ensures that private and social costs become the same. It works for both production and consumption externalities whether they are positive or negative.

Internalization seems a simple solution, but it is not without its difficulties. To highlight the first of these, consider an industry in which the productive activity of each firm causes an externality for the other firms in the industry. In this situation the internalization argument would suggest that the firms become a single monopolist. If this were to occur, welfare loss would then arise due to the ability of the single firm to exploit its monopoly position, and this may actually be greater than the initial loss due to the externality. Although this is obviously an extreme example, the internalization argument always implies the construction of larger economic units and a consequent increase in market power. The welfare loss due to market power then has to be offset against the gain from eliminating the effect of the externality.

The second difficulty is that the economic agents involved may simply not wish to be amalgamated into a single unit. This objection is particularly true when applied to consumption externalities. That is, if a household generates an externality for their neighbor, it is not clear that they would wish to form a single household unit, particularly if the externality is a negative one.

In summary, internalization will eliminate the consequences of an externality in a very direct manner by ensuring that private and social costs are equated. However, it is unlikely to be a practical solution when many distinct economic agents contribute separately to the total externality, and it has the disadvantage of leading to increased market power.

### 8.8 The Coase Theorem

After identifying externalities as a source of market failure, this chapter has taken the standard approach of discussing policy remedies. In contrast to this, there has developed a line of reasoning that questions whether such intervention is necessary. The focal point for this is the Coase theorem, which suggests that economic agents may resolve externality problems themselves without the need for government intervention. This
conclusion runs against the standard assessment of the consequences of externalities and explains why the Coase theorem has been of considerable interest.

The Coase theorem asserts that if the market is allowed to function freely, then it will achieve an efficient allocation of resources. This claim can be stated formally as follows.

**Theorem 8.2** (Coase theorem) In a competitive economy with complete information and zero transaction costs, the allocation of resources will be efficient and invariant with respect to legal rules of entitlement.

The legal rules of entitlement, or property rights, are of central importance to the Coase theorem. Property rights are the rules that determine ownership within the economy. For example, property rights may state that all agents are entitled to unpolluted air or the right to enjoy silence (they may also state the opposite). Property rights also determine the direction in which compensation payments will be made if a property right is violated.

The implication of the Coase theorem is that there is no need for policy intervention with regard to externalities except to ensure that property rights are clearly defined. When they are, the theorem presumes that to eliminate any market failure, those affected by an externality will find it in their interest to reach private agreements with those causing it. These agreements will involve the payment of compensation to the agent whose property right is being violated. The level of compensation will ensure that the right price emerges for the externality and a Pareto-efficient outcome will be achieved. These compensation payments can be interpreted in the same way as the personalized prices discussed in section 8.5.

As well as claiming that the outcome will be efficient, the Coase theorem asserts that the equilibrium will be invariant to the how property rights are assigned. This is surprising since a natural expectation is, in the example, for the level of pollution under a polluter-pays system (i.e., giving property rights to pollutees) to be less than that under a pollutee-pays (i.e., giving property rights to the polluter). To show how the invariance argument works, consider the example of a factory that is polluting the atmosphere of a neighboring house. When the firm has the right to pollute, the householder can only reduce the pollution by paying the firm a sufficient amount of compensation to make it worthwhile to stop production or to find an alternative means of production. Let the amount of compensation the firm requires be $C$. Then the cost to the householder of the pollution, $G$, will either be greater than $C$, in which case the householder will be willing to compensate the firm and the externality will cease, or it will be less than $C$ and the
externality will be left to continue. Now consider the outcome with the polluter-pays principle. The cost to the firm for stopping the externality now becomes $C$ and the compensation required by the household is $G$. If $C$ is greater than $G$, the firm will be willing to compensate the household and continue producing the externality; if it is less than $G$, it stops the externality. Considering the two cases, it can be seen the outcome is determined only by the value of $G$ relative to $C$ and not by the assignment of property rights, which is essentially the content of the Coase theorem.

There is a further issue before invariance can be confirmed. The change in property rights between the two cases will cause differences in the final distribution of income due to the direction of compensation payments. Invariance can only hold if this redistribution of income does not cause a change in the level of demand. This requires there to be no income effects, or to put it another way, the marginal unit of income must be spent in the same way by both parties.

When the practical relevance of the Coase theorem is considered, a number of issues arise. The first lies with the assignment of property rights in the market. With commodities defined in the usual sense, it is clear who is the purchaser and who is the supplier, and therefore the direction in which payment should be transferred. This is not the case with externalities. For example, with air pollution it may not be clear that the polluter should pay, with the implicit recognition of the right to clean air, or whether there is a right to pollute, with clean air something that should have to be paid for. This leaves the direction in which payment should go unclear. Without clearly specified property rights, the bargaining envisaged in the Coase theorem does not have a firm foundation: neither party would willingly accept that they were the party that should pay.

If the exchange of commodities would lead to mutually beneficial gains for two parties, the commodities will be exchanged unless the cost of doing so outweighs the benefits. Such transactions costs may arise from the need for the parties to travel to a point of exchange or from the legal costs involved in formalizing the transactions. They may also arise due to the search required to find a trading partner. Whenever they arise, transactions costs represent a hindrance to trade and, if sufficiently great, will lead to no trade at all taking place. The latter results in the economy having a missing market.

The existence of transactions costs is often seen as the most significant reason for the nonexistence of markets in externalities. To see how they can arise, consider the problem of pollution caused by car emissions. If the reasoning of the Coase theorem is applied literally, then any driver of a car must purchase pollution rights from all of the agents that are affected by the car emissions each time, and every time, that the car is used. Obviously this would take an absurd amount of organization, and since
considerable time and resources would be used in the process, transactions costs would be significant. In many cases it seems likely that the welfare loss due to the waste of resources in organizing the market would outweigh any gains from having the market.

When external effects are traded, there will generally only be one agent on each side of the market. This thinness of the market undermines the assumption of competitive behavior needed to support the efficiency hypothesis. In such circumstances the Coase theorem has been interpreted as implying that bargaining between the two agents will take place over compensation for external effects and that this bargaining will lead to an efficient outcome. Such a claim requires substantiation.

Bargaining can be interpreted as taking the form of either a cooperative game between agents or as a noncooperative game. When it is viewed as cooperative, the tradition since Nash has been to adopt a set of axioms that the bargain must satisfy and to derive the outcomes that satisfy these axioms. The requirement of Pareto-efficiency is always adopted as one of the axioms so that the bargained agreement is necessarily efficient. If all bargains over compensation payments were placed in front of an external arbitrator, then the Nash bargaining solution would have some force as descriptive of what such an arbitrator should try and achieve. However, this is not what is envisaged by the Coase theorem, which focuses on the actions of markets free of any regulation. Although appealing as a method for achieving an outcome agreeable to both parties, the fact that Nash bargaining solution is efficient does not demonstrate the correctness of the Coase theorem.

The literature on bargaining in a noncooperative context is best divided between games with complete information and those with incomplete information, since this distinction is of crucial importance for the outcome. One of the central results of noncooperative bargaining with complete information is due to Rubinstein who considers the division of a single object between two players. The game is similar to the fundraising game presented in the public goods chapter. The players take turns to announce a division of the object, and each period an offer and an acceptance or rejection are made. Both players discount the future, so they are impatient to arrive at an agreed division. Rubinstein shows that the game has a unique (subgame perfect) equilibrium with agreement reached in the first period. The outcome is Pareto-efficient.

The important point is the complete information assumed in this representation of bargaining. The importance of information for the nature of outcomes will be extensively analyzed in chapter 10, and complete information is equally important for bargaining. In the simple bargaining problem of Rubinstein the information that must be known are the preferences of the two agents, captured by their rates of time discount. When these discount rates are private information, the attractive properties of the complete
information bargain are lost, and there are many potential equilibria whose nature is
dependent on the precise specification of the structure of bargaining.

In the context of externalities it seems reasonable to assume that information will be
incomplete, since there is no reason why the agents involved in bargaining an agreement
over compensation for an external effect should be aware of each other’s valuations
of the externality. When they are not aware, there is always the incentive to try to
exploit a supposedly weak opponent or to pretend to be strong and make excessive
demands. This results in the possibility that agreement may not occur even when it is
in the interests of both parties to trade.

To see this more clearly, consider the following bargaining situation. There are two
agents: a polluter and a pollutee. They bargain over the decision to allow or not the
pollution. The pollutee cannot observe the benefit of pollution $B$ but knows that it is
drawn from a distribution $F(B)$, which is the probability that the benefit is less than or
equal to $B$. Similarly the polluter cannot observe the cost of pollution $C$ but knows that
it is drawn from a distribution $G(C)$. Obviously the benefit is known to the polluter and
the cost is known to the pollutee. Let us give the property rights to the pollutee so that he
has the right to a pollution-free environment. Pareto-efficiency requires that pollution
be allowed whenever $B \geq C$. Now the pollutee (with all the bargaining power) can
make a take-it-or-leave-it offer to the polluter. What will be the bargaining outcome?

The pollutee will ask for compensation $T > 0$ (since $C > 0$) to grant permission
to pollute. The polluter will only accept to pay $T$ if his benefit from polluting exceeds
the compensation he has to pay, so $B \geq T$. Hence the probability that the polluter will
accept the offer is equal to $1 - F(T)$, that is, the probability that $B \geq T$. The best deal
for the pollutee is to ask for compensation that maximizes her expected payoff defined
as the probability that the offer is accepted times the net gain if the offer is accepted.
Therefore the pollutee asks for compensation $T^*$, which solves

$$\max_{\{T\}} \left(1 - F(T)\right) [T - C]. \quad (8.21)$$

Clearly, the optimal value, $T^*$, is such that

$$T^* > C. \quad (8.22)$$

But then bargaining can result (with strictly positive probability) in an inefficient out-
come. This is the case for all realizations of $C$ and $B$ such that $C < B < T^*$, which
implies that the offer is rejected (since the compensation demanded exceeds the benefit)
and thus pollution is not allowed, while Pareto-efficiency requires permission to pollute
to be granted (since its cost is less than its benefit).
The efficiency thesis of the Coase theorem relies on agreements being reached on the compensation required for external effects. The results above suggest that when information is incomplete, bargaining between agents will not lead to an efficient outcome.

8.9 Nonconvexity

One of the basic assumptions that supports economic analysis is that of convexity. Convexity gives indifference curves their standard shape, so consumers always prefer mixtures to extremes. It also ensures that firms have nonincreasing returns so that profit maximization is well defined. Without convexity, many problems arise with the behavior of the decisions of individual firms and consumers, and with the aggregation of these decisions to find an equilibrium for the economy.

Externalities can be a source of nonconvexity. Consider the case of a negative production externality. The panel at the left in figure 8.10 displays a firm whose output is driven to zero by an externality regardless of the level of other inputs. An example would be a fishery where sufficient pollution of the fishing ground by another firm can kill all the fish. In the panel at the right side of the figure, a zero output level is not reached, but output tends to zero as the level of the externality is increased. In both situations the production set of the firm is not convex.

In either case the economy will fail to have an equilibrium if personalized taxes are employed in an attempt to correct the externality. Suppose that the firm were to receive a subsidy for accepting externalities. Its profit-maximizing choice would be to produce

![Figure 8.10](#) Nonconvexity
an output level of zero and to offer to accept an arbitrarily large quantity of externalities. Since its output is zero, the externalities can do it no further harm, so this plan will lead to unlimited profits. If the price for accepting externalities were zero, the same firm would not accept any. The demand for externalities is therefore discontinuous, and an equilibrium need not exist.

There is also a second reason for nonconvexity with externalities. It is often assumed that once all inputs are properly accounted for, all firms will have constant returns to scale, since behavior can always be replicated. That is, if a fixed set of inputs (i.e., a factory and staff) produces output $y$, doubling all those inputs must produce output $2y$, since they can be split into two identical subunits (e.g., two factories and staff) producing an amount $y$ each. Now consider a firm subject to a negative externality, and assume that it has constant returns to all inputs including the externality. From the perspective of society, there are constant returns to scale. Now let the firm double all its inputs, but with the externality held at a constant level. Since the externality is a negative one, it becomes diluted by the increase in other inputs, and output must more than double. The firm therefore faces private increasing returns to scale. With such increasing returns, the firm’s profit-maximizing decision may not have a well-defined finite solution and market equilibrium may again fail to exist.

These arguments provide some fairly powerful reasons why an economy with externalities may not share some of the desirable properties of economies without. The behavior that follows from nonconvexity can prevent some of the pricing tools that are designed to attain efficiency from functioning in a satisfactory manner. At worst, nonconvexity can even cause there to be no equilibrium in the economy.

8.10 Conclusions

Externalities are an important feature of economic activity. They can arise at a local level between neighbors and at a global level between countries. The existence of externalities can lead to inefficiency if no attempt is made to control their level. The Coase theorem suggests that well-defined property rights will be sufficient to ensure that private agreements can resolve the externality problem. In practice, property rights are not well defined in many cases of externality. Furthermore the thinness of the market and the incomplete information of market participants result in inefficiencies that undermine the Coase theorem.

The simplest policy solution to the externality problem is a system of corrective Pigouvian taxes. If the tax rate is proportional to the marginal damage (or benefit)
caused by the externality, then efficiency will result. However, for this argument to apply when there are many consumers and firms requires that the taxes be so differentiated between economic agents that they become equivalent to a system of personalized prices. The optimal system then becomes impractical due to its information limitations. An alternative policy response is the use of marketable licenses that limit the emission of externalities. Licenses have some administrative advantages over taxes and will produce the same outcome when costs and benefits are known with certainty. With uncertainty, licenses and taxes have different effects and combining the two can lead to a superior outcome.

Further Reading

The classic analysis of externalities is in:

The externality analysis is carried further in a more rigorous and complete treatment in:

A persuasive argument for the use of corrective taxes is in:

The problem of social cost and the bargaining solution with many legal examples is developed in:

An illuminating classification of externalities and nonmarket interdependences is in:

A comprehensive and detailed treatment of the theory of externalities can be found in:

The efficient noncooperative bargaining solution with perfect information is in:

The general theory of bargaining with complete and incomplete information and many applications is in:

An extremely simple exposition of the conflict between individual motives and collective efficiency is in:

The bandwagon effect and technology adoption is in:


A summary of the arguments on the Tragedy of the Commons appears first in:


The nonconvexity problem with externalities was first pointed out in:


Exercises

8.1 “Smoke from a factory dirties the local housing and poisons crops.” Identify the nature of the externalities in this statement.

8.2 How would you describe the production function of a laundry polluted by a factory?

8.3 Let \( U = [x_1]^{\alpha} [x_2 y]^{1-\alpha} \), where \( y \) is an externality. Is this externality positive or negative? How does it affect the demand for good 1 relative to the demand for good 2?

8.4 The two consumers in the economy have preferences \( U_1 = [x_1^{\alpha}] [x_2 y^{1-\alpha}] \) and \( U_2 = [x_1^{\alpha}] [x_2^{1-\alpha}] \), where \( x_i \) is consumption of good \( i \) by consumer \( h \). Show that the equilibrium is efficient despite the externality. Explain this conclusion.

8.5 Consider a group of \( n \) students. Suppose that each student \( i \) puts in \( h_i \) hours of work on her classes that involves a disutility of \( h_i^2 \). Her benefits depend on how she performs relative to her peers and take the form \( u(h_i) \) for all \( i \), where \( \bar{h} = \frac{1}{n} \sum h_i \) denotes the average number of hours put in by all students in the class and \( u(\cdot) \) is an increasing and concave function.

a. Calculate the symmetric Nash equilibrium.

b. Calculate the Pareto-efficient level of effort.

c. Explain why the equilibrium involves too much effort compared to the Pareto-efficient outcome.

8.6 There are 4 students registered for the class “Introduction to externalities” at the University of Life. The professor is lazy so decides to implement a simple grading system. No lectures are given, no exercises are set, but there is a final paper. The paper requests each student, \( i \), to choose a single number \( z_i \geq 0 \) and record the choice on the answer sheet. The next day the final grade \( G^i \), with a maximum of 20 for each student, is released using the following rule (which is public information before the exam): \( G^i = 10 + \sqrt{z_i} - \frac{1}{2} \sum_{i=1}^{4} z_i \). The exam is passed if a grade of 10 out of 20, or more, is obtained.

a. State the maximization problem faced by each student.

b. If each student is maximizing his expected grade, what is the desired grade chosen?
c. If each student is maximizing his expected grade, what is the final grade received by each student? (Assume that the equilibrium is symmetric.)

d. Is the outcome Pareto efficient? If not, provide an example of an outcome that is a Pareto improvement.

8.7 Graduate student A smokes, but his office mate B hates smoking. A and B have the following utility functions: 
\[ U^A = 100 + 10z - 0.1z^2 \]  and \[ U^B = 100 - 10z \], where \( z \) is the number of cigarettes smoked by A (and \( U^A \) includes the cost of cigarettes). Determine:

a. The number of cigarettes smoked by A when the external effect on B is ignored.

b. The socially optimal level of cigarettes that should be smoked by A.

c. The optimal Pigouvian tax needed to decentralize the social optimum.

d. The outcome with Coasian bargaining when the property right is assigned to the smoker.

e. The outcome with Coasian bargaining when the property right is assigned to the nonsmoker.

8.8 Consider two firms, X and Y, located on the same river bank. The two firms produce paper for a printing works, and the production of paper requires the use of chemicals. Firm X is located upstream from firm Y and has the production function \( x = 1,000 \ell^X \) per day, and the amount of paper produced by firm X. Firm Y, located downstream, has a similar production function, but its output can be affected by the chemical waste produced by firm X. Hence firm Y has the following production function:

\[ y = \begin{cases} 
1000 \ell^Y \frac{1}{2} (x - x_0)^{-\alpha} & \text{if } x > x_0, \\
1000 \ell^Y \frac{1}{2} & \text{if } x < x_0, 
\end{cases} \]

where \( x_0 \) is the natural capacity for absorption of pollution by the river, \( \ell^Y \) the labor used, and \( y \) the amount of paper produced by the firm. The selling price of a unit of paper is \( p = 1 \) and the wage per unit of labor is \( w = 50 \). Each firm is a profit maximizer.

a. Assume \( \alpha = 0 \). Determine the quantity of paper produced by each firm in equilibrium.

b. Assume \( x_0 = 19,000 \) and \( \alpha = 0.1 \). How does this modify the result in part a? What is the interpretation of \( \alpha \)?

c. Assume that the two firms merge. Does the new firm have an incentive to reallocate some labor from X to Y? Explain the result.

8.9 There is a large number of commuters who decide to use either their car or the train. Commuting by train takes 70 minutes whatever the number of commuters taking the train. Commuting by car takes \( C(x) = 20 + 60x \) minutes, where \( x \) is the proportion of commuters taking their cars, \( 0 \leq x \leq 1 \).

a. Plot the curves of the commuting time by car and the commuting time by train as a function of the proportion of car users.

b. What is the proportion of commuters who will take their car if everyone is taking her decision freely and independently so as to minimize her own commuting time?

c. What is the proportion of car users that minimizes the total commuting time?

d. Compare this with your answer given in part b. Interpret the difference. How large is the deadweight loss from the externality?
e. Explain how a toll could achieve the efficient allocation of commuters between train and car and be beneficial for everyone.

8.10 Re-do the previous problem by replacing the train by a bus and assuming that commuting time by bus is increasing with the proportion of commuters using car (traffic congestion). Let the commuting time by bus be \( B(x) = 40 + 20x \) and the commuting time by car be \( C(x) = 20 + 60x \), where \( x \) is the proportion of commuters taking their car, \( 0 \leq x \leq 1 \).

8.11 Consider a binary choice to allow or not the emission of pollutants. The cost to consumers of allowing the pollution is \( C = 2,000 \), but this cost is only observable to the consumers. The benefit for the polluter of allowing the externality is \( B = 2,500 \), and only the polluter knows this benefit. Clearly, optimality requires this externality to be allowed, since \( B > C \). However, the final decision must be based on what each party chooses to reveal.

a. Construct a tax-subsidy revelation scheme such that it is a dominant strategy for each party to report truthfully their private information.

b. Show that this revelation scheme induces the optimal production of the externality.

c. Show that this revelation scheme is unbalanced in the sense that the given equilibrium reports the tax to be paid by the polluter is less than the subsidy paid to the pollutee.

8.12 How can licenses be used to resolve the tragedy of the commons?

8.13 If insufficient abatement is very costly, which of taxation or licenses is preferable?

8.14 Are the following statements true or false? Explain why.

a. If your consumption of cigarettes produces negative externalities for your partner (which you ignore), then you are consuming more cigarettes than is Pareto efficient.

b. It is generally efficient to set an emission standard allowing zero pollution.

c. A tax on cigarettes induces the market for cigarettes to perform more efficiently.

d. A ban on smoking is necessarily efficient.

e. A competitive market with a negative externality produces more output than is efficient.

f. A snob effect is a negative (network) externality from consumption.

8.15 A chemical factory produces a product that is sold at the price of $10 per ton. The cost of production is \( CF(y) = 0.5y^2 \), where \( y \) is the number of tons of production. For each ton of production, the factory produces 1 kg of pollutants that are either dispersed into the atmosphere at a cost of $0 to the factory or captured and stored at a cost of $2 for each kg of pollutant. The amount of dispersant pollutant is denoted \( e \). The pollution causes a nearby textile firm to have to wash its products twice with an additional cost of \( C_e = 0.5e^2 \), which does not depend on the amount of output, \( m \), produced. The unit price of the output of the textile firm is $5, and the cost of production \( CT = 0.02m^2 \).

a. Determine the values of \( y \), \( m \), and \( e \) in the competitive equilibrium. How much does the chemical factory spend on capturing the pollutant? What level of environmental cost does the factory inflict on the textile firm?

b. Why is the competitive equilibrium inefficient from a social perspective?

c. What is the socially optimal level of \( y \)?
d. Assume that the externality is internalized. Determine the values of $y$, $m$, and $e$, the expenditure on capture, and the additional cost for the textile firm caused by the environmental pollution.

8.16 Consider two consumers with utility functions

$$U^A = \log(x_1^A) + x_2^A - \frac{1}{2} \log(x_1^B), \quad U^B = \log(x_1^B) + x_2^B - \frac{1}{2} \log(x_1^A).$$

Both consumers have income $M$, and the (before-tax) price of both goods is 1.

a. Calculate the market equilibrium.

b. Calculate the social optimum for a utilitarian social welfare function.

c. Show that the optimum can be sustained by a tax placed on good 1 (so that the after-tax price becomes $1 + t$) with the revenue returned equally to the consumers in a lump-sum manner.

d. Assume now that preferences are given by

$$U^A = \rho^A \log(x_1^A) + x_2^A - \frac{1}{2} \log(x_1^B), \quad U^B = \rho^B \log(x_1^B) + x_2^B - \frac{1}{2} \log(x_1^A).$$

Calculate the taxes necessary to decentralize the optimum.

e. For preferences of part d and income $M = 20$, contrast the outcome when taxes can and cannot be differentiated between consumers.

8.17 A competitive refining industry releases one unit of waste into the atmosphere for each unit of refined product. The inverse demand function for the refined product is $p^d = 20 - q$, which represents the marginal benefit curve where $q$ is the quantity consumed when the consumers pay price $p^d$. The inverse supply curve for refining is $MPC = 2 + q$, which represents the marginal private cost curve when the industry produces $q$ units. The marginal external cost curve is $MEC = 0.5q$, where $MEC$ is the marginal external cost when the industry releases $q$ units of waste. Marginal social cost is given by $MSC = MPC + MEC$.

a. What are the equilibrium price and quantity for the refined product when there is no correction for the externality?

b. How much of the refined product should the market supply at the social optimum?

c. How large is the deadweight loss from the externality?

d. Suppose that the government imposes an emission fee of $T$ per unit of emissions. How large must the emission fee be if the market is to produce the socially efficient amount of the refined product?

8.18 Discuss the following statement: “A tax is a fine for doing something right. A fine is a tax for doing something wrong.”

8.19 Suppose that the government issues tradable pollution permits.

a. Is it better for economic efficiency to distribute the permits among polluters or to auction them?

b. If the government decides to distribute the permits, does the allocation of permits among firms matter for economic efficiency?
A chemical producer dumps toxic waste into a river. The waste reduces the population of fish, reducing profits for the local fishery industry by $150,000 per year. The firm could eliminate the waste at a cost of $100,000 per year. The local fishing industry consists of many small firms.

a. Apply the Coase theorem to explain how costless bargaining will lead to a socially efficient outcome, no matter to whom property rights are assigned (either to the chemical firm or the fishing industry).

b. Verify the Coase theorem if the cost of eliminating the waste is doubled to $200,000 (with the benefit for the fishing industry unchanged at $150,000).

c. Discuss the following argument: “A community held together by ties of obligation and mutual interest can manage the local pollution problems.”

d. Why might bargaining not be costless?

A firm, $S$, produces steel but also produces waste that contaminates a nearby river. Steel can be sold for $10 per ton. The cost function of the steel firm is given by $c_S(s) = s^2$, where $s$ is the output of steel in tons. The level of waste, $x$, is related to output by $x = 0.1s$. A fish farm, $F$, is located downstream and is negatively affected by the waste polluting the water. Every fish produced by the farm can be sold for $2$. The cost function of the fish farm is given by $c_F(f; x) = f + x^2$. The fish farm has a capacity constraint $f \leq 10$.

a. Compute the optimal output of the steel firm if it makes its decision without any constraint. Assume now that the firm must compensate the fishing club members an amount $q$ per unit of waste that is produced.

b. Compute the optimum value of $q$, the optimal steel output, and the profit levels of the steel firm and the fishery.

c. Compare the solutions to parts a and b, and explain any differences.

It is often used as an objection to market-based policies of pollution abatement that they place a monetary value on cleaning up our environment. Economists reply that society implicitly places a monetary value on environmental cleanup even under command-and-control policies. Explain why this is true.

Use examples to answer whether the externalities related to common resources are generally positive or negative. Is the free-market use of common resources greater or less than the socially optimal use?

Why is there more litter along highways than in people’s yards?

Evaluate the following statement: “Since pollution is bad, it would be socially optimal to prohibit the use of any production process that creates pollution.”

Why is it not generally efficient to set an emissions standard allowing zero pollution?

Education is often viewed as a good with positive externalities.

a. Explain how education might produce positive external effects.

b. Suggest a possible action of the government to induce the market for education to perform more efficiently.
9 Imperfect Competition

9.1 Introduction

The analysis of economic efficiency in chapter 2 demonstrated the significance of the competitive assumption that no economic agent has the ability to affect market prices. Under this assumption prices reveal true economic values and act as signals that guide agents to mutually consistent decisions. As the Two Theorems of Welfare Economics showed, they do this so well that Pareto-efficiency is attained. Imperfect competition arises whenever an economic agent has the ability to influence prices. To be able to do so requires that the agent be large relative to the size of the market in which they operate. It follows from the usual application of economic rationality that those agents who can affect prices will aim to do so to their own advantage. This must be detrimental to other agents and to the economy as a whole. This basic feature of imperfect competition, and its implications for economic policy, will be explored in this chapter.

Imperfect competition can take many forms. It can arise due to monopoly in product markets and through monopsony in labor markets. Firms with monopoly power will push prices above marginal cost in order to raise their profits. This will reduce the equilibrium level of consumption below what it would have been had the market been competitive and will transfer surplus from consumers to the owners of the firm. Unions with monopoly power can ensure that the wage rate is increased above its competitive level and secure a surplus for their members. The increase in wage rate reduces employment and output. Firms (and even unions) can engage in non-price competition by choosing the quality and characteristics of their products, undertaking advertising, and blocking the entry of competitors.

Each of these forms of behavior can be interpreted as an attempt to increase market power and obtain a greater surplus. When they can occur, the assumption of price-taking behavior used to prove the Two Theorems is violated, and an economy with imperfect competition will not achieve an efficient equilibrium (with one special exception as is detailed later). It then becomes possible that policy intervention can improve on the unregulated outcome. The purpose of this chapter is to investigate how the conclusions derived in earlier chapters need to be modified and to look at some additional issues specific to imperfect competition.
The first part of the chapter focuses on imperfect competition in product markets. After categorizing types of imperfect competition, defining the market structure, and measuring the intensity of competition, the failure of efficiency is demonstrated when there is a lack of competition. This is followed by a discussion of tax incidence in competitive and imperfectly competitive markets. The effects of specific and ad valorem taxes are then distinguished, and their relative efficiency is assessed. The policies used to regulate monopoly and oligopoly in practice are also described. There is next a discussion of the recent European policy on the regulation of mergers. The final part of the chapter focuses on market power on the two sides of the labor market. Market power from the supply side (monopoly power of a labor union) is contrasted with monopsony power from the demand side. It is shown that both cases lead to inefficient underemployment with wages, respectively above and below competitive wages.

9.2 Concepts of Competition

Imperfect competition arises whenever an economic agent exploits the fact that they have the ability to influence the price of a commodity. If the influence on price can be exercised by the sellers of a product, then there is monopoly power. If it is exercised by the buyers, then there is monopsony power, and if by both buyers and sellers, there is bilateral monopoly. A single seller is a monopolist and a single buyer a monopsonist. Oligopoly arises with two or more sellers who have market power, with duopoly being the special case of two sellers.

An agent with market power can set either the price at which it sells, with the market choosing quantity, or the quantity it supplies, with the market determining price. When there is either monopoly or monopsony, it does not matter whether price or quantity is chosen: the equilibrium outcome will be the same. If there is more than one agent with market power, then the choice variable does make a difference. In oligopoly markets Cournot behavior refers to the use of quantity as the strategic variable and Bertrand behavior to the use of prices. Typically Bertrand behavior is more competitive in that it leads to a lower market price. Entry by new firms may either be impossible, so that an industry is composed of a fixed number of firms, and not be hindered, or incumbent firms may follow a policy of entry deterrence.

Forms of imperfect competition also vary with respect to the nature of products sold. Products may be homogeneous, so that the output of different firms is indistinguishable by the consumer, or differentiated, so that each firm offers a different variant. With homogeneous products, at equilibrium there must be a single price in the market. Product
differentiation can either be vertical (whereby products can be unambiguously ranked in terms of quality) or horizontal (whereby consumers differ in which specification they prefer). Equilibrium prices can vary across specifications in markets with differentiated products. The notion of product differentiation captures the idea that consumers make choices among competing products on the basis of factors other than price. The exact nature of the differentiation is very important for the market outcome. What differentiation implies is that purchases of a product do not fall off to zero when its price is raised above that of competing products. The greater the differentiation, the lower is the willingness of consumers to switch among sellers when one seller changes its price. The theory of monopolistic competition relates to this competition among many differentiated sellers who can enjoy some limited monopoly power if tastes differ markedly from one consumer to the next.

When products are differentiated, firms may engage in non-price competition. This is the use of variables other than price to gain profit. For example, firms may compete by choosing the specification of their product and the quantity of advertising used to support it. The level of investment can also be a strategic variable if this can deter entry by making credible a threat to raise output.

To limit the number of cases to be considered, this chapter will focus on Cournot behavior, so that quantity is the strategic variable, with homogeneous products. Although only one of many possible cases, this perfectly illustrates most of the significant implications of imperfect competition. It also has monopoly as a special case (when there is a single firm) and competition as another (when the number of firms tends to infinity).

9.3 Market Structure

The structure of the market describes the number and size of firms that compete within it and the intensity of this competition. To describe the structure of the market, it is first necessary to define the market.

9.3.1 Defining the Market

A market consists of the buyers and sellers whose interaction determines the price and quantity of the good that is traded. Generally, two sellers will be considered to be in the same market if their products are close substitutes. Measuring the own-price elasticity of demand for a product tells us whether there are close substitutes available, but it does not identify what those substitutes might be. To identify the close substitutes, one
must study cross-price elasticities of demand between products. When the cross-price elasticity is positive, it indicates that consumers are increasing their demand for one good as the price of the other good increases. The two products are thus close substitutes. Another approach to defining markets is to use the standard industry classification that identifies products as close competitors if they share the same product characteristics. Although products with the same classification number are often close competitors, this is not always true. For example, all drugs share the same classification number but not all drugs are close substitutes for each other.

Markets are also defined by geographic areas, since otherwise identical products will not be close substitutes if they are sold in different areas and the cost of transporting the product from one area to another is large. Given this reasoning, one would expect close competitors to locate as far as possible from each other, and it therefore may seem peculiar to see them located close to one another in some large cities. This reflects a common trade-off between market size and market share. For instance, antique stores in Brussels are located next to one another around the Place du Grand Sablon. The reason is that the bunching effect helps to attract customers in the first place (market size), even if they become closer competitors in dividing up the market (market sharing). By locating close together, Brussels’ antique stores make it more convenient for shoppers to come and browse around in search of some antiques. In other words, the bunching of sellers creates a critical mass that makes it easier to attract shoppers.

9.3.2 Measuring Competition

We now proceed on the basis that the market has been defined. What does it then mean to say that there is “more” or “less competition” in this market? Three distinct dimensions are widely used and need to be clearly distinguished.

The first dimension is *contestability*, which represents the freedom of rivals to enter an industry. It depends on legal monopoly rights (patent protection, operating licenses, etc.) or other barriers to entry (economies of scale and scope, the marketing advantage of incumbents, entry-deterring strategies, etc.). Entry barriers protect the market leader from serious competition from newcomers. Contestability theory shows how the threat of entry can constrain incumbents from raising prices even if there is only one firm currently operating in the market. However, when markets are not perfectly contestable, the threat of potential competition is limited, which allows the incumbents to reap additional profits.

A second dimension is the degree of *concentration* that represents the number and distribution of rivals currently operating in the same market. As we will see, the
performance of a market depends on whether it is concentrated (having few sellers) or unconcentrated (having many sellers). A widespread measure of market concentration is the \( n \)-firm concentration ratio. This is defined as the consolidated market share of the \( n \) largest firms in the market. For example, the four-firm concentration ratio in the US cigarette industry is 0.92, which means that the four largest cigarette firms have a total market share of 92 percent (with the calculation of market share usually based on sales revenue). Table 9.1 shows the four-firm concentration ratios for some US industries in 1987.

The problem with the \( n \)-firm concentration ratio is that it is insensitive to the distribution of market shares among the largest firms. For example a four-firm concentration ratio does not change if the first-largest firm increases its market share at the expense of the second-largest firm. To capture the relative size of the largest firms, another commonly used measure is the Herfindahl index. This index is defined as the sum of the squared market shares of all the firms in the market. Letting \( s_i \) be the market share of firm \( i \), the Herfindahl index is given by \( H = \sum_i s_i^2 \). Notice that the Herfindahl index in a market with two equal-size firms is \( \frac{1}{4} \) and with \( n \) equal-size firms is \( \frac{1}{n} \). For this reason a market with Herfindahl index of 0.20 is also said to have a number’s equivalent of 5. For example, if there is one dominant firm with a market share of 44 percent and 100 identical small firms with a total market share of 56 percent, the Herfindahl is

\[
H = \sum_i s_i^2 = (0.44)^2 + 100 \left( \frac{0.56}{100} \right)^2 = 0.197. \tag{9.1}
\]

Table 9.1

<table>
<thead>
<tr>
<th>Industry</th>
<th>Number of firms</th>
<th>Four-firm concentration ratio</th>
<th>Herfindahl index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cereal breakfast foods</td>
<td>33</td>
<td>0.87</td>
<td>0.221</td>
</tr>
<tr>
<td>Pet food</td>
<td>130</td>
<td>0.61</td>
<td>0.151</td>
</tr>
<tr>
<td>Book publishing</td>
<td>2,182</td>
<td>0.24</td>
<td>0.026</td>
</tr>
<tr>
<td>Soap and detergents</td>
<td>683</td>
<td>0.65</td>
<td>0.170</td>
</tr>
<tr>
<td>Petroleum refining</td>
<td>200</td>
<td>0.32</td>
<td>0.044</td>
</tr>
<tr>
<td>Electronic computers</td>
<td>914</td>
<td>0.43</td>
<td>0.069</td>
</tr>
<tr>
<td>Refrigerators/freezers</td>
<td>40</td>
<td>0.85</td>
<td>0.226</td>
</tr>
<tr>
<td>Laundry machines</td>
<td>11</td>
<td>0.93</td>
<td>0.286</td>
</tr>
<tr>
<td>Greeting cards</td>
<td>147</td>
<td>0.85</td>
<td>0.283</td>
</tr>
</tbody>
</table>

This market structure is then interpreted as being equivalent to one with 5 identical firms. Herfindahls associated to some US industries are indicated in table 9.1. These numbers show that the market for laundry firms, which has a number’s equivalent less than 4, is more concentrated than the market for book publishers, which has a number’s equivalent of 38.

The third dimension of the market structure is *collusiveness*. This is related to the degree of independence of firms’ strategies within the market or its reciprocal, which is the possibility for sellers to agree to raise prices in unison. Collusion can either be explicit (e.g., a cartel agreement) or tacit (when it is in each firm’s interest to refrain from aggressive price cutting). Explicit collusion is illegal and more easily detected than tacit collusion. However, tacit collusion is more difficult to sustain. Experience has shown that it is unusual for more than a handful of sellers to raise prices much above costs for a sustained period. One common reason is that a small firm may view the collusive bargain among larger rivals as an opportunity to steal their market shares by undercutting the collusive price, which in turn triggers a price war. The airline industry is a good example in recent years of frequent price wars. The additional problem with the airline industry is that fixed cost is high relative to variable cost. This means that once a flight is scheduled, airlines face tremendous pressure to fill their planes, and they are willing to fly passengers at prices close to marginal cost but far below average cost. Thus with such pricing practices, airlines can take large financial losses during price wars.

The three dimensions of market structure and the resulting intensity of competition may be related. The freedom to enter a market may result in a larger number of firms operating and thus a less concentrated market, which in turn may lead to the breakdown of collusive agreement to raise prices.

### 9.4 Welfare

Imperfect competition, along with public goods, externalities, and asymmetric information, is one of the standard forms of market failure that leads to the inefficiency of equilibrium. It is the inefficiency that provides the motivation for economic policy in relation to imperfect competition. To provide the context for the discussion of policy, this section demonstrates the source of the inefficiency and reports measures of its extent.

#### 9.4.1 Inefficiency

The most important fact about imperfect competition is that it invariably leads to inefficiency. The cause of this inefficiency is now isolated in the profit-maximizing
behavior of firms that have an incentive to restrict output so that price is increased above the competitive level.

In a competitive economy equilibrium will exist where the price of each commodity is equal to its marginal cost of production. This is due to the argument that firms will always move to increase supply whenever price is above marginal cost, since price is taken as given and additional supply will raise profit. As all firms raise supply, prices will fall until there is no incentive for further supply increases. This argument shows that the profit-maximizing behavior of competitive firms drives price down to marginal cost. If marginal cost is constant at value $c$, then competition results in a price, $p$, satisfying

$$p = c. \quad (9.2)$$

To see the cause of inefficiency with imperfect competition, consider first the case of monopoly. Assume that the monopolist produces with a constant marginal cost, $c$, and chooses its output level, $y$, to maximize profit. The market power of the monopolist is reflected in the fact that as its output is increased, the market price of the product will fall. This relationship is captured by the inverse demand function, $p(y)$, which determines price as a function of output. As $y$ increases, $p(y)$ decreases. Using the inverse demand function, which the monopolist is assumed to know, we have the profit level of the firm as

$$\pi = [p(y) - c] y. \quad (9.3)$$

The first-order condition describing the profit-maximizing output level is

$$p + y \frac{dp}{dy} - c = 0, \quad (9.4)$$

which, since $\frac{dp}{dy} < 0$ (price falls as output increases), implies that $p > c$. The condition in (9.4) shows that the monopolist will set price above marginal cost and that the monopolist’s price does not satisfy the efficiency requirement of being equal to marginal cost. The fact that the monopolist perceives that its output choice affects price (so $\frac{dp}{dy}$ is not zero) is reflected directly in the divergence of price and marginal cost.

The condition describing the choice of output can be re-arranged to provide further insight into degree of divergence between price and marginal cost. Using the elasticity of demand, $\varepsilon = \frac{dy}{dp} \frac{p}{y} < 0$, the profit-maximization condition can be written as

$$\frac{p - c}{p} = \frac{1}{|\varepsilon|}. \quad (9.5)$$
This equilibrium condition for the monopoly is called the **inverse elasticity pricing rule**. In words, the condition says that the percentage deviation between the price and the marginal cost is equal to the inverse of the elasticity of demand. The expression \( \frac{p - c}{p} \) is the **Lerner index**. The Lerner index will be shown shortly to be strictly between zero and one (i.e., \( |\varepsilon| > 1 \)). The monopoly pricing rule can also be written as

\[
p = \mu c,
\]

where \( \mu = \frac{1}{1 - (1/|\varepsilon|)} > 1 \) is called the **monopoly markup** and measures the extent to which price is raised above marginal cost. This pricing rule shows that the markup above marginal cost is inversely related to the absolute value of the elasticity of demand. The higher the absolute value of the elasticity, the smaller is the monopoly markup.

In the extreme case of perfectly elastic demand, which equates to the firm having no market power, price would be equal to marginal cost. For the markup \( \mu \) to be finite (i.e., price is well defined), it must be the case that \( |\varepsilon| > 1 \) so that the monopolist locates on the elastic part of the demand curve. If demand is inelastic, with \( |\varepsilon| \leq 1 \), then the monopolist makes maximum profit by selling the smallest possible quantity at an arbitrarily high price. Since the monopolist operates on the elastic part of the demand curve with \( |\varepsilon| > 1 \), the Lerner index, \( \frac{p - c}{p} = \frac{1}{|\varepsilon|} \in (0, 1) \), provides a simple measure of market power ranging from zero to one for maximal market power. Therefore a firm might have a monopoly, but its market power might still be low because it is constrained by competition from substitute products outside the market. By differentiating its product, a monopolist can insulate its product from the competition of substitute products and thereby expands its market power.

This relation of the monopoly markup to the elasticity of demand can be easily extended from monopoly to oligopoly. Assume that there are \( m \) firms in the market, and denote the output of firm \( j \) by \( y_j \). The market price is now dependent on the total output of the firms, \( y = \sum_{j=1}^{m} y_j \). With output level \( y_j \), the profit level of firm \( j \) is

\[
\pi_j = (p - c) y_j.
\]

(9.7)

Adopting the Cournot assumption that each firm regards its competitors’ outputs as fixed when it optimizes, the choice of output for firm \( j \) satisfies

\[
p + y_j \frac{dp}{dy} - c = 0.
\]

(9.8)

Now assume that the firms are identical and each produces the same output level, \( \frac{y}{m} \). The first-order condition for choice of output (9.8) can then be re-arranged to obtain the Lerner index.
Chapter 9: Imperfect Competition

\[
\frac{p - c}{p} = \frac{1}{m} \frac{1}{|\varepsilon|},
\]

(9.9)

and the oligopoly price is given by

\[
p = \mu c,
\]

(9.10)

where \(\mu = m \frac{1}{m - (1/|\varepsilon|)} > 1\) is the oligopoly markup. Thus, in the presence of several firms in the market, the Lerner index of market power is deflated according to the market share. As for monopoly, the value of the markup is related to the inverse of the elasticity of demand. The Lerner index can be used to show that an oligopoly becomes more competitive as the number of firms in the industry increases. This claim follows from the fact that \(\frac{p - c}{p}\) must tend to zero as \(m\) tends to infinity. Hence, as the number of firms increases, the Cournot equilibrium becomes more competitive and price tends to marginal cost. The limiting position with an infinite number of firms can be viewed as the idealization of the competitive model.

There is one special case of monopoly for which the equilibrium is efficient. Let the firm be able to charge each consumer the maximum price that the consumer is able to pay. To do so obviously requires the firm to have considerable information about its customers. The consequence is that the firm extracts all consumer surplus and translates it into profit. It will keep supplying the good until price falls to marginal cost and there is no more surplus to extract. So total supply will be equal to that under the competition. This scenario, known as perfect price discrimination, results in all the potential surplus in the market being turned into monopoly profit. No surplus is lost due to the monopoly, but all surplus is transferred from the consumers to the firm. Of course, this scenario can only arise with an exceedingly well-informed monopolist.

9.4.2 Incomplete Information

Monopoly inefficiency can also arise from the firm having incomplete information, even in situations where there would be efficiency with complete information. To see this, suppose that a monopolist with constant marginal cost \(c\) faces a buyer whose willingness to pay for a unit of the firm’s output is \(v\). If there is complete information, the firm and buyer will agree to a price between \(c\) and \(v\), and the product will be traded. The surplus from the transaction is shared between the two parties and no inefficiency arises.

The difference that imperfect information can make is that trade will sometimes not take place even though both parties would gain if they did trade. Suppose now that
the monopolist cannot observe \( v \) but knows from experience that it is drawn from a distribution \( F(v) \), which is the probability that the buyer’s valuation is less or equal to \( v \). The function \( (1 - F(v)) \) is analogous to the expected demand when a purchaser buys at most one unit because the probability that there is a demand at price \( v \) is the probability that the buyer’s valuation is higher than the price. Assume that there are potential gains from trade so \( v > c \) for at least a range of \( v \). Pareto-efficiency requires trade to occur if and only if \( v \geq c \).

The monopolist’s problem is to offer a price \( p \) that maximizes its expected profit (anticipating that the buyer will not accept the offer if \( v < p \)). This price must fall between \( c \) and \( v \) for trade to occur. The monopolist sets a price \( p^* \) that solves

\[
\max_{\{p\}} \left[ 1 - F(p) \right] (p - c). \tag{9.11}
\]

From the assumption that there is a potential gain from trade, there must be a range of values of \( v \) higher than \( c \), and thus it is possible for the monopolist to charge a price in excess of the marginal cost and the offer being accepted. Clearly, the price that maximizes expected profit must be \( p^* > c \), so the standard conclusion of monopoly holds that price is in excess of marginal cost. When trade takes place (so a value of \( v \) occurs with \( c < p^* < v \)), the outcome is an efficient trade. However, when a value of \( v \) occurs with \( c < v < p^* \), trade does not take place. This is inefficient because trade should occur when the benefit exceeds the cost (\( v > c \)). The effect of the monopolist setting price above marginal cost is to eliminate some of the potential trades.

For instance, assume that the willingness to pay \( v \) is uniformly distributed on the interval \([0, 1]\) with the marginal cost \( 0 < c < 1 \). Then the probability that trade takes place at price \( p \) (expected demand) is \( 1 - F(p) = 1 - p \), which gives expected revenue \( [1 - F(p)] p = [1 - p] p \) and marginal revenue \( MR = 1 - 2p \). The expected profit is \( \pi = [1 - p] (p - c) \), and the profit-maximizing pricing satisfies the first-order condition \( [1 - 2p] + c = 0 \), which can be re-arranged to give monopoly price of \( p^* = \frac{1 + c}{2} > c \). The parallel between this monopoly choice under incomplete information and the standard monopoly problem is illustrated in figure 9.1.

### 9.4.3 Measures of Welfare Loss

It has been shown that the equilibrium of an imperfectly competitive market is not Pareto-efficient, except in the special case of perfect price discrimination. This makes it natural to consider what the degree of welfare loss may actually be. The assessment
of monopoly welfare loss has been a subject of some dispute in which calculations have provided a range of estimates from the effectively insignificant to considerable percentages of potential welfare.

The inefficiency of monopoly will be described in chapter 12 and part of that argument is now briefly provided. Figure 9.2 assumes that the marginal cost of production is constant at value $c$ and that there are no fixed costs. The equilibrium price if the industry were competitive, $p^c$, would be equal to marginal cost, so $p^c = c$. This price leads to output level $y^c$ and generates consumer surplus $ADc$. The inverse demand function facing the firm, $p(y)$, determines price as a function of output and is also the average revenue function for the firm. This is denoted by $AR$. The marginal revenue function is denoted $MR$. The monopolist’s optimal output, $y^m$, occurs where marginal revenue and marginal cost are equal. At this output level the price with monopoly is $p^m$. Consumer surplus is $ABp^m$ and profit is $p^mBEc$.

Contrasting the competitive and the monopoly outcomes shows that some of the consumer surplus under competition is transformed into profit under monopoly. This is the area $p^mBEc$, and it represents a transfer from consumers to the firm. However, some of the consumer surplus is simply lost. This loss is the area $BDE$, which is termed the *deadweight loss of monopoly*. Since the total social surplus under monopoly ($ABp^m + p^mBEc$) is less than that under competition ($ADc$), the monopoly is inefficient.
This inefficiency is reflected in the fact that consumption is lower under monopoly than competition.

When the demand function is linear so that the AR curve is a straight line, the welfare loss area $BDE$ is equal to half of the area $p^m BEc$. The area $p^m BEc$ is monopoly profit, which is equal to $(p^m - c)y^m$. This implies that the loss $BDE$ is $\frac{1}{2}(p^m - c)y^m$. From the first-order condition for the choice of monopoly output, (9.5), $p^m - c = -\frac{1}{\varepsilon}p^m$. By this result it follows that a measure of the deadweight loss is

$$\text{Deadweight loss} = -\frac{p^m y^m}{2\varepsilon} = -\frac{R^m}{2\varepsilon},$$

(9.12)

where $R^m$ is the total revenue of the monopolist. This formula is especially simple to evaluate to obtain an idea of the size of the deadweight loss. For example, if the elasticity of demand is $-2$, then the welfare loss is 25 percent of sales revenue and is therefore quite large.

Numerous studies have been published that provide measures of the degree of monopoly welfare loss. A selection of these results is given in table 9.2. The smaller values are obtained by calculating only the deadweight loss triangle. If these were correct, then we could conclude that monopoly power is not a significant economic issue. This was the surprising conclusion of the initial study by Harberger in 1954, as it challenged the conventional wisdom that monopoly must be damaging to the economy. In contrast, the larger values of loss are obtained by including the costs of defending the
monopoly position. Chapter 12 considers the arguments proposed in the rent-seeking literature for the inclusion of these additional components of welfare loss. These values reveal monopoly loss to be very substantial.

It can be appreciated from table 9.2 that a broad range of estimates of monopoly welfare loss have been produced. Some studies conclude that welfare loss is insignificant; others conclude that it is very important. What primarily distinguishes these differing estimates is whether it is only the deadweight loss that is counted or the deadweight loss plus the cost of defending the monopoly. Which one is correct is an unresolved issue that involves two competing perspectives on economic efficiency.

There is one further point that needs to be made. The calculations above have been based on a static analysis in which there is a single time period. The demand function, the product traded, and the costs of production are all given. The firm makes a single choice and equilibrium is attained. What this ignores are all the dynamic aspects of economic activity such as investment and innovation. When these factors are taken into account, as Schumpeter forcefully argued, it is even possible for a monopoly to generate dynamic welfare gains rather than losses. This claim is based on the argument that investment and innovation will only be undertaken if firms can expect to earn a sufficient return. In a competitive environment, all gains are competed away, so the incentives to innovate are eliminated. Conversely, holding a monopoly position allows gains to be realized. This creates the incentive to invest and innovate. Furthermore the incentive is strengthened by the monopoly’s desire to maintain its strong hold. Dynamic gains can more than offset static losses, reinforcing the argument for allowing monopoly. We return to this issue in the discussion of regulation in section 9.7.

Table 9.2
Monopoly welfare loss

<table>
<thead>
<tr>
<th>Author</th>
<th>Sector</th>
<th>Welfare loss (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harberger</td>
<td>US manufacturing</td>
<td>0.08</td>
</tr>
<tr>
<td>Gisser</td>
<td>US manufacturing</td>
<td>0.11–1.82</td>
</tr>
<tr>
<td>Peterson and Connor</td>
<td>US food manufacturing</td>
<td>0.16–5.15</td>
</tr>
<tr>
<td>Masson and Shaanan</td>
<td>37 US industries</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
</tr>
<tr>
<td>McCorriston</td>
<td>UK agricultural inputs</td>
<td>1.6–2.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20–40</td>
</tr>
<tr>
<td>Cowling and Mueller</td>
<td>United States</td>
<td>4–13</td>
</tr>
<tr>
<td></td>
<td>United Kingdom</td>
<td>3.9–7.2</td>
</tr>
</tbody>
</table>
The study of tax incidence is about determining the changes in prices and profits that follow the imposition of a tax. The formal or legal incidence of a tax refers to who is legally responsible for paying the tax. The legal incidence can be very different from the economic incidence, which relates to who ultimately has to alter his behavior because of the tax.

To see this distinction, consider the following example. A tax of $1 is levied on a commodity that costs $10, and this tax must be paid by the retailer. The legal incidence is simple: for each unit sold the retailer must pay $1 to the tax authority. The economic incidence is much more complex. The first question has to be: What does the price of the commodity become after the tax? It may change to $11, but this would be an exception rather than the norm. It may, for example, rise instead only to $10.50. If it does, $0.50 of the tax falls on the consumer to pay. What of the other $0.50? This depends on how the producer responds to the tax increase. The producer may lower the price of the commodity to the retailer from $9 to $8.75 and then bear $0.25 of the tax. The remaining $0.25 of the tax is paid by the retailer. The economic incidence of the tax is then very distinct from the legal incidence.

This example raises the question of what determines the economic incidence. The answer is found in the demand and supply curves for the good that is taxed. Economic incidence will first be determined for the competitive case, and then it is shown how the conclusions are modified by imperfect competition. Imperfect competition can in fact result in very interesting conclusions concerning tax incidence.

Tax incidence analysis is at its simplest when there is competition and the marginal cost of production is constant. In this case the supply curve in the absence of taxation must be horizontal at a level equal to marginal cost; see figure 9.3. This gives the before-tax price \( p = c \). The introduction of a tax of amount \( t \) will raise this curve by exactly the amount of the tax. The after-tax price, \( q \), is at the intersection of the demand curve and the new supply curve. It can be seen that \( q = p + t \), so price will rise by an amount equal to the tax. Hence the tax is simply passed forward by the firms onto consumers, since price is always set equal to marginal cost plus tax.

When marginal cost is not constant and the supply curve slopes upward, the introduction of a tax still shifts the curve vertically upward by the amount equal to the tax. The extent to which price rises is then determined by the slopes of the supply and demand curve. If the demand curve is vertical, price rises by the full amount of the tax; otherwise, it will rise by less; see figure 9.4.
Figure 9.3
Tax incidence with perfectly elastic supply

Figure 9.4
Tax incidence in the general case
In summary, if the supply curve is horizontal (so supply is infinitely elastic) or the demand curve is vertical (so demand is completely inelastic), then price will rise by exactly the amount of the tax. In all other cases it will rise by less, with the exact rise being determined by the elasticities of supply and demand. When the price increase is equal to the tax, the entire tax burden is passed by the firm onto the consumers. Otherwise, the burden of the tax is shared between firms and consumers. Consequently the extent to which the price is shifted forward from the producer onto the consumers is dependent on the elasticities of supply and demand.

There are two reasons why tax incidence with imperfect competition is distinguished from the analysis for the competitive case. First, prices on imperfectly competitive markets are set at a level above marginal cost. Second, imperfectly competitive firms may also earn nonzero profits, so taxation can affect profit. To trace the effects of taxation, it is necessary to work through the profit-maximization process of the imperfectly competitive firms. Such an exercise involves characterizing the optimal choices of the firms and then seeing how they are affected by a change in the tax rate.

The incidence of a tax on output can be demonstrated by returning to the diagram for monopoly profit maximization. A tax of value \( t \) on output changes the tax-inclusive marginal cost from \( c \) to \( c + t \). In figure 9.5 this is shown to move the intersection between the marginal revenue curve and the marginal cost curve from \( a \) to \( b \). Output falls from \( y^o \) to \( y' \), and price rises from \( p \) to \( q \). In this case price rises by less than the
tax imposed—the difference between \( q \) and \( p \) is less than \( t \). This is called the case of tax \textit{undershifting}. What it means is that the monopolist is absorbing some of the tax and not passing it all on to the consumer.

With competition, the full value of the tax may be shifted to consumers but never more. With monopoly, the proportion of the tax that is shifted to consumers is determined by the shape of the AR curve (and hence the MR curve). In contrast to competition, for some shapes of AR curve it is possible for the imposition of a tax to be met by a price increase that exceeds the value of the tax. This is the case of tax \textit{overshifting} and is illustrated in figure 9.6. The imposition of the tax, \( t \), leads to a price increase from \( p \) to \( q \). As is clear in the figure, \( q - p > t \). This outcome could never happen in the competitive case.

The feature that distinguishes the cases of overshifting and undershifting is the shape of the demand function. Figure 9.5 has a demand function that is convex—it becomes increasingly steep as quantity increases. In contrast, figure 9.6 involves a concave demand function with a gradient that decreases as output increases. Either of these shapes for the demand function is entirely consistent with the existence of monopoly.

The overshifting of taxation is also a possibility with oligopoly. To illustrate this, consider the constant elasticity demand function \( X = p^\varepsilon \), where \( \varepsilon < 0 \) is the elasticity of demand. Since the elasticity is constant, so must be the markup at \( \mu^o = \frac{m}{m-(1/|\varepsilon|)} \). Furthermore, because \( \varepsilon < 0 \) it follows that \( \mu^o > 1 \). Applying the markup to marginal

![Figure 9.6](image)

**Figure 9.6**

Tax overshifting
cost plus tax obtains the equilibrium price of the oligopoly, \( q = \mu^o[c + t] \). The effect of an increase in the tax is then

\[
\frac{\partial q}{\partial t} = \mu^o > 1, \tag{9.13}
\]

so there is always overshifting with the constant elasticity demand function. This holds for any value of \( m \geq 1 \), and hence applies to both monopoly (\( m = 1 \)) and oligopoly (\( m \geq 2 \)). In addition, as \( m \) increases and the market becomes more competitive, \( \mu^o \) will tend to 1, as will \( \frac{\partial q}{\partial t} \), so the competitive outcome of complete tax shifting will arise.

Some estimates of the value of the tax-shifting term are given in table 9.3 for the beer and tobacco industries. Both of these industries have a small number of dominant firms and an oligopolistic market structure. The figures show that although undershifting arises in most cases, there is evidence of overshifting in the tobacco industry.

There is an even more surprising effect that can occur with oligopoly: an increase in taxation can lead to an increase in profit. The analysis of the constant elasticity case can be extended to demonstrate this result. Since the equilibrium price is \( q = \mu^o[c + t] \), we use the demand function to obtain the output of each firm as

\[
x = \frac{[\mu^o]^\varepsilon [c + t]^\varepsilon}{m}. \tag{9.14}
\]

Using these values for price and output results in a profit level for each firm of

\[
\pi = \frac{[\mu^o - 1][\mu^o]^\varepsilon [c + t]^{\varepsilon + 1}}{m}. \tag{9.15}
\]

The effect of an increase in the tax on the level of profit is then given by

\[
\frac{\partial \pi}{\partial t} = \frac{[\mu^o - 1][\mu^o]^\varepsilon [\varepsilon + 1][c + t]^\varepsilon}{m}. \tag{9.16}
\]

The possibility of the increase in tax raising profit follows by observing that if \( \varepsilon > -1 \), then \( [\varepsilon + 1] > 0 \), so \( \frac{\partial \pi}{\partial t} > 0 \). When the elasticity satisfies this restriction, an increase
in the tax will raise the level of profit. Put simply, the firms find the addition to their costs to be profitable.

It should be observed that such a profit increase cannot occur with monopoly because a monopolist must produce on the elastic part of the demand curve with $\varepsilon < -1$. With oligopoly the markup remains finite provided that $m - \frac{1}{m} > 0$ or $\varepsilon < -\frac{1}{m}$. Therefore profit can be increased by an increase in taxation if there is oligopoly.

The mechanism that makes this outcome possible is shown in figure 9.7, which displays the determination of the Cournot equilibrium for a duopoly. The figure is constructed by first plotting the isoprofit curves. The curves denote sets of output levels for the two firms that give a constant level of profit. The profit of firm 1 is highest on the curves closest to the horizontal axis, and it reaches its maximum at the output level, $m_1$, which is the output firm 1 would produce if it were a monopolist. Similarly the level of profit for firm 2 is higher on the isoprofit curves closest to the vertical axis, and is maximized at its monopoly output level, $m_2$. The assumption of Cournot oligopoly is that each firm takes the output of the other as given when it maximizes. So for any fixed output level for firm 2, firm 1 will maximize profit on the isoprofit curve that is horizontal at the output level of firm 2. Connecting the horizontal points gives the best-reaction function for firm 1, which is labeled $r_1(y_2)$. Similarly, setting a fixed output level for firm 1, we have that firm 2 maximizes profit on the isoprofit curve that is vertical at this level of firm 1’s output. Connecting the vertical points gives its best-reaction function $r_2(y_1)$.

Figure 9.7
Possibility of a profit increase
The Cournot equilibrium for the duopoly is where the best-reaction functions cross, and the isoprofit curves are locally horizontal for firm 1 and vertical for firm 2. This is point \( c \) in the figure. The Cournot equilibrium is not efficient for the firms, and a simultaneous reduction in output by both firms, which would be a move from \( c \) in the direction of \( b \), would raise both firms' profits. Further improvement in profit can be continued until the point that maximizes joint profit, \( \pi_1 + \pi_2 \), is reached. Joint profit maximization occurs at a point of tangency of the isoprofit curves, which is denoted by point \( b \) in figure 9.7. The firms could achieve this point if they were to collude, but such collusion would not be credible because both the firms would have an incentive to deviate from point \( b \) by increasing output.

It is this inefficiency that opens the possibility for a joint increase in profit to be obtained. Intuitively, how taxation raises profit is by shifting the isoprofit curves in such a way that the duopoly equilibrium moves closer to the point of joint profit maximization. Although total available production must fall as the tax increases, the firms secure a larger fraction of the gains from trade. Unlike collusion, the tax is binding on the firms and produces a credible reduction in output.

### 9.6 Specific and Ad valorem Taxation

The analysis of tax incidence has so far considered only specific taxation. With specific taxation, the legally responsible firm has to pay a fixed amount of tax for each unit of output. The amount that has to be paid is independent of the price of the commodity. Consequently the price the consumer pays is the producer price plus the specific tax. This is not the only way in which taxes can be levied. Commodities can alternatively be subject to ad valorem taxation so that the tax payment is defined as a fixed proportion of the producer price. Consequently, as price changes, so does the amount paid in tax.

The fact that tax incidence has been analyzed only for specific taxation is not a limitation when firms are competitive, since the two forms are entirely equivalent. The meaning of equivalence here is that a specific tax and an ad valorem tax that lead to the same consumer price will raise the same amount of tax revenue. Their economic incidence is therefore identical.

This equivalence can be shown as follows. Let \( t \) be the specific tax on a commodity. Then the equivalent ad valorem tax rate \( \tau \) must satisfy the equation

\[
q = p + t = [1 + \tau] p.
\]

(9.17)
Solving this equation, we have that \( \tau = \frac{t}{p} \) is the ad valorem tax rate that leads to the same consumer price as the specific tax. In terms of the incidence diagrams, both taxes would shift the supply curve for the good in exactly the same way. The demonstration of equivalence is completed by showing that the taxes raise identical levels of tax revenue. The revenue raised by the ad valorem tax is \( R = \tau p X \). Using the fact that \( \tau = \frac{t}{p} \), we can write this revenue level as \( \frac{t}{p} p X = t X \), which is the revenue raised by the specific tax. This completes the demonstration that the specific and ad valorem taxes are equivalent.

With imperfect competition this equivalence between the two forms of taxation breaks down: specific and ad valorem taxes that generate the same consumer price generate different levels of revenue. The reason for this breakdown of equivalence, and its consequences, are now explored.

The fact that specific and ad valorem taxes have different effects can be seen very easily in the monopoly case. Assume that the firm sells at price \( q \) and that each unit of output is produced at marginal production cost, \( c \). With a specific tax the consumer price and producer price are related by \( q = p + t \). This allows the profit level with a specific tax to be written as

\[
\pi = [q - t] x - cx = qx - [c + t] x. \tag{9.18}
\]

The expression for this profit level shows that the specific tax acts as an addition to the marginal cost for the firm. Now consider instead the payment of an ad valorem tax at rate \( \tau \). Since an ad valorem tax is levied as a proportion of the producer price, the consumer price and producer price are related by \( q = (1 + \tau) p \); hence the consumers pay price \( q \) and the firm receives \( p = \frac{1}{1 + \tau} q \). The profit level with the ad valorem tax is then

\[
\pi = \frac{1}{1 + \tau} qx - cx. \tag{9.19}
\]

The basic difference between the two taxes can be seen by comparing these alternative specifications of profit. From the perspective of the firm, the specific tax raises marginal production cost from \( c \) to \( c + t \). In contrast, the ad valorem tax reduces the revenue received by the firm from \( qx \) to \( \frac{1}{1 + \tau} qx \). Hence the specific tax works via the level of costs, whereas the ad valorem tax operates via the level of revenue. With competition this difference is of no consequence. But the very basis of imperfect competition is that the firms recognize the effect their actions has on revenue—so the ad valorem tax interacts with the expression of monopoly power.

The consequence of this difference is illustrated in figure 9.8. In the left-hand panel, the effect of a specific tax is shown. In the right-hand panel, the effect of an ad valorem
tax is shown. The specific tax leads to an upward shift in the tax-inclusive marginal cost curve. This moves the optimal price from $p$ to $q$. The ad valorem tax leads to a downward shift in average and marginal revenue net of tax as shown in figure 9.8. The ad valorem tax leads from price $p$ in the absence of taxation to $q$ with taxation. The resulting price increase is dependent on the slope of the marginal revenue curve.

What is needed to make a firm comparison between the effects of the two taxes is some common benchmark. The benchmark chosen is a given consumer price. The values of the specific and ad valorem taxes that lead to this consumer price are found. The taxes are then contrasted by determining which raises the most tax revenue. This comparison is easily conducted by returning to the definition of profit in (9.19). With the ad valorem tax, the profit level can be expressed as

$$\pi = \frac{1}{1+\tau}qx - cx = \frac{1}{1+\tau} [qx - [c + \tau c] x].$$

(9.20)

The second term of (9.20) shows that the ad valorem tax is equivalent to the combined use of a specific tax of value $\tau c$ plus a profit tax at rate $\frac{\tau}{1+\tau}$. A profit tax has no effect on the firm’s choice, but it does raise revenue. Hence an ad valorem tax with its rate is set so that

$$\tau c = t$$

(9.21)

Figure 9.8
Contrasting taxes
leads to the same after-tax price as the specific tax. However, the ad valorem tax raises more revenue. This is because the component $τc$ collects the same revenue as the specific tax $t$, but the ad valorem tax also collects revenue from the profit-tax component. Hence the ad valorem tax must collect more revenue for the same consumer price. This result can, alternatively, be expressed as the fact that for a given level of revenue, an ad valorem tax leads to lower consumer price than a specific tax.

In conclusion, ad valorem taxation is more effective than specific taxation when there is imperfect competition. The intuition behind this conclusion is that the ad valorem tax lowers marginal revenue, and this reduces the perceived market power of the firm. Consequently the ad valorem tax has the helpful effect of reducing monopoly power, offsetting some of the costs involved in raising revenue through commodity taxation.

9.7 Regulation of Monopoly

Up until this point the focus has been placed on the welfare loss caused by imperfect competition and on tax incidence. As we have shown, there are two competing views about the extent of the welfare loss, but even if the lower values are accepted, it is still beneficial to reduce the loss as far as possible. This raises the issue of the range of policies that are available to reduce the adverse effects of monopoly.

When faced with imperfect competition, the most natural policy response is to encourage an enhanced degree of competition. There are several ways in which this can be done. The most dramatic example is US antitrust legislation, which has been used to enforce the division of monopolies into separate competing firms. This policy was applied to the Standard Oil Company, which was declared a monopoly and broken up into competing units in 1911. More recently the Bell System telephone company was broken up in 1984. This policy of breaking up monopolists represents extreme legislation and, once enacted, leaves a major problem of how the system should be organized following the breakup. Typically the industry will require continuing regulation, a theme to which we return below.

Less dramatic than directly breaking up firms is to provides aids to competition. A barrier to entry is anything that allows a monopoly to sustain its position and prevent new firms from competing effectively. Barriers to entry can be legal restrictions such as the issue of a single license permitting only one firm to be active. They can also be technological in the sense of superior knowledge, the holding of patents, or the structure of the production function. Furthermore some barriers can be erected deliberately by the
incumbent monopolist specifically to deter entry. For a policy to encourage competition, it must remove or at least reduce the barriers to entry. The appropriate policy response depends on the nature of the barrier.

If a barrier to entry is created by a legal restriction, it can equally be removed by a change to the law. But here it is necessary to inquire as to why the restriction was created initially. One possible answer would take us to the concept of rent-creation, which is discussed in chapter 12. In that chapter the introduction of a restriction is seen as a way of generating rent. An interesting example of the creation of such restrictions are the activities of MITI (the Ministry of International Trade and Industries) in Japan. In 1961 MITI produced its “Concentration Plan,” which aimed to concentrate the mass-production automakers into two to three groups. The intention behind this was to sustain the international competition that ensued after the liberalization of auto imports into Japan and to place the Japanese car industry in a stronger position for exporting. These intentions were never fully realized, and the plan was ultimately undermined by developments in the auto industry, especially the emergence of Honda as a major manufacturer. Despite this, the example still stands as a good illustration of a deliberate policy attempt to restrict competition.

If barriers to entry relate to technological knowledge, then it is possible for the government to insist on the sharing of this knowledge. Both the concerns over the bundling of Internet Explorer with Windows in the United States and the bundling of Media Player with Windows in Europe are pertinent examples. In the United States the outcome has been that Microsoft is obliged to provide rival software firms with information that allows them to develop competing products, and to ensure that these products work with the Windows operating system. Microsoft’s rivals are pushing for a similar solution in the European Union. The existence of patents to protect the use of knowledge is also a barrier to entry. The reasoning behind patents is that they allow a reward for innovation: new discoveries are only valuable if the products in which they are embedded can be exploited without competitors immediately copying them. The production of generic drugs is one of the better-known examples of product copying. Without patents, the incentive to innovate would be much reduced and aggregate welfare would fall. The policy issue then becomes the choice of the length of a patent. It must be long enough to allow innovation to be adequately rewarded but not so long that it stifles competition. Current practice in the United States is that the term of a patent is twenty years from the date at which the application is filed.

Barriers to entry can also be erected as a deliberate part of a corporate strategy designed to deter competitors. Entry barriers can be within the law, such as sustained advertising campaigns to build brand loyalty or the building of excess capacity to deter
entry, or they can be illegal such as physical intimidation, violence, and destruction of property. Obviously the latter category can be controlled by recourse to the law if potential competitors wish to do so. Potentially limitations could be placed on advertising. The limitations on tobacco advertisements is an example of such a policy, but this has been motivated on health grounds and not competition reasons. The role of excess capacity is to provide a credible threat that the entry of a competitor will be met by an increase in output from the incumbent with a consequent reduction in market price. The reduction in price can make entry unprofitable, so sustaining the monopoly position. Although the economic reasoning is clear, it is difficult to see how litigation could ever demonstrate that excess capacity was being held as an entry deterrent, and this limits any potential policy response.

The enhancement of competition only works if it is possible for competitors to be viable. The limits of the argument that monopoly can be tackled by the encouragement of competition are confronted when the market is characterized by natural monopoly. The essence of natural monopoly is that there are increasing returns in production and that the level of demand is such that only a single firm can be profitable. This is illustrated in figure 9.9 where the production technology of the two firms involves a substantial fixed cost but a constant marginal cost. Consequently the average cost curve, denoted $AC$, is decreasing while the marginal cost curve, $MC$, is horizontal. When there is a monopoly, the single firm faces the demand curve $AR^1$. Corresponding to this average revenue curve is the marginal revenue curve $MR^1$. The profit-maximizing price for the

![Figure 9.9](image-url)

Natural monopoly
monopoly is $p$ and output is $y^1$. It should be observed that the price is above the level of average cost at output $y^1$, so the monopolist earns a profit.

Now consider the consequence of a second firm entering the market. The cost conditions do not change, so the $AC$ and $MC$ curves are unaffected. Demand conditions do change since the firms have to share the market. The simplest assumption to make is that the two firms share exactly half the market each. This would hold if the total market consists of two geographical areas each of which could be served by one firm. Furthermore this is the most beneficial situation for the firms since it keeps them from competing. Any other way of sharing the market would lead to them to earning less profit. With the market shared equally, the demand facing each firm becomes $AR^2$ (equal to the old $MR^1$) and marginal revenue, $MR^2$. The profit-maximizing price remains at $p$, but now at output $y^2$ this is below average cost. The two firms must therefore both take a loss. Since this market sharing is the most profitable way for the two firms to behave, any other market behavior must lead to an even greater loss.

What this argument shows is that a market in which one firm can be profitable cannot support two firms. The problem is that the level of demand does not generate enough revenue to cover the fixed costs of two firms operating. The examples that are usually cited of natural monopolies involve utilities such as water supply, electricity, gas, telephone, and railways where a large infrastructure has to be in place to support the market and is very costly to replicate. If these markets do conform to the situation in the figure, then without government intervention, only a single firm could survive in the market. Furthermore any policy to encourage competition will not succeed unless the government can fundamentally alter the structure of the industry. It is not enough just to try to get another firm to operate.

The two policy responses to natural monopoly most widely employed have been public ownership and private ownership with a regulatory body controlling behavior. When the firm is run under public ownership, its price should be chosen to maximize social welfare subject to the budget constraint placed on the firm—the resulting price is termed the Ramsey price. The budget constraint may require the firm to break even or to generate income above production cost. Alternatively, the firm may be allowed to run a deficit that is financed from other tax revenues. Assume that all other markets in the economy are competitive. The Ramsey price for a public firm subject to a break-even constraint will then be equal to marginal cost if this satisfies the constraint. If losses arise at marginal cost, then the Ramsey price will be equal to average cost. The literature on public sector pricing has extended this reasoning to situations where marginal cost and demand vary over time such as in the supply of electricity. Doing this leads into
the theory of peak-load pricing. When other markets are not competitive, the Ramsey price will reflect the distortions elsewhere in the economy.

Public ownership was practiced extensively in the United Kingdom and elsewhere in Europe. All the major utilities including gas, telephone, electricity, water, and trains were taken into public ownership. This policy was eventually undermined by the problems of the lack of incentive to innovate, invest, or limit costs. Together, these produced a very poor outcome with the lack of market forces producing industries that were overmanned and inefficient. As a consequence the United Kingdom has undertaken a privatization program that has returned all these industries to the private sector.

The treatment of the various industries since the return to private ownership illustrates different responses to the regulation of natural monopoly. The water industry is broken into regional suppliers that do not compete directly but are closely regulated. With telephones, the network is owned by British Telecom, but other firms are permitted access agreements to the network. This can allow them to offer a service without the need to undertake the capital investment. In the case of the railways, the ownership of the track, which is the fixed cost, has been separated from the rights to operate trains, which generates the marginal cost. Both the track owner and the train operators remain regulated. With gas and electricity, competing suppliers are permitted to supply using the single existing network.

The most significant difference between public ownership and private ownership with regulation is that under public ownership the government is as informed as the firm about demand and cost conditions. This allows the government to determine the behavior of the firm using the best available information. Policy can only maximize the objective function in an expected sense. So, although the available information may not be complete, the best that is possible will be achieved. As an alternative to public ownership, a firm may remain under private ownership but be made subject to the control of a regulatory body. This introduces possible asymmetries in information between the firm and the regulator. Faced with limited information, one approach considered in the theoretical literature is for the regulator to design an incentive mechanism that achieves a desirable outcome. An example of such a regulatory scheme is the two-part tariff in which the payment for a commodity involves a fixed fee to permit consumption followed by a price per unit of consumption, with these values being set by the regulator. Alternatively, the regulator may impose a constraint on some observable measure of the firm’s activities such as that it must not exceed a given rate of return on the capital employed. Even more simple are the regulatory schemes in the United Kingdom that involve restricting prices to rise at a slower rate than an index of the general price level.
The analysis has looked at a range of issues concerned with dealing with monopoly power and how to regulate industries. The essence of policy is to move the economy closer to the competitive outcome, but there can be distinct problems in achieving this. Monopoly can arise because of the combination of cost and demand conditions, and this can place limitations on what policies are feasible. Natural monopoly results in the need for regulation.

9.8 Regulation of Oligopoly

9.8.1 Detecting Collusion

In an oligopolistic market firms can collectively act as a monopolist and are consequently able to increase their prices. The problem for a regulatory agency is that such collusion is often tacit and so difficult to detect. However, from an economic viewpoint there is no real competition, and a high price is the prima facie evidence of collusion. The practical question for the regulator is whether a high price is the natural outcome of competition in a market where there is significant product differentiation (and so little pricing constraint from substitute products) or whether it reflects price collusion.

Nevo (2001) studied this question for the breakfast cereal industry where the four leaders Kellogg, Quaker, General Mills, and Post were accused by Congressman Chuck Schumer (March 1995) of charging “caviar prices for cornflakes quality.” After estimating price elasticities of demand for each brand of cereal, Nevo (2001) used these price elasticities to calculate the Lerner index for each brand, \( \frac{p - c}{p} \), that would prevail in the industry if producers were colluding and acting as a monopolist. Nevo then calculated the Lerner index for each brand if producers were really competing with each other.

Given the estimated demand elasticities, Nevo found that with collusion, the Lerner index of each brand would be on average around 65 to 75 percent. With the firms competing, the Lerner index would be on average around 40 to 44 percent. The next step was to compare these estimates of the Lerner index for the hypothetical collusive and competing industry with the actual Lerner index for the breakfast cereal industry to see which hypothesis is the most likely. According to Nevo, the actual Lerner index for the breakfast cereal market was about 45 percent in 1995. This market power index is far below the 65 to 75 percent hypothetical Lerner index that would prevail in a colluding industry and much closer to the Lerner index in the competing hypothesis. Nevo concludes that market power is significant in this industry, not because of collusion.
but because of product differentiation that limits competition from substitute products (after all, what is the substitute for a “healthy” cereal breakfast?).

### 9.8.2 Merger Policy

In its recent reform of merger regulation, the European Commission has recognized that in oligopolistic markets a merger may harm competition and consequently increase prices. Under the original European Commission Merger Regulation (ECMR) a merger was incompatible with the common market if and only if it “creates or strengthens a dominant position as a result of which competition would be significantly impeded.” The problem with this two-part cumulative test was that unless a merger was likely to create or strengthen a dominant position, the question of whether it could lessen competition did not arise and so could not be used to challenge a merger. However, one can easily think of oligopoly situations where a merger would substantially lessen competition without giving any individual firm a dominant position. Moreover the concept of dominance is not easily established especially in the presence of tacit collusion. In practice, the concept of dominance had different meanings depending on the circumstances. In particular, when there was some presumption of collusion, the European Commission could use the concept of “collective” dominance, taking as a single unit a group of sellers suspected to collude in their pricing policy. Just as Alice said in *Through the Looking Glass*, the question comes to “whether you can make words mean so many different things.”

In the 2004 reform of merger policy the European Commission shifted the attention to the second part of the original regulation. The key article in the new ECMR says that “a concentration which would significantly impede effective competition, in the common market or in a substantial part of it, in particular as a result of the creation or strengthening of a dominant position, shall be declared incompatible with the common market” (Article 2). Thus the European Commission has recognized that reducing competition is not necessarily dominance but rather a result of how much competition is left. The fundamental idea is that in oligopolistic markets a merger of two or more rivals raises competitive concerns if the merging firms sell products that are close substitutes. By removing the competitive constraint, merging firms are able to increase their prices. This is the “unilateral effect” theory of competitive harm that has been commonly used in the US merger regulation.

Economists have developed a large number of simulation methods, mostly based on estimated demand elasticities, to determine the possible change in price resulting from a merger. Simulation models combine market data on market shares, the own-price
elasticity of demand, and the cross-price elasticities of demand with a model of firm behavior and anticipated reductions in cost from the merger to predict the likely price effects. A practical example will be useful to illustrate the method. The example is drawn from Hausman and Leonard (1997) and concerns the market for bath tissue. In 1995 the producer of the Kleenex brand acquired the producer of two competing brands (Cottonelle and ScotTissue). The market shares for these products and other brands are shown in table 9.4.

Using weekly retail scanner data that tracks household purchases in retail stores in major US cities, it was possible to estimate own-price elasticities as shown in table 9.4. The key cross-price elasticities were estimated to be 0.19 (Kleenex relative to Cottonelle), 0.18 (Kleenex relative to ScotTissue), 0.14 (Cottonelle relative to Kleenex), and 0.06 (ScotTissue relative to Kleenex). In addition it was anticipated that the acquisition would reduce the marginal cost of production for ScotTissue, Cottonelle, and Kleenex by 4, 2.4, and 2.4 percent respectively. With these estimates of demand elasticities, information about market shares, and the anticipated cost saving from the acquisition of Cottonelle and ScotTissue by the Kleenex brand, it was possible to evaluate the price effects of the merger. A simulation model based on these market estimates and other assumptions about firm and market behavior (Nash equilibrium and constant marginal costs) produced the following prices changes. The acquisition would lead to a reduction in the price of ScotTissue and Cottonelle by 2.6 and 0.3 percent respectively, and an increase in the price of Kleenex by 1.0 percent. Not surprisingly, the Antitrust did not challenge the merger.

Table 9.4
Estimating the effect of merger in the bath tissue market

<table>
<thead>
<tr>
<th>Bath tissue brand</th>
<th>Market share</th>
<th>Own-price elasticity</th>
<th>Price change [cost change]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kleenex</td>
<td>7.5%</td>
<td>−3.38</td>
<td>+1.0% [−2.4%]</td>
</tr>
<tr>
<td>Cottonelle</td>
<td>6.7%</td>
<td>−4.52</td>
<td>−0.3 [−2.4]</td>
</tr>
<tr>
<td>ScotTissue</td>
<td>16.7%</td>
<td>−2.94</td>
<td>−2.6 [−4.0]</td>
</tr>
<tr>
<td>Charmin</td>
<td>30.9%</td>
<td>−2.75</td>
<td></td>
</tr>
<tr>
<td>Northern</td>
<td>12.4%</td>
<td>−4.21</td>
<td></td>
</tr>
<tr>
<td>Angel Soft</td>
<td>8.8%</td>
<td>−4.08</td>
<td></td>
</tr>
<tr>
<td>Private label</td>
<td>7.6%</td>
<td>−2.02</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>9.4%</td>
<td>−1.98</td>
<td></td>
</tr>
<tr>
<td>Market demand</td>
<td></td>
<td>−1.17</td>
<td></td>
</tr>
</tbody>
</table>

Source: Data from tables 1 and 2 in Hausman and Leonard (1997).
Chapter 9: Imperfect Competition

9.9 Unions and Taxation

As well as monopoly on product markets, it is possible to have unions creating market power for their members on input markets. By organizing labor into a single collective organization, unions are able to raise the wage above the competitive level and generate a surplus for their members. The issue of tax incidence is also of interest when there are unions, since they can employ their market power to reduce the effect of a tax on the welfare of members.

The role of trade unions is to ensure that they secure the best deal possible for their members. In achieving this, the union faces a trade-off between the wage rate and the level of employment, since a higher wage will invariably lead to lower employment. This trade-off has to be resolved by the union’s preferences.

A standard way of representing the preferences of a union is to assume that it has a fixed number, \( m \), of members. Each employed member receives a wage \( w(1 - t) \), where \( t \) is the tax on wage income. The unemployed members receive a payment of \( b \), which can represent either unemployment benefit or the payment in a nonunionized occupation. The level of employment is determined by a labor demand function \( n(w) \), with higher values of \( w \) leading to lower levels of employment. If the wage rate is \( w \), the probability of any particular member being employed and receiving \( w(1 - t) \) is \( \frac{n(w)}{m} \). Consequently, if all members are assumed to have the same preferences, the expected utility of a typical union member is

\[
U = \frac{n(w)}{m} u(w(1 - t)) + \frac{m - n(w)}{m} u(b). \tag{9.22}
\]

Since all union members have identical preferences, this utility function can also be taken to represent the preferences of the union.

The union chooses the wage rate to maximize utility, so that the chosen wage satisfies the first-order condition

\[
n'(w)[u(w(1 - t)) - u(b)] + n(w)[1 - t] u'(w(1 - t)) = 0. \tag{9.23}
\]

The interpretation of this condition is that the optimal wage rate balances the marginal utility of a higher wage against the value of the marginal loss of employment. Now define the elasticity of labor demand by \( \varepsilon_n = \frac{\Delta n}{\Delta w \frac{n}{w}} < 0 \) and the elasticity of utility by

\[
\varepsilon_u = \frac{\Delta u}{\Delta w(1 - t) \frac{u}{u}} > 0.
\]

The first-order condition (9.23) can then be written as

\[
u(w(1 - t)) = \mu u(b). \tag{9.24}
\]
where \( \mu^u = \frac{1}{1 - (\varepsilon_u / |\varepsilon_n|)} > 1 \) is the union markup relating the utility of an employed member to that of an unemployed member. This markup is a measure of the unions market power. Given a value for the utility elasticity, \( \varepsilon_u \), the markup increases the lower is the elasticity of labor demand, \( \varepsilon_n \). As labor demand becomes perfectly elastic, as it does if the labor market is perfectly competitive, then \( \mu^u \) tends to 1, and the union can achieve no advantage for its members.

The incidence of taxation can now be determined. To simplify, assume that the two elasticities—and hence the markup—are constant. Then the utility of the after-tax wage must always bear the same relation to the utility of unemployment benefit. Consequently \( u[1 - t] \) must be constant whatever the tax rate. This can only be achieved if the union negates any tax increase by securing an increase in the wage rate that exactly offsets the tax change. Consequently those who retain employment are left unaffected by the tax change, but since the wage has risen, employment must fall. Overall, the union members must be worse off. This argument can easily be extended to see that if the elasticities are not constant, there is the potential for the overshifting, or undershifting, of any tax increase. In this respect tax incidence with trade unions has very similar features to incidence with monopoly.

### 9.10 Monopsony

A monopsony market is a market consisting of a single buyer who can purchase from many sellers. The single buyer (or monopsonist) could be a firm that constitutes the only potential buyer of an input. It could also be an individual or public organization that is the only buyer of a product. For example, in many countries the government is the monopsonist in the teaching and nursing markets. In local markets with only one large employer, the local employer might literally be the only employment option in the local community (a coal mine, supermarket, government agency, etc.), so it might make sense that the local employer acts as a monopsonist in reducing the wage below the competitive level. In larger markets with more than one employer, employers association often have opportunities to coordinate their wage offers. This wage coordination allows employers to act as a “demand” cartel in the labor market and thus replicate the monopsony outcome. Just as monopoly results in supply reduction with a price or wage above competitive levels, monopsony will result in demand reduction with price or wage below competitive levels.

In a perfectly competitive market in which many firms purchase labor services, each firm takes the price of labor as given. Each firm maximizes its profits by choosing the
employment level that equates the marginal revenue product of labor with the wage rate. In contrast, in a monopsony labor market, the monopsony firm pays a wage below the competitive wage. The result is a shortage of employment relative to the competitive level. The idea is that since the marginal revenue product from additional employment exceeds the wage cost in a monopsony labor market the monopsonist employer might want to hire more people at the prevailing wage. However, it would not want to increase the wage to attract more workers because the gain from hiring additional workers (the marginal revenue product) is outweighed by the higher wage bill it would face for its existing workforce.

Figure 9.10 shows the equilibrium in a monopsony labor market. The competitive equilibrium occurs at a market-clearing wage, \( w^c \), where the labor supply curve intersects the demand curve. Suppose now there is a single buyer on this labor market. The marginal revenue of labor is the additional revenue that the firm gets when it employs an additional unit of labor. Suppose that the firm’s output as a function of its labor use is \( Q(L) \) and that the firm is a price taker on the output market, so its output price \( p \) is independent of the amount of output \( Q \). Then the marginal revenue of labor is \( MR_L = p \frac{dQ}{dL} \), which is decreasing due to decreasing returns to labor. This marginal revenue is depicted in figure 9.10 as the downward-sloping labor demand curve. The supply of labor is described by the “inverse” supply curve. The inverse supply curve

![Figure 9.10](image)

Monopsony in the labor market
\( w(L) \) describes the wage required to induce any given quantity of labor to be supplied. Since the supply curve is upward sloping, \( \frac{dw}{dL} > 0 \). The total labor cost of the monopsonist is \( Lw(L) \), and the marginal cost of labor is the extra cost that comes from hiring one more worker \( MC_L = w + L \frac{dw}{dL} \). This additional cost can be decomposed into two parts: the cost from employing more workers at the existing wage, \( w \), and the cost from raising the wage for all workers, \( L \frac{dw}{dL} \). Since \( \frac{dw}{dL} > 0 \), the marginal labor cost curve lies everywhere above the labor supply curve, as indicated in figure 9.10. The monopsonist will maximize profit, \( \pi = pQ(L) - w(L)L \), at the point where the marginal revenue of labor is equal to marginal cost, \( p \frac{dQ}{dL} = w + L \frac{dw}{dL} \).

The choice that gives maximum profit occurs in figure 9.10 at the intersection between the marginal cost curve and the labor demand curve, yielding employment level \( L^m \) and wage rate \( w^m \). Therefore in a monopsony labor market the monopsony firm pays a wage that is less than the competitive wage with employment level below the competitive level. The monopsony equilibrium condition can also be expressed as an inverse elasticity pricing rule. Indeed the elasticity of labor supply is \( \varepsilon_L = \frac{dL}{dw} \frac{w}{L} \) and the profit maximization condition \( MR_L = MC_L \) can be re-arranged to give

\[
\frac{MR_L - w}{w} = \frac{1}{\varepsilon_L}. \tag{9.25}
\]

This inverse pricing rule says that the percentage deviation from the competitive wage is inversely proportional to the elasticity of labor supply. In contrast to monopoly, the key elasticity is the supply elasticity. Just as monopoly results in a deadweight loss, so does monopsony leading to underemployment and underpricing of the input (in this case labor) relative to the competitive outcome.

### 9.11 Conclusions

This chapter has shown how imperfect competition leads to a failure to attain Pareto-efficiency. As with all such failures, this opens a potential role for government intervention to promote efficiency. Estimates of the welfare loss due to imperfect competition vary widely from the almost insignificant to considerable proportions of welfare, depending on the perspective taken upon expenditures on securing the monopoly position. These static losses have to be set against the possible dynamic gains.

Economic tax incidence relates to whoever ultimately has to change his behavior as a consequence of taxation. With competition the outcome is fairly straightforward: the...
cost of a commodity tax is divided between producers and consumers, with the division depending on the elasticities of supply and demand. Imperfect competition introduces two additional factors. Taxes may be overshifted so that price rises by more than the value of the tax. In addition an increase in taxation may even raise the profits of firms. In contrast to the competitive case, specific and ad valorem taxation are not equivalent with imperfect competition. In a choice between the instruments, ad valorem taxation is more effective, since it has the effect of reducing perceived monopoly power.

To reduce the welfare loss, policy should attempt to encourage competition. In some circumstances this can work, but when there is natural monopoly, this policy has to be carefully considered. A natural monopoly could be taken into public ownership or run as a private firm with regulation. Recent policy has concentrated on the latter.

Further Reading

The measurement of welfare loss began with:
The other values in table 8.1 are taken from:
The basics of oligopoly theory are covered in:
The analysis of tax incidence with oligopoly can be traced back to:
The results in table 8.3 are compiled from:
Part III: Departures from Efficiency


Results on comparison of specific and ad valorem tax are in:


The example on detecting collusion is drawn from:


The merger simulation model for bath tissue is drawn from:


A further discussion of merger simulation analysis can be found in:


A presentation of the various concepts of competition is in:


A good perspective on the inefficiency resulting from market power with special attention on information problems is:


The basic and first paper on product differentiation is:


The other classic paper is:


An economic analysis of regulation policies with special attention to the United Kingdom is:


A good account of antitrust law and economics is in:

Exercises

9.1 What should be the objective of a monopoly labor union?

9.2 An industry is known to face a market price elasticity of demand $\varepsilon = -3$. Suppose that this elasticity is approximately constant as the industry moves along its demand curve. The marginal cost in this industry is $10$ per unit, and there are five firms in the industry. What would the Lerner index be at the Cournot equilibrium in this industry?

9.3 Consider a monopolist operating the underground in Europa city with a total cost curve given by $c(x) = 15 + 5x$. The monopolist sets two prices: a high price $p_h$ and a low price $p_l$. Everyone is eligible for the high price, but only by taking the tube outside the peak hours is anyone eligible for the discount price. Suppose that the only off-peak travelers are those who are not willing to buy the ticket at $p_h$.

a. If the monopolist faces the inverse demand curve given by $p(x) = 20 - 5x$, what are the profit-maximizing values of $p_h$ and $p_l$? [Hint: Let $x_h$ and $x_l$ denote the high-price and low-price quantities respectively. Then profit for the price discriminating monopolist is $\pi = p(x_h)x_h + p(x_h + x_l)x_l - c(x_h + x_l)$.

b. How much economic profit does the monopolist take?

c. How much profit would be made if the same price were charged to all buyers (no price discrimination)? Discuss the difference from part b.

9.4 Demonstrate that monopoly is Pareto-inefficient. Must it always lead to a lower level of social welfare than competition?

9.5 Consider an economy with one good and a linear inverse demand $p(x) = a - bx$. Suppose that there is a single firm operating in this market and that this firm faces a linear cost function $C(x) = cx$ (with $c < a$).

a. Show that the profit maximizing output with monopoly is $x_m = \frac{a-c}{2b}$ and the resulting price is $p_m = \frac{a-c}{2}$.

b. Show that the efficient competitive output level is $x^c = \frac{a-c}{2b} = 2x_m$.

c. Calculate the monopoly profit and the monopoly deadweight loss, and show that these are respectively $\pi_m = \frac{1}{b} \left( \frac{a-c}{2} \right)^2$ and $\lambda_m = \frac{x_m^2}{2}$.

d. Consider a quantity subsidy $s$ to the monopolist so that its cost function is $C(x) = [c-s]x$. Show that a subsidy rate of $s = a - c$ induces the monopolist to produce the efficient amount of output.

e. What is the monopolist’s profit resulting from a government intervention imposing marginal cost pricing?

9.6 The inverse demand function for a product is given by $p = a - bX$. The cost function for each firm producing the product is $C(x) = F + cx$.

a. Assume that the industry is monopolized. For what value of $F$ can the monopoly be profitable?

b. Assume that the industry is a Cournot duopoly. For what values of $F$ can both duopolists be profitable?
c. Use the solutions to parts a and b to determine a range of $F$ for which the industry is a natural monopoly.

d. Assume that the two duopolists decide to act collusively. Determine the values of $F$ for which both can be profitable. Does this give higher or lower social welfare than monopoly?

e. The government can either provide a production subsidy to the monopolist (an amount $s$ per unit of output) or subsidize the fixed costs of the Cournot duopolists. Which policy is most cost effective?

9.7 Assume that a monopolist can identify two distinct markets. Find the profit-maximizing prices if the demand functions for the two markets are

$$x_1 = 100 - 2p_1, \quad x_2 = 150 - 3p_2.$$ 

What is the level of consumer surplus in each market? If the monopolist is forced by legislation to charge a single price, what will this price be? Contrast the level of consumer surplus with and without price discrimination.

9.8 Consider two monopolists operating in separate markets with identical and constant marginal cost. Are the following statements true or false?

a. If both face different linear demand curves that are parallel, the monopolist that will have the higher markup is the one whose demand curve is farther from the origin.

b. If both face linear demand curves with identical vertical intercepts but different slopes, the monopolist with the higher markup is the one with the steeper demand curve.

c. If both face linear demand curves with identical horizontal intercepts but different slopes, the monopolist with the higher markup is the one with the steeper demand curve.

9.9 Discuss how brand promotion can increase inefficiency. Is brand proliferation good or bad?

9.10 Demand is assumed to be unit-elastic: $X(p) = \frac{1}{p}$. There are $m \geq 2$ firms operating in the market with constant marginal cost levels $c_1 \leq c_2 \leq \ldots \leq c_m$. They engage in Cournot competition.

a. Show that the equilibrium price implies Lerner indexes $\frac{P-C}{P} = s_i$, where $s_i$ is the market share of firm $i$.

b. Using the equilibrium price, show that the profit of firm $i$ is equal to $[s_i]^2$.

c. Show that the industry profit is equal to the Herfindahl index $H = \sum_i [s_i]^2$.

d. What is the effect of a specific tax $t$ on equilibrium price? How does this tax affect the industry profit and the Herfindahl index?

9.11 Consider a standard Cournot oligopoly with $n = 2k$ identical firms (with $k \geq 1$), an inverse demand $P(X)$, and a cost function $C(x)$ with no fixed costs. Consider only two possible cases: $C(x)$ convex and $C(x)$ concave. Assume that there is always a unique symmetric equilibrium with per firm output $x_k$ and profit $\pi_k$. Assume that there are $k$ two-firm mergers.

a. List all conditions on the primitives of the model such that each firm is better off after these mergers. Explain your answer (no proof needed).

b. Can such a set of mergers be expected to take place without regulatory intervention? Explain.

c. Under what conditions can such a set of mergers increase social welfare?
9.12 Consider a standard Cournot oligopoly with \( n \geq 2 \) identical firms, \( P(x) = a - bX, \ X \geq 0, \) and \( C(x) = cx^2. \)

a. Find the Cournot equilibrium output and profit.

b. If \( m \) firms wish to merge, what would be their cost function, assuming that they can use all their \( m \) production plants but that they otherwise do not have any efficiency gains as a result of the merger?

c. Given the cost function from part b, when is an \( m \)-firm merger profitable to the merged entity? To the nonmerging firms?

d. Give a precise economic intuition explaining your answer relative to the usual (linear cost) case.

9.13 Consider two firms, \( i = 1, 2, \) producing differentiated products and engaged in Cournot competition. The inverse demand for firm \( i \) is given by \( p_i = a - b q_i - dq_j, \) where \( q_i \) is the amount of its own output and \( q_j \) is firm \( j \)'s level of output (with \( a > c, \ b > \frac{1}{2} \) and \( -1 < d < 1). \) Similarly the inverse demand for firm \( j \) is given by \( p_j = a - b q_j - dq_i. \) The goods are substitutes for \( d > 0 \) and complements for \( d < 0. \) The marginal cost of each firm is zero.

a. Given the market demands, what are the best-response functions of the two firms?

b. Draw the best-response functions both for complements (\( d < 0 \)) and substitutes (\( d > 0 \)).

c. Compute the Cournot equilibrium quantities and prices in this market.

d. Compare the outcome between substitutes and complements goods.

e. What are the profit-maximizing quantities and prices if firm \( i \) is a monopolist in this market? Compare with part c.

9.14 Consider a standard Cournot oligopoly with \( n \geq 2 \) identical firms, an inverse demand function \( n \geq 2, \) and cost function \( C(x) = K + cx \) if \( x > 0, \) and 0 if \( x = 0, \) meaning \( K \) is a fixed cost.

a. Find the Cournot equilibrium output and profit. How many firms (as a function of \( K \)) can survive at the equilibrium?

b. When is an \( m \)-firm merger profitable to the merged entity? To the nonmerging firms?

c. Give a precise economic intuition as to why most mergers are not profitable in the usual model with \( K = 0. \) How is it different when \( K > 0? \)

9.15 Consider a homogeneous-good Cournot oligopoly with \( n \geq 2 \) identical firms with cost \( C(x) = 0 \) and inverse demand \( P(X) = e^{-X}. \)

a. Find a firm’s best-response function, the Cournot equilibrium output, price, and profit. What type of equilibrium is this?

b. Find all the merger sizes \( m (2 \leq m \leq n) \) that are profitable to the merged entity. Are these mergers also profitable to the nonmerging firms?

c. Give an economic intuition, and compare it to the case of linear demand.

9.16 Consider Cournot competition with \( n \) identical firms. Suppose that the inverse demand function is linear with \( P(X) = a - b X, \) where \( X \) is total industry output, \( a, b > 0. \) Each firm has a linear cost function of the form \( C(x) = cx, \) where \( x \) stands for per firm output. It is assumed that \( a > c. \)
a. At the symmetric equilibrium, what are the industry output and price levels? What are the equilibrium per firm output and profit levels? What is the equilibrium social welfare (defined as the difference between the area under the demand function and total cost)?

b. Now let m out of n firms merge. Show that the merger is profitable for the m merged firms if and only if it involves a pre-merger market share of 80 percent.

c. Show that each of the \((n - m)\) nonmerged firms is better off after the merger.

d. Show that the \(m\)-firm merger increases industry price and also lowers consumer welfare.

9.17 What is the difference between vertical and horizontal product differentiation? Provide an example of each.

9.18 A monopolist faces the inverse demand function \(P(x) = a - bx\) and produces with constant marginal cost \(c\).

a. Determine the effect on equilibrium price of the introduction of a specific tax of value \(t\). Is the tax overshifted?

b. Calculate the effect on profit of the tax. Show that \(d\pi = -x\), where \(x\) is the equilibrium output level. Explain this result.

c. Now replace the specific tax with an ad valorem tax at rate \(\tau\). Find a pair of taxes that lead to the same level of tax revenue. Which gives a lower price?

9.19 A Cournot oligopoly of \(n\) firms faces an inverse demand function \(p = X^{-1/v}\), where \(X\) is aggregate industry output \(X = \sum_{i=1}^{n} x_i\). The cost function for each firm is \(C(x_i) = cx_i\) and a specific tax \(t\) is charged on each unit of output.

a. Show that the output of each firm at the symmetric Cournot equilibrium is \(x = \frac{1}{n} \left[ \frac{a - x}{c + t} \right]^{v} \).

b. Show that the rate of tax shifting, \(\frac{dp}{dt}\), is constant at \(\frac{dp}{dt} = \frac{v}{vn - 1}\).

c. Comment on the effect of increasing the number of firms and increasing the elasticity of demand.

9.20 For the same market description used in exercise 9.19:

a. Can a tax increase raise profit if \(n = 1\)?

b. What conditions are required for a tax increase to raise profit?

9.21 (Mixed oligopoly) Consider a market with one public firm, denoted 0, and one private firm, denoted 1. Both firms produce a homogeneous good with identical and constant marginal \(c\) per unit of output, and face the same linear demand function \(P(X) = a - bX\) with \(X = x_0 + x_1\). It is assumed \(a > c\). The private firm maximizes profit \(\pi_1 = P(X)x_1 - cx_1\), and the public firm maximizes a combination of welfare and profit \(V_0 = \theta W + [1 - \theta]\pi_0\) with welfare given by consumer surplus less cost, \(W = \int_0^X P(y) dy - c(x_0 + x_1)\). Both firms choose output as the strategic variable.

a. Calculate the best-response functions of the public and the private firms. Use a graph of the best-response functions to illustrate what would happen if \(\theta\) changed from 0 to 1.

b. Calculate the equilibrium quantities for the private and public firms. Derive the aggregate output in equilibrium as a function of \(\theta\).
c. Calculate the socially optimal output level (by using the marginal cost pricing rule), and compare with the equilibrium outcome.
d. Show that an increase in $\theta$ must increase the equilibrium industry output, and so equilibrium price must fall and welfare increase. Verify that the equilibrium outcome converges to the socially optimal outcome when $\theta = 1$.
e. Consider $\theta < 1$ and calculate the quantity subsidy $s$ (with marginal cost after subsidy $c - s$) such that the firms will produce the socially optimal output level. What impact does a change in $\theta$ have on the optimal subsidy? Why?

9.22 Define natural monopoly. Draw the demand, marginal revenue, marginal cost, and average cost curves for a natural monopoly.
a. What does the size of a market have to do with whether an industry is a natural monopoly?
b. What are the two problems that arise when the government regulates a natural monopoly by limiting price to be equal to marginal cost?
c. Suppose that a natural monopoly is required to charge average total cost. On your diagram, label the price charged and the deadweight loss to society relative to marginal-cost pricing.

9.23 What gives the government the power to regulate mergers between firms? From the viewpoint of the welfare to society, give a good reason and a bad reason why two firms might want to merge.

9.24 Assume that a monopolist’s marginal cost is positive at all output levels. Are the following true or false?
a. When the monopolist operates on the inelastic part of the demand curve, it can increase profit by producing less.
b. When the monopolist operates on the inelastic part of the demand curve, it can increase profit by producing more.
c. The monopolist’s marginal revenue can be negative for some levels of output.

9.25 (Varian) A daily dose of the AIDS drug PLC sells for $18 in the United States and $9 in Uganda (New York Times, September 21, 2000). Even at $9 a dose the drug company makes a profit on additional sales. But if the drug were sold at $9 to everyone, profits would decline. Price discrimination is not popular with consumers, especially those paying the higher price. To evaluate whether differential pricing is good or bad, the critical question from the viewpoint of economics is whether uniform price or differential pricing leads to more people getting the drug. In general, there is no easy answer. Imagine that there are only two countries involved, the United States and Uganda:
a. Imagine the US market for the PLC drug is more than five times the Ugandan market, and the drug sells respectively for $18 and $9. What price is likely to prevail if only one price can be charged? What would be the effect on total consumption and, especially, for drug consumers in Uganda? What would be the effect on US drug consumers?
b. Imagine an anti-malarial drug that many people in Uganda would buy at $2 a dose and few people in the United States would buy at $10. If the Ugandan market is more than ten times the US market, what price is likely to prevail if drug company can set only one price? What would be the effect on total consumption and for drug consumers in United States and Uganda?
c. Based on this example, discuss when price discrimination is likely to be socially useful and when it does not have much to recommend it.

9.26 A company is considering building a bridge across a river. The bridge would cost $3 million to build and nothing to maintain. The anticipated demand over the lifetime of the bridge is \( x = 800 - 100p \), where \( x \) is the number of crossings (in thousands) given the price per crossing \( p \).

a. If the company builds the bridge, what will be the profit-maximizing price?
b. Will that price lead to the efficient number of crossings? Why or why not?
c. What will be the company’s profit or loss? Should it build the bridge?
d. If the government were to build the bridge, what price should it charge?
e. Should the government build the bridge? Why or why not?

9.27 The jazz singer Nora Jones has monopoly power over a scarce resource: herself on stage. She is the only person who can perform a Nora Jones concert. Does this fact imply that the government should regulate ticket prices for her concerts? Explain.
10 Asymmetric Information

10.1 Introduction

A key feature of the real world is asymmetric information. Most people want to find the right partner, one who is caring, kind, healthy, intelligent, attractive, trustworthy, and so on. While attractiveness may be easily verified at a glance, many other traits people seek in a partner are difficult to observe, and people usually rely on behavioral signals that convey partial information. There may be good reasons to avoid a potential mate who is too eager to start a relationship with you, as this may suggest unfavorable traits. Similarly it is hard not to infer that people who participate in dating services must be on average less worth meeting, and the consensus appears to be that these services are a bad investment. The reason is that the decision to resort to a dating agency identifies people who have trouble initiating their own relationships, which is indicative of other unwelcome traits. The lack of information causes caution in dating, which can result in good matches being missed.

Asymmetric information arises in economics when the two sides of the market have different information about the goods and services being traded. In particular, sellers typically know more about what they are selling than buyers do. This can lead to adverse selection where bad-quality goods drive out good-quality goods, at least if other actions are not taken. Adverse selection is the process by which buyers or sellers with “unfavorable” traits are more likely to participate in the exchange. Adverse selection is important in economics because it often eliminates exchange possibilities that would be beneficial to both consumers and sellers alike. There might seem some easy way to resolve the problem of information asymmetry: let everyone reveal what they know. Unfortunately, individuals do not necessarily have the incentive to tell the truth (think about the mating example or the market identification of high- and low-ability people).

Information imperfections are pervasive in the economy, and in some sense, it is an essential feature of a market economy that different people know different things. While such information asymmetries inevitably arise, the extent to which they do so and their consequences depends on how the market is organized. The anticipation that they will arise also affects market behavior. In this chapter we discuss the ways in which information asymmetries affect market functioning and how they can be partially overcome through policy intervention. We do not consider how the agents can create information problems, for example, in an attempt to exploit market power by
differentiating products or by taking actions to increase information asymmetries as in the general governance problem.

One fundamental lesson of information imperfection is that actions convey information. This is a commonplace observation in life, but it took some time for economists to fully appreciate its profound effects on how markets function. Many examples can be given. A willingness to purchase insurance at a given price conveys information to an insurance company, because those most likely to decide that the insurance is not worthwhile are those who are least likely to have an accident. The quality of a guarantee offered by a firm conveys information about the quality of its products as only firms with reliable products are willing to offer a good guarantee. The number of years of schooling may also convey information about the ability of an individual. More able people may go to school longer and the higher wage associated with more schooling may simply reflect the sorting that occurs rather than the ability-augmenting effect of schooling itself. The willingness of an investor to self-finance a large fraction of the cost of a project conveys information about his belief in the project. The size of deductibles and co-payments that an individual chooses in an insurance contract may convey information that he is less risk prone. The process by which individuals reveal information about themselves through the choices that they make is called self-selection.

Upon recognizing that actions convey information, two important results follow. First, when making decisions, agents will not only think about what they prefer, but they will also think about how their choice will affect others’ beliefs about them. So I may choose longer schooling not because I value what is being taught, but because it changes others’ beliefs concerning my ability. Second, it may be possible to design a set of choices that would induce those with different characteristics to effectively reveal their characteristics through their choices. As long as some actions are more costly for some types than others, it is an easy matter to construct choices that separate individuals into classes: self-selection mechanisms could, and would, be employed to screen. For example, insurance companies may offer a menu of transaction terms that will separate out different classes of risk into preferring different parts of the menu.

In equilibrium both sides of the market are aware of the informational consequences of their actions. In the case where the insurance company or employer takes the initiative, self-selection is the main screening device. In the case where the insured, or the employee, takes the initiative to identify himself as a better type, it is usually considered as a signaling device. So the difference between screening and signaling lies in whether the informed or uninformed side of the market moves first.

Whatever the actions taken, the theory predicts that the types of transactions that will arise in practice are different from those that would emerge in a perfect-information
context. The fact that actions convey information affects equilibrium outcomes in a profound way. On the one hand, since quality increases with price in adverse selection models, it may be profitable to pay a price in excess of the market-clearing price. In credit markets, the supply of loans may be rationed. In the labor market, the wage rate may be higher than the market-clearing wage, leading to unemployment. There may exist multiple equilibria. Two forms of equilibria are possible: pooling equilibria, in which the market cannot distinguish among the types, and separating equilibria, in which the different types separate out by taking different actions. On the other hand, under plausible conditions, equilibrium might not exist (in particular, if the cost of separation is too great).

Another set of issues arise when actions are not easily observable. An employer would like to know how hard his employee is working; a lender would like to know the actions the borrower will undertake that might affect the chance of reimbursement. These asymmetries of information about actions are as important as the situations of hidden knowledge. They lead to what is referred to as the moral hazard problem. This term originates from the insurance industry, which recognized early that more insurance reduces the precautions taken by the insured (and not taking appropriate precautions was viewed to be immoral, hence the name). One way to solve this problem is to try to induce desired behavior through the setting of contract terms. A borrower’s risk-taking behavior may be controlled by the interest rate charged by the lender. The insured will exert more care when facing contracts with large deductibles. But, in competing for risk-averse customers, the insurance companies face an interesting trade-off. The insurance has to be complete enough so that the individual will purchase. At the same time deductibles have to be significant enough to provide adequate incentives for insured parties to take care.

This chapter will explore the consequences of asymmetric information in a number of different market situations. It will describe the inefficiencies that arise and discuss possible government intervention to correct these. Interpreted in this way, asymmetric information is one of the classic reasons for market failure and will prevent trading partners from realizing all the gains of trade. In addition to asymmetric information between trading parties, it can also arise between the government and the consumers and firms in the economy. When it does, it restricts the policies that the government can implement. Some aspects of how this affects the effectiveness of the government will be covered in this chapter; others will become apparent in later chapters. The main implication that will emerge for public intervention is that even if the government also faces informational imperfections, the incentives and constraints it faces often differ from those facing the private sector. Even when government faces exactly the same
informational problems, welfare can be improved by market intervention. There are interventions in the market that can make all parties better off.

### 10.2 Hidden Knowledge and Hidden Action

There are two basic forms of asymmetric information that can be distinguished. *Hidden knowledge* refers to a situation where one party has more information than the other party on the quality (or “type”) of a traded good or contract variable. *Hidden action* is when one party can affect the “quality” of a traded good or contract variable by some action, and this action cannot be observed by the other party.

Examples of hidden knowledge abound. Workers know more about their own abilities than the firm does; doctors know more about their own skills, the efficacy of drugs, and what treatment patients need than do either the patients themselves or the insurance companies; the person buying life insurance knows more about his health and life expectancy than the insurance firm; when an automobile insurance company insures an individual, the individual may know more than the company about her inherent driving skill and hence about her probability of having an accident; the owner of a car knows more about the quality of the car than potential buyers; the owner of a firm knows more about the firm than a potential investor; the borrower knows more about the riskiness of his project than the lender does; and not least, in the policy world, policy makers know more about their competence than the electorate.

Hidden knowledge leads to the *adverse selection* problem. To introduce this, suppose that a firm knows that there are high-productivity and low-productivity workers and that it offers a high wage with the intention of attracting high-productivity workers. Naturally this high wage will also prove attractive to low-productivity workers, so the firm will attract a combination of both types. If the wage is above the average productivity, the firm will make a loss and be forced to lower the wage. This will result in high-productivity workers leaving and average productivity falling. Consequently the wage must again be lowered. Eventually the firm will be left with only low-productivity workers. The adverse selection problem is that the high wage attracts the workers the firm wants (the high-productivity) and the ones it does not (the low-productivity). The observation that the firm will eventually be left with only low-productivity workers reflects the old maxim that “The bad drives out the good.”

There are also plenty of examples of hidden action. The manager of a firm does not seek to maximize the return for shareholders but instead trades off her remuneration for less work effort. Firms may find it most profitable to make unsafe products when
quality is not easily observed. Employers also want to know how hard their workers work. Insurers want to know what care their insured take to avoid an accident. Lenders want to know what risks their borrowers take. Patients want to know if doctors provide the correct treatment or if, in an attempt to protect themselves from malpractice suits, they choose conservative medicine, ordering tests and procedures that may not be in the patient’s best interests, and surely not worth the costs. The tax authority wants to know if taxing more may induce people to work less or to conceal more income. Government wants to know if more generous pension replacement rates may induce people to retire earlier. A welfaristic government will worry about the recipients of welfare spending too much and investing too little, thus being more likely to be in need again in the future. This concern will also be present among altruistic parents who cannot commit not to help out their children when needy and governments who cannot commit not to bail out firms with financial difficulties.

From hidden actions arises the moral hazard problem. This refers to the inefficiency that arises due to the difficulties in designing incentive schemes that ensure the right actions are taken. For instance, the price charged for insurance must take into account the fact that an insured person may become more careless once they have the safety net of insurance cover.

10.3 Actions or Knowledge?

Although the definitions given above make moral hazard and adverse selection seem quite distinct, in practice, it may be quite difficult to determine which is at work. The following example, due to Milgrom and Roberts, serves to illustrate this point.

A radio story in the summer of 1990 reported a study on the makes and models of cars that were observed going through intersections in the Washington, DC, area without stopping at the stop signs. According to the story, Volvos were heavily overrepresented: the fraction of cars running stop signs that were Volvos was much greater than the fraction of Volvos in the total population of cars in the DC area. This is initially surprising because Volvo has built a reputation as an especially safe car that appeals to sensible, safety-conscious drivers. In addition Volvos are largely bought by middle-class couples with children. How then is this observation explained?

One possibility is that people driving Volvos feel particularly safe in this sturdy, heavily built, crash-tested car. Thus they are willing to take risks that they would not take in another, less safe car. This implies that driving a Volvo leads to a propensity to run stop signs. This is essentially a moral hazard explanation: the car is a form of
insurance, and having the insurance alters behavior in a way that is privately rational but socially undesirable.

A second possibility is that the people who buy Volvos know that they are bad drivers who are apt, for example, to be paying more attention to their children in the back seat than to stop signs. The safety that a Volvo promises is then especially attractive to people who have this private information about their driving, so they buy this safe car in disproportionately large numbers. Hence a propensity for running stop signs leads to the purchase of a Volvo. This is essentially a self-selection story: Volvo buyers are privately informed about their driving habits and abilities and choose the car accordingly.

This self-selection is not necessarily adverse selection. It only becomes adverse selection if it imposes costs on Volvo. Quite the opposite may in fact be true, and the self-selection of customers can be very profitable.

It is also typically difficult to disentangle the moral hazard problem from the adverse selection problem in antipoverty programs because it is difficult to decide whether poverty is due to a lack of productive skill (adverse selection) or rather to a lack of effort from the poor themselves who know they will get welfare assistance anyway (moral hazard).

10.4 Market Unraveling

10.4.1 Hazard Insurance

In the Introduction we noted that asymmetric information can lead to a breakdown in trade as the less-informed party began to realize that the least desirable potential partners are those who are more willing to exchange. This possibility is now explored more formally in a model of the insurance market in which individuals differ in their accident probabilities. The basic conclusion to emerge is that in equilibrium some consumers do not purchase insurance, even though they could profitably be sold insurance if accident probabilities were observable to insurance companies.

Assume that there is a large number of insurance companies and that the insurance market is competitive. The insurance premium is based on the level of expected risk among those who accept offers of insurance. Competition ensures that profits are zero in equilibrium through entry and exit. Furthermore, if there is any new insurance contract that can be offered that will make a positive profit given the contracts already available, then one of the companies will choose to offer it.
The demand for insurance comes from a large number of individuals. These can be broken down into many different types of individual who differ in their probability of incurring damage of value $d = 1$. The probability of damage for an individual is given by $\theta$. Different individuals have different values of $\theta$, but all values lie between 0 and 1. If $\theta = 1$, the individual is certain to have an accident. Asymmetric information is introduced by assuming that each individual knows their own value of $\theta$ but that it is not observable by the insurance companies. The insurance companies do know (correctly) that risks are uniformly distributed in the population over the interval $[0, 1]$.

All of the individuals are risk averse, meaning that they are willing to pay an insurance premium to avoid facing the cost of damage. For each type the maximal insurance premium that they are willing to pay, $\pi(\theta)$, is given by

$$\pi(\theta) = [1 + \alpha] \theta,$$

where $\alpha > 0$ measures the level of risk aversion.

The assumption of competition among the insurance companies implies that in equilibrium they must earn zero profits. Now assume that insurance companies just offer a single insurance policy to all customers. Given the premium (or price) of the policy, $\pi$, the policy will be purchased by all the individuals whose expected value of damage is greater than or equal to this. That is, an individual will purchase the policy if

$$\pi(\theta) \geq \pi,$$

(10.2)

If a policy is to break even with zero profit, the premium for this policy must just equal the average value of damage for those who choose to purchase the policy. Hence (10.2) can be used to write the break-even condition as

$$\pi = E(\theta : \pi(\theta) \geq \pi),$$

(10.3)

which is just the statement that the premium equals expected damage. Returning to (10.1), the condition that $\pi(\theta) \geq \pi$ is equivalent to $[1 + \alpha] \theta \geq \pi$ or $\theta \geq \frac{\pi}{1 + \alpha}$. Using the fact that the $\theta$ is uniformly distributed gives

$$E(\theta : \pi(\theta) \geq \pi) = E\left(\theta : \frac{\pi}{1 + \alpha} \leq \theta \leq 1 \right) = \frac{1}{2} \left[ \frac{\pi}{1 + \alpha} + 1 \right].$$

(10.4)

The equilibrium premium then satisfies

$$\pi = \frac{1}{2} \left[ \frac{\pi}{1 + \alpha} + 1 \right],$$

(10.5)

or
\[ \pi = \frac{1 + \alpha}{1 + 2\alpha}. \]  

(10.6)

This equilibrium is illustrated in figure 10.1. It occurs where the curve \( E(\theta : \pi(\theta) \geq \pi) \) crosses the 45° line—this intersection is the value given in (10.6). It can be seen from the figure that insurance is only taken by those with high risks, namely all those with risk \( \theta \geq \frac{1}{1 + 2\alpha} \). This reflects the process of market unraveling through which only a small fraction of the potential consumers are actually served in equilibrium. The level of the premium is too high for the low-risk to find it worthwhile to take out the insurance. This outcome is clearly inefficient, since the first-best outcome requires insurance for all consumers. To see this, note that the premium a consumer of type \( \theta \) is willing to pay satisfies

\[ \pi(\theta) = (1 + \alpha)\theta > \theta \quad \text{for all } \theta. \]  

(10.7)

Therefore everyone is willing to pay more than the price the insurance companies need to break even if they could observe probabilities of accident.

This finding of inefficiency is a consequence of the fact that the insurance companies cannot distinguish the low-risk consumers from the high-risk. When a single premium is offered to all consumers, the high-risk consumers force the premium up, and this drives the low-risk out of the market. This is a simple example of the mechanism of
adverse selection in which the bad types always find it profitable to enter the market at the expense of the good. Without any intervention in the market, adverse selection will always lead to an inefficient equilibrium.

10.4.2 Government Intervention

There is a simple way the government can avoid the adverse selection process by which only the worst risks purchase insurance: it is by forcing all individuals to purchase the insurance. Compulsory insurance is then a policy that can make many consumers better off. With this, high-risk consumers benefit from a lower premium than the actual risk they face and lower than the level in (10.6)—it will actually be $\pi = \frac{1}{2} < \frac{1+\alpha}{1+2\alpha}$. The benefit for some of the low-risk is that they can now purchase a policy at a more favorable premium than that offered if only high-risk people purchased it. This benefits those close to the average who, although paying more for the policy than the level of the actual risk they face. Only the very low-risk are made worse off—they would rather have no insurance than pay the average premium.

The imposition of compulsory insurance may seem to be a very strong policy, since in few circumstances are consumers forced by the government to make specific purchases. But it is the policy actually used for many insurance markets. For instance, both automobile insurance and employee protection insurance are compulsory. Health care insurance and unemployment insurance are also compulsory. Aircraft have to be insured. Pleasure boats have to be compulsorily insured in some countries (e.g., France) but not in others (e.g., the United Kingdom), despite their representing a much greater capital investment than automobiles. One argument that could be advanced to explain this difference is the operation of self-selection into boating as a leisure activity: those who choose to do it are by their nature either low-risk or sufficiently cautious to insure without compulsion.

There is another role for government intervention. So far the arguments have concentrated on one of the simplest cases. Particularly restrictive was the assumption that the probability of damage was uniformly distributed across the population. It was this assumption (together with the proportional reservation premium) that ensured that the curve $E(\theta : \pi(\theta) \geq \pi)$ is a straight line with a single intersection with the 45 degree line. When the uniform distribution assumption is relaxed, $E(\theta : \pi(\theta) \geq \pi)$ will have a different shape, and the nature of equilibrium may be changed. There exist in fact functions for the distribution of types that lead to multiple equilibria. Such a case is illustrated in figure 10.2. In this figure $E(\theta : \pi(\theta) \geq \pi)$ crosses the 45 degree line three times so that there are three equilibria that differ in the size of the premium. At the
low-premium equilibrium, $E_1$, most of the population is able to purchase insurance, but at the high-premium equilibrium, $E_3$, very few can.

Each of these equilibria is based on correct but different self-fulfilling beliefs. For example, if the insurance companies are pessimistic and expect that only high-risk consumers will take out insurance, they will set a high premium. Given a high premium, only the high-risk will choose to accept the policy. The beliefs of the insurance companies are therefore confirmed, and the economy becomes trapped in a high-premium equilibrium with very few consumers covered by insurance. This is clearly a bad outcome for the economy, since there are also equilibria with lower premiums and wider insurance coverage.

When there are multiple equilibria, the one with the lowest premium is Pareto-preferred—it gives more consumers insurance cover and at a lower price. Consequently, if one of the other equilibria is achieved, there is a potential benefit from government intervention. The policy the government should adopt is simple: it can induce the best equilibrium (that with the lowest premium) by imposing a limit on the premium that can be charged. If we are at the wrong equilibrium, the corresponding premium reduction (from $E_2 \rightarrow E_1$ or $E_3 \rightarrow E_1$) will attract the good risks, making the cheaper insurance policy $E_1$ sustainable. This policy is not without potential problems. To see these, assume that the government slightly miscalculates and sets the maximum premium below the premium of policy $E_1$. No insurance company can make a profit at this price, and all offers of insurance will be withdrawn. The policy will then worsen the outcome.
If set too high, one of the other equilibria may be established. To intervene successfully in this way requires considerable knowledge on the part of the government.

This analysis of the insurance market has shown how asymmetric information can lead to market unraveling with the bad driving out the good, and eventually to a position where fewer consumers participate in the market than is efficient. In addition asymmetric information can lead to multiple equilibria. These equilibria can also be Pareto-ranked. For each of these problems, a policy response was suggested. The policy of making insurance compulsory is straightforward to implement and requires little information on the part of the government. Its only drawback is that it cannot benefit all consumers, since the very low risk consumers are forced to purchase insurance they do not find worthwhile. In contrast, the policy of a maximum premium requires considerable information and has significant potential pitfalls.

10.5 Screening

If insurance companies are faced with consumers whose probabilities of having accidents differ, then it will be to the companies’ advantage if they can find some mechanism that permits them to distinguish between the high-risk and low-risk. Doing so allows them to tailor insurance policies for each type and hence avoid the pooling of risks that causes market unraveling.

The mechanism that can be used by the insurance companies is to offer a menu of different contracts designed so that each risk type self-selects the contract designed for it. By self-select, we mean that the consumers find it in their own interest to select the contract aimed at them. As we will show, self-selection will involve the high-risks being offered full insurance coverage at a high premium, while the low-risks are offered partial coverage at a low premium requiring them to bear part of the loss. The portion they have to bear consists of a deductible (an initial amount of the loss) and co-insurance (an extra fraction of the loss beyond the deductible). An equilibrium like this where different types purchase different contracts is called a separating equilibrium. This should be contrasted to the pooling equilibrium of the previous section in which all consumers of insurance purchased the same contract. Obviously the high-risks will lose from this separation, since they will no longer benefit from the lower premium resulting from their pooling with the low-risks.

To model self-selection, we again assume that the insurance market is competitive so that in equilibrium insurance companies will earn zero profits. Rather than have a continuous range of different types, we now simplify by assuming there are just two
types of agents. The high-risk agents have a probability of an accident occurring of $p_h$, and the low-risks a probability $p_\ell$, with $p_h > p_\ell$. The two types form proportions $\lambda_h$ and $\lambda_\ell$ of the total population, where $\lambda_h + \lambda_\ell = 1$. Both types have the same fixed income, $r$, and suffer the same fixed damage, $d$, in the case of an accident.

If a consumer of type $i$ buys an insurance policy with a premium $\pi$ and payout (or coverage) $\delta$, the expected utility of this consumer type is given by

$$V_i(\delta, \pi) = p_i u(r - d + \delta - \pi) + (1 - p_i) u(r - \pi).$$

(10.8)

When the consumer purchases no insurance (so $\pi = 0$ and $\delta = 0$), expected utility is

$$V_i(0, 0) = p_i u(r - d) + (1 - p_i) u(r).$$

(10.9)

It is assumed that the consumer is risk averse, so the utility function, $u(\cdot)$, is concave.

The timing of the actions in the model is described by the following two stages:

• **Stage 1** Firms simultaneously choose a menu of insurance contracts $S_i = (\delta_i, \pi_i)$ with contract $i$ intended for consumers of type $i$.

• **Stage 2** Consumers choose their most preferred contract (not necessarily the one the insurance companies intended for them!).

We now analyze the equilibrium of this insurance market under a number of different assumptions on information.

### 10.5.1 Perfect Information Equilibrium

In the perfect information equilibrium the insurance companies are assumed to be able to observe the type of each consumer; that is, they know exactly the accident probability of each customer. This case of perfect information is used as a benchmark to isolate the consequences of the asymmetric information that is soon to be introduced.

Figure 10.3 illustrates the equilibrium with perfect information. The curved lines are indifference curves—one curve is drawn for each type. The steeper curve is that of the high-risk. The indifference curves are positively sloped because consumers are willing to trade off greater coverage for a higher premium. They are concave because of risk aversion. It is assumed that willingness to pay for extra coverage increases with the probability of having an accident. This makes the indifference curves of the high-risk steeper at any point than those of the low-risk so that the indifference curves satisfy the single-crossing property. Single crossing means that any pair of indifference curves—one for the low-risk and one for the high-risk—can only cross once. With full information the insurance companies know the accident probability. They can then
offer contracts that trade off a higher premium for increased coverage at the rate of the accident probability. That is, low-risk types can be offered any contract \{\pi, \delta\} satisfying \(\pi = p_\ell \delta\), and the high-risk any contract satisfying \(\pi = p_h \delta\). These equations give the two straight lines in figure 10.3. These are the equilibrium contracts that will be offered. To see this, note that if an insurance company offers a contract that is more generous (charges a lower premium for the same coverage), this contract must make a loss, and it will be withdrawn. Conversely, if a less generous contract is offered (so has a higher premium for the same coverage), other companies will be able to better it without making a loss. Therefore it will never be chosen.

Given this characterization of the equilibrium contracts, the final step is to observe that when these contracts are available, both types will choose to purchase full insurance coverage. They will choose \(\delta = d\) and pay the corresponding premium. Hence the competitive equilibrium when types are observable by the companies is a pair of insurance contracts \(S_h^*, S_\ell^*\), where

\[
S_h^* = (d, p_h d) \tag{10.10}
\]

and

\[
S_\ell^* = (d, p_\ell d) \tag{10.11}
\]

so there is full coverage and actuarially fair premia are charged. As for any competitive equilibrium with full (hence symmetric) information, this outcome is Pareto-efficient.
10.5.2 Imperfect Information Equilibrium

Imperfect information is introduced by assuming that the insurance companies cannot distinguish a low-risk consumer from a high-risk. We also assume that it cannot employ any methods of investigation to elicit further information. As we will discuss later, insurance companies routinely do try to obtain further information. The reasons why they do and the consequences of doing so will become clear once it is understood what happens if they don’t.

Given these assumptions, the insurance companies cannot offer the contracts that arose in the full-information competitive equilibrium. The efficient contract for the low-risk provides any given degree of coverage at a lower premium than the contract for the high-risk. Hence both types will prefer the contract intended for the low-risk (this is adverse selection again!). If offered, an insurance company will charge a premium based on the low-risk accident probability but have to pay claims at the population average probability. It will therefore make a loss and have to be withdrawn. This argument suggests what the insurance companies have to do: if they wish to offer a contract that will attract the low-risk type, the contract must be designed in such a way that it does not also attract the high-risk. This requirement places constraints on the contracts that can be offered and is what prevents the attainment of the efficient outcome.

Assume now that insurance companies offer a contract \( S_h \) designed for the high-risk and a contract \( S_\ell \) designed for the low-risk. To formally express the comments in the previous paragraph, we say that when types are not observable, the contracts \( S_h \) and \( S_\ell \) have to satisfy the self-selection (or incentive-compatibility) constraints. These constraints require the low-risk to find that the contract \( S_\ell \) offers them at least as much utility as the contract \( S_h \), with the converse holding for the high-risk. If these constraints are satisfied, the low-risk will choose the contract designed for them, as will the high-risk. The self-selection constraints can be written as

\[
V_\ell(S_\ell) \geq V_\ell(S_h) \quad (IC_u) \tag{10.12}
\]

and

\[
V_h(S_h) \geq V_h(S_\ell) \quad (IC_d). \tag{10.13}
\]

(These are labeled \( IC_u \) and \( IC_d \) because the first has the low-risk types looking “up” at the contract of the high-risk, the second has the high-risk looking “down” at the contract of the low-risk. This becomes clear in figure 10.4.) As we have already remarked, the contracts \( S_h^*, S_\ell^* \) arising in the full-information equilibrium do not satisfy \( IC_d \): the high-risk will always prefer the low-risk’s contract \( S_\ell^* \).
There is only one undominated pair of contracts that achieves the desired separation. By undominated, we mean that no other pair of separating contracts can be introduced that makes a positive profit in competition with the undominated contracts. The properties of the pair are that the high-risk type receives full insurance at an actuarially fair rate. The low-risk do not receive full insurance. They are restricted to partial coverage, with the extent of coverage determined by where the indifference curve of the high-risk crosses the actuarially fair insurance line for the low-risk. In addition the constraint (10.13) is binding while the constraint (10.12) is not. This feature, that the “good” type (here the low-risk) are constrained by the “bad” type (here the high-risk), is common to all incentive problems of this kind.

It can easily be seen that the insurance contracts are undominated by any other pair of separating contracts and make zero profit for the insurance companies. To see that no contract can be introduced that will appeal to only one type and yield positive profit, assume that such a contract was aimed at the high-risk. Then it must be more favorable than the existing contract; otherwise, it will never be chosen. But the existing contract is actuarially fair, so any contract that is more favorable must make a loss. Alternatively, a contract aimed at the low-risk will either attract the high-risk too, and so not separate, or, if it attracts only low-risk, will be unprofitable. There remains, though, the possibility that a pooling contract can be offered that will attract both types and be profitable.

To see how this can arise, consider figure 10.5. A pooling contract will appeal to both types if it lies below the indifference curves attained by the separating contracts (lower premium and possibly greater coverage). Since the population probability of
an accident occurring is \( p = \lambda_h p_h + \lambda_\ell p_\ell \), an actuarially fair pooling contract \( \{\pi, \delta\} \) will relate premium and coverage by \( \pi = p\delta \). When \( \lambda_h \) is large, the pooling contract will lie close to the actuarially fair contract of the high-risk and hence will be above the indifference curve attained by the low-risks in the separating equilibrium. In this case the separating contracts will form an equilibrium. Conversely, when \( \lambda_\ell \) is large, the pooling contract will lie close to the actuarially fair contract for the low-risk. It will therefore be below the indifference curves of both types in the separating equilibrium and, when offered, will attract both low- and high-risk types. When this arises, the separating contracts cannot constitute an equilibrium, since an insurance company can offer a contract marginally less favorable than the actuarially fair pooling contract, attract all consumers, and make a profit.

To summarize, there exists a pair of contracts that separate the population and are not dominated by any other separating contracts. On the one hand, they constitute an equilibrium if the proportion of high-risk consumers in the population is sufficiently large (so that the low-risks prefer to separate and choose partial coverage rather than be pooled with many high-risks and pay a higher premium). On the other hand, if the proportion of low-risk is sufficiently large, there will be a pooling contract that is preferred by both types and profitable for an insurance company. In this latter case there can be no separating equilibrium.

By using the same kind of argument, it can be shown that there is no pooling equilibrium. Consider a pooling contract \( S \) with full coverage and average risk premium. Any
contract $S^\circ = (\delta^\circ, \pi^\circ)$ in the wedge formed by the two indifference curves in figure 10.6 attracts only low-risks and makes a positive profit. It will therefore be offered and attract the low-risk away from the pooling contract. Without the low-risk the pooling contract will make a loss.

In conclusion, there is no pooling equilibrium in this model of the insurance market. There may be a separating equilibrium, but this depends on the population proportions. When there is no separating equilibrium, there is no equilibrium at all. Asymmetric information either causes inefficiency by leading to a separating equilibrium in which the low-risk have too little insurance cover, or it results in there being no equilibrium at all. In the latter case we cannot predict what the outcome will be.

10.5.3 Government Intervention

Government intervention in this insurance market is limited by the same information restriction that affects firms: they cannot tell who is low-risk or high-risk directly but can only make inferences from observing choices. This has the consequence that it restricts policy intervention to be based on the same information as that available to the insurance companies. Even under these restrictions the government can achieve a Pareto improvement by imposing a cross-subsidy from low-risks to high-risks. It does this by subsidizing the premium of the high-risk and taxing the premium of the low-risk. It can do that without observing risk by imposing a minimal coverage for all at the average risk premium.
The reason why this policy works is that the resulting transfer from the low-risks to the high-risks relaxes the incentive constraint ($IC_d$). This makes the set of insurance policies that satisfies the constraints larger and so benefits both types. This equilibrium cannot be achieved by the insurance companies because it would require them all to act simultaneously. This is an example of a coordination failure that prevents the attainment of a better outcome.

This policy is illustrated in figure 10.7. Let the subsidy to the high-risk be given by $t_h$ and the tax on the low-risk be $t_L$. The tax and subsidy are related to the transfer, $t$, by the relationships $t_h = \frac{t}{\lambda_h}$ and $t_L = \frac{t}{\lambda_L}$. The premium for the low-risk then becomes $p_L + t_L$ and for the high-risks $p_h - t_h$. As figure 10.7 shows, the high-risks are strictly better off and the low-risks are as well off as before because higher coverage is now incentive compatible. The policy intervention has therefore engineered a Pareto improvement. It should be noted that the government has improved the outcome, even though it has the same information as the insurance companies. Government achieves this improvement through its ability to coordinate the transfer—something the insurance companies cannot do.

10.6 Signaling

The fundamental feature at the heart of asymmetric information is the inability to distinguish the good from the bad. This is to the detriment of both the seller of a good
article, who fails to obtain its true value, and to the purchaser, who would rather pay a higher price for something that is known to be good. It seems natural that this situation would be improved if the seller could convey some information that convinces the purchaser of the quality of the product. For instance, the seller may announce the names of previous satisfied customers (employment references can be interpreted in this way) or provide an independent guarantee of quality (e.g., a report on the condition of a car by a motoring organization). Warranties can also serve as signals of quality for durable goods because, if a product is of higher quality, it is less costly for the seller to offer a longer warranty. Such information, generally termed signals, can be mutually beneficial.

It is worth noting the difference between screening and signaling. The less-informed players (like the insurance companies) use screening (different insurance contracts) to find out what the better-informed players (insurance customer) know (their own risk). In contrast, more-informed players use signals to help the less-informed players find out the truth.

For a signal to work it must satisfy certain criteria. First, it must be verifiable by the receiver (i.e., the less-informed agent). Being given the name of a satisfied customer is not enough—it must be possible to check back that they are actually satisfied. Second, it must be credible. In the case of an employment reference this is dependent partly on the author of the reference having a reputation to maintain and partly on the possibility of legal action if false statements are knowingly made. Finally the signal must also be costly for the sender (i.e., the better-informed agent) to obtain and the cost must differ between various qualities of sender. In the case of an employment reference this is obtained by a record of quality work. Something that is either costlessly obtainable by both the senders of low- and high-quality or equally costly cannot have any value in distinguishing between them. We now model such signals and see the effect that they have on the equilibrium outcome.

The modeling of signaling revolves around the timing of actions. The basic assumption is that the informed agent moves first and invests in acquiring a costly signal. The uninformed party then observes the signals of different agents and forms inferences about quality on the basis of these signals. An equilibrium is reached when the chosen investment in the signal is optimal for each informed agent and the inferences of the uninformed about the meaning of signals are justified by the outcomes. As we will see, the latter aspect involves self-supporting beliefs: they may be completely irrational, but the equilibrium they generate does not provide any evidence to falsify them.
10.6.1 Educational Signaling

To illustrate the consequences of signaling, we will consider a model of productivity signaling in the labor market. The model has two identical firms that compete for workers through the wages they offer. The set of workers can be divided into two types according to their productivity levels. Some of the workers are innately low-productivity in the form of employment offered by the firms, while the others are high-productivity. Without any signaling, the firms are assumed to be unable to judge the productivity of a worker.

The firms cannot directly observe a worker’s type before hiring, but high-productivity workers can signal their productivity by being educated. Education itself does not alter productivity, but it is costly to acquire. Firms can observe the level of education of a potential worker and condition their wage offer on this. Hence education is a signal. Investment in education will be worthwhile if it earns a higher wage. To make it an effective signal, it must be assumed that obtaining education is more costly for the low-productivity than it is for the high-productivity; otherwise, both will have the same incentive for acquiring it.

Formally, let $\theta_h$ denote the productivity of a high-productivity worker and $\theta_\ell$ that of a low-productivity worker, with $\theta_h > \theta_\ell$. The workers are present in the population in proportions $\lambda_h$ and $\lambda_\ell$, so $\lambda_h + \lambda_\ell = 1$. The average productivity in the population is given by

$$E(\theta) = \lambda_h \theta_h + \lambda_\ell \theta_\ell. \tag{10.14}$$

Competition between the two firms ensures that this is the wage that would be paid if there were no signaling and the firms could not distinguish between workers. For a worker of productivity level $\theta$, the cost of obtaining education level $e$ is

$$C(e, \theta) = e/\theta, \tag{10.15}$$

which satisfies the property that any given level of education is more costly for a low-productivity worker to obtain.

The firms offer wages that are (potentially) conditional on the level of education; “potentially” is added because there may be equilibria in which the firms ignore the signal. The wage schedule is denoted by $w(e)$. Given the offered wage schedule, the workers aim to maximize utility, which is defined as wages less the cost of education. Hence their decision problem is

$$\max_{\{e\}} w(e) - e/\theta. \tag{10.16}$$
As shown in figure 10.8, the preferences in (10.16) satisfy the single-crossing property when defined over wages and education. Here $V_\ell$ denotes an indifference curve of a low-productivity worker and $V_h$ that of a high-productivity type. At any point the greater marginal cost of education for the low-productivity type implies that they have a steeper indifference curve.

An equilibrium for this economy is a pair \( \{e^*(\theta), w^*(e)\} \), where \( e^*(\theta) \) determines the level of education as a function of productivity and \( w^*(e) \) determines the wage as a function of education. In equilibrium these functions must satisfy three properties:

1. No worker wants to change his education choice given the wage schedule \( w^*(e) \).
2. No firm wants to change its wage schedule given its beliefs about worker types and education choices \( e^*(\theta) \).
3. Firms have correct beliefs given the education choices.

The first candidate for an equilibrium is a separating equilibrium in which low- and high-productivity workers choose different levels of education. Any separating equilibrium must satisfy

\begin{align*}
& \text{(i) } e^*(\theta_\ell) \neq e^*(\theta_h), \\
& \text{(ii) } w^*(e^*(\theta_\ell)) = \theta_\ell, \\
& \hspace{1cm} w^*(e^*(\theta_h)) = \theta_h,
\end{align*}
(iii) $w^*(e^* (\theta_\ell)) - \frac{e^*(\theta_\ell)}{\theta_\ell} \geq w^*(e^* (\theta_h)) - \frac{e^*(\theta_h)}{\theta_h}$,
(iv) $w^*(e^* (\theta_h)) - \frac{e^*(\theta_h)}{\theta_h} \geq w^*(e^* (\theta_\ell)) - \frac{e^*(\theta_\ell)}{\theta_\ell}$.

Condition (i) is the requirement that low- and high-productivity workers choose different education levels, (ii) that the wages are equal to the marginal products, and (iii) that the choices are individually rational for the consumers. The values of the wages given in (ii) are a consequence of signaling and competition between firms. Signaling implies workers of different productivities are paid different wages. If a firm paid a wage above the marginal product, it would make a loss on each worker employed. This cannot be profit maximizing. Alternatively, if one firm paid a wage below the marginal productivity, the other would have an incentive to set its wage incrementally higher. This would capture all the workers of that productivity level and would be the more profitable strategy. Therefore the only equilibrium values for wages when signaling occurs are the productivity levels. This leaves only the levels of education to be determined.

The equilibrium level of education for the low-productivity workers is found by noting that if they choose not to act like the high-productivity, then there is no point in obtaining any education—education is simply a cost that does not benefit them. Hence $e^*(\theta_\ell) = 0$. By this fact and that wages are equal to productivities, the level of education for the high-productivity workers can be found from the incentive compatibility constraints. From (iiia),

$$\theta_\ell \geq \theta_h - \frac{e^*(\theta_h)}{\theta_\ell},$$
(10.17)

or

$$e^*(\theta_h) \geq \theta_h[\theta_h - \theta_\ell].$$
(10.18)

Condition (10.18) provides the minimum level of education that will ensure that the low-productivity workers choose not to be educated. Now from (iiib) it follows that

$$\theta_h - \frac{e^*(\theta_h)}{\theta_h} \geq \theta_\ell,$$
(10.19)

or

$$\theta_h[\theta_h - \theta_\ell] \geq e^*(\theta_h).$$
(10.20)
Hence a complete description of the separating equilibrium is

\[ e^*(\theta_\ell) = 0, \theta_\ell [\theta_h - \theta_\ell] \leq e^*(\theta_h) \leq \theta_h [\theta_h - \theta_\ell], \quad (10.21) \]

\[ w(e^*(\theta_\ell)) = \theta_\ell, \quad w(e^*(\theta_h)) = \theta_h, \quad (10.22) \]

so the low-productivity workers obtain no education, the high-productivity have education somewhere between the two limits and both are paid their marginal products. An equilibrium satisfying these conditions is illustrated in figure 10.9.

Since there is a range of possible values for \( e^*(\theta_h) \), there is not a unique equilibrium but a set of equilibria differing in the level of education obtained by the high-productivity. This set of separating equilibria can be ranked according to criterion of Pareto-preference. Clearly, changing the level of education \( e^*(\theta_h) \) within the specified range does not affect the low-productivity workers. However, the high-productivity workers always prefer a lower level of education, since education is costly. Therefore equilibria with lower \( e^*(\theta_h) \) are Pareto-preferred, and the most preferred equilibrium

![Figure 10.9](image-url)

**Figure 10.9**
Separating equilibrium
is that with \( e^*(\theta_h) = \theta_\ell [\theta_h - \theta_\ell] \). The Pareto-dominated separating equilibria are supported by the high-productivity worker’s fear that choosing less education will give an unfavorable impression of their productivity to the firm and thus lead to a lower wage.

There are arguments (called refinements of equilibrium) to suggest that this most-preferred equilibrium will actually be the one that emerges. Let the equilibrium level of education for the high-productivity type, \( e^*(\theta_h) \), be above the minimum required to separate. Denote this minimum \( e^0 \). Now consider the firm observing a worker with an education level at least equal to \( e^0 \) but less than \( e^*(\theta_h) \). What should a firm conclude about this worker? Clearly, the worker cannot be low-productivity, since such a choice is worse for them than choosing no education. Hence the firm must conclude that the worker is of high productivity. Realizing this, it then pays the worker to deviate, since it would reduce the cost of an education. This argument can be repeated until \( e^*(\theta_h) \) is driven down to \( e^0 \).

Signaling allows the high-productivity to distinguish themselves from the low-productivity. It might be thought that this improvement in information transmission would make signaling socially beneficial. However, this need not be the case, since the act of signaling is costly and does not add to productivity. The alternative to the signaling equilibrium is pooling where both types purchase no education and are paid a wage equal to the average productivity. The low-productivity would prefer this equilibrium as it raises their wage from \( \theta_\ell \) to \( E(\theta) = \lambda_h \theta_h + \lambda_\ell \theta_\ell \). For the high-productivity pooling is preferred if

\[
E(\theta) = \lambda_h \theta_h + \lambda_\ell \theta_\ell > \theta_\ell [\theta_h - \theta_\ell] / \theta_h. \tag{10.23}
\]

Since \( \lambda_\ell = 1 - \lambda_h \), this inequality will be satisfied if

\[
\lambda_h > 1 - \frac{\theta_\ell}{\theta_h}. \tag{10.24}
\]

Hence, when there are sufficiently many high-productivity workers so that the average wage is close to the high productivity level, the separating equilibrium is Pareto-dominated by the pooling equilibrium. In these cases signaling is individually rational but socially unproductive. Again, the Pareto-dominated separating equilibrium is sustained by the high-productivity workers’ fear that lowering their education would give a bad impression of their ability to the firms and thus lead to lower wage. Actually the no-signaling pooling equilibrium is not truly available to the high-productivity workers.
If they get no education, firms will believe they are low-productivity workers and then offer a wage of $\theta_L$. So we get the paradoxical situation that high-productivity workers choose to signal, although they are worse off when signaling.

If the government were to intervene in this economy, it has two basic policy options. The first is to allow signaling to occur but to place an upper limit on the level of education equal to $\theta_L [\theta_h - \theta_L]$. It might choose to do this in those cases where the pooling equilibrium does not Pareto-dominate the separating equilibrium. There is, though, one problem with banning signaling and enforcing a pooling equilibrium. The pooling equilibrium requires the firms to believe that all workers have the same ability. If the firms were to “test” this belief by offering a higher wage for a higher level of education, they would discover that the belief was incorrect. This is illustrated in figure 10.10. A low-productivity worker would be better off getting no education than getting education above $e^*$ whatever the firm’s belief and the resulting wage. Therefore the firm should believe that any worker choosing an education level above $e^*$ has high productivity and should be offered a wage $\theta_h$. But, if this is so, the high-productivity worker could do better than the pooling equilibrium by deviating to an education level slightly in excess of $e^*$ to get a wage $\theta_h$. Therefore the pooling equilibrium is unlikely, since it involves unreasonable beliefs from the firms.

10.6.2 Implications

The model of educational signaling shows how an unproductive but costly signal can be used to distinguish between quality levels through a set of self-supporting beliefs. There will be a set of Pareto-ranked equilibria with the lowest level of signal the most preferred. Although there is an argument that the economy must achieve the Pareto-dominating signaling equilibrium, it is possible that this may not happen. If it does not, the economy may become settled in a Pareto-inferior separating equilibrium. Even if this does not happen, it is still possible for the pooling equilibrium to Pareto-dominate the separating equilibrium. This will occur when the high-productivity workers are relatively numerous in the population, since in that case almost every worker is getting unproductive but costly education to separate themselves from the few bad workers.

There are several policy implications of these results. In a narrow interpretation, they show how the government can increase efficiency and make everyone better off by restricting the size of signals that can be transmitted. Alternatively, the government could improve the welfare of everyone by organizing a cross-subsidy from the good to
the bad workers. This can take the form of a minimum wage for the low-productivity workers in excess of their productivity financed by wage limit for the high-productivity workers that is below their productivity. Notice that a ban on signaling is an extreme form of such cross-subsidization, since it forces the same wage for all. When the pooling equilibrium is Pareto-preferred, signals should be eliminated entirely. More generally, the model demonstrates how market solutions may endogenously arise to combat the problems of asymmetric information. These solutions can never remove the problems entirely—someone must be bearing the cost of improving information flows—and can even exacerbate the situation.

The basic problem for the government in responding to these kinds of problems is that it does not have a natural informational advantage over the private agents. In the model of education there is no reason to suppose that the government is any more able to tell the low-productivity workers from the high-productivity (in fact there is every reason to suspect that the firms would be better equipped to do this). Faced with these kinds of problems, the government would have little to offer beyond the cross-subsidization we have just mentioned.

![Diagram](image_url)

**Figure 10.10**
Unreasonable beliefs
10.7 Moral Hazard (Hidden Action)

A moral hazard problem arises when an agent can affect the “quality” of a traded good or contract variable by some action that is not observed by other agents. For instance, a houseowner once insured may become lax in her attention to security, such as leaving windows open, in the knowledge that if burgled she will be fully compensated. Or a worker, once in employment, may not fully exert himself, reasoning that his lack of effort may be hidden among the effort of the workforce as a whole. Such possibilities provide the motive for contracts to be designed that embody incentives to lessen these effects.

In the case of the worker, the employment contract could provide for a wage that is dependent on some measure of the worker’s performance. Ideally the measure would be his exact productivity, but except for the simplest cases, this could be difficult to measure. Difficulties can arise because production takes place in teams (a production line can often be interpreted as a team) with the effort of the individual team member impossible to distinguish from the output of the team as a whole. They can also arise through randomness in the relation between effort and output. As examples, agricultural output is driven by the weather, maintenance tasks can depend on the (variable) condition of the item being maintained, and production can be dependent on the random quality of other inputs.

We now consider the design of incentive schemes in a situation with moral hazard. The model we choose embodies the major points of the previous discussion: effort cannot be measured directly, so a contract has to be based on some observable variable that roughly measures effort.

10.7.1 Moral Hazard in Insurance

The moral hazard problem that can arise in an insurance market is that effort on accident prevention is reduced when consumers become insured. If accident-prevention effort is costly, for instance, driving more slowly is time-consuming or eating a good diet is less enjoyable, then a rational consumer will seek to reduce such effort when it is beneficial to do so (and the benefits are raised once insurance is offered). Insurance companies must counteract this tendency through the design of their contracts.

To model this situation, assume an economy populated by many identical agents. The income of an agent is equal to $r$ with probability $1 - p$ and $r - d$ with probability...
Here $p$ is interpreted as the probability of an accident occurring and $d$ the monetary equivalent of the accident damage. Moral hazard is introduced by assuming that the agents are able to affect the accident probability through their prevention efforts.

To simplify, it is assumed that effort, $e$, can take one of two values. If $e = 0$, an agent is making no effort at accident prevention and the probability of an accident is $p(0)$. Alternatively, if $e = 1$, the agent is making maximum effort at accident prevention and the probability is $p(1)$. In line with these interpretations, it is assumed that $p(0) > p(1)$, so the probability of the accident is higher when no effort is undertaken. The cost of effort for the agents, measured in utility terms, is $c(e) \equiv ce$.

In the absence of insurance, the preferences of the agent are described by the expected utility function

$$U^o(e) = p(e)u(r - d) + (1 - p(e))u(r) - ce,$$

where $u(r - d)$ is the utility if there is an accident and $u(r)$ is the utility if there is no accident. It is assumed that the agent is risk averse, so the utility function $u(\cdot)$ is concave.

The value of $e$, either 0 or 1, is chosen to maximize this utility. Effort to prevent the accident will be undertaken ($e = 1$) if

$$U^o(1) > U^o(0).$$

Evaluating the utilities and rearranging shows that $e = 1$ if

$$c \leq c_0 \equiv [p(0) - p(1)] [u(r) - u(r - d)].$$

Here $c_0$ is the critical value of effort cost. If effort cost is below this value, effort will be undertaken. Therefore, in the absence of insurance, effort will be undertaken to prevent accidents if the cost of doing so is sufficiently small.

Consider now the introduction of insurance contracts. A contract consists of a premium $\pi$ paid by the consumer and an indemnity $\delta$, $\delta \leq d$, paid to the consumer if they are subject to an accident. The consumer’s preferences over insurance policies (meaning different combinations of $\pi$ and $\delta$) and effort are given by

$$U(e, \delta, \pi) = p(e)u(r - \pi + \delta - d) + (1 - p(e))u(r - \pi) - ce,$$

with $U(e, 0, 0) = U^o(e)$. 
10.7.2 Effort Observable

To provide a benchmark with which to measure the effects of moral hazard, we first analyze the choice of insurance contract when effort is observable by the insurance companies. In this case there can be no efficiency loss, since there is no asymmetry of information.

If the insurance company can observe $e$, it will offer an insurance contract that is conditional on effort choice. The contract will therefore be of the form $\{\delta(e), \pi(e)\}$, with $e = 0, 1$. Competition among the insurance companies ensures that the contracts on offer maximize the utility of a representative consumer subject to constraint that the insurance companies at least break even. To meet this latter requirement the premium must be no lower than the expected payment of indemnity. For a given $e$ (recall this is observed) the policy therefore solves

$$\max_{\{\delta, \pi\}} U(e, \delta, \pi) \quad \text{subject to} \quad \pi \geq p(e) \delta.$$  \hfill (10.29)

The solution to this is a policy

$$\{\delta^*(e) = d, \pi^*(e) = p(e)d\},$$  \hfill (10.30)

so that the damage is fully covered and the premium is fair given the effort level chosen. This is illustrated in figure 10.11. The straight line is the set of contracts that are fair (so $\pi = p(e) \delta$), and $I$ is the highest indifference curve that can be achieved given these contracts. (Note that utility increases with a lower premium and greater coverage.) The first-best contract is therefore full insurance with $\delta^*(e) = d$ and $\pi^*(e) = p(e)d$.

At the first-best contract, the resulting utility level is

$$U^*(e) = u(r - p(e)d) - ce.$$  \hfill (10.31)

Effort will be undertaken ($e = 1$) if

$$U^*(1) \geq U^*(0),$$  \hfill (10.32)

which holds if

$$c \leq c_1 \equiv u(r - p(1)d) - u(r - p(0)d).$$  \hfill (10.33)

That is, the cost of effort is less than the utility gain resulting from the lower premium.

An interesting question is whether the first-best contract encourages the supply of effort, in other words, whether the level of effort cost below which effort is supplied in the absence of the contract, $c_0$, is less than that with the contract, $c_1$. Calculations show
that the outcome may go in either direction depending on the accident probabilities associated with effort and no effort.

10.7.3 **Effort Unobservable**

When effort is unobservable, the insurance companies cannot condition the contract on it. Instead, they must evaluate the effect of the policies on the choices of the consumers and choose the policy taking this into account.

The preferences of the consumer over contracts are determined by the highest level of utility they can achieve with that contract, given that they have made the optimal choice of effort. Formally, the utility $V(\delta, \pi)$ arising from contract $(\delta, \pi)$ is determined by

$$V(\delta, \pi) \equiv \max_{e=0,1} U(e, \delta, \pi).$$

(10.34)

The basic analytical difficulty in undertaking the determination of the contract is the nonconvexity of preferences in the contract space $(\delta, \pi)$. This nonconvexity arises at the point in the contract space where the consumers switch from no effort $(e = 0)$ to full effort $(e = 1)$. When supplying no effort their preferences are determined by $U(0, \delta, \pi)$ and when they supply effort by $U(1, \delta, \pi)$. At any point $(\hat{\delta}, \hat{\pi})$ where $U(0, \hat{\delta}, \hat{\pi}) = U(1, \hat{\delta}, \hat{\pi})$, the indifference curve of $U(0, \hat{\delta}, \hat{\pi})$ is steeper than that of $U(1, \hat{\delta}, \hat{\pi})$ because the willingness to pay for extra coverage is higher when there is no

![Figure 10.11](image-url)

**Figure 10.11**

First-best contract
effort and thus a high risk of accident. This is illustrated in figure 10.12, where $\delta^*(\pi)$ denotes the locus of points where the consumer is indifferent to $e = 0$ and $e = 1$. This locus separates those who make effort from those who make no effort. For each premium $\pi$, there is an indemnity level $\delta^*(\pi)$ such that if $\delta < \delta^*(\pi)$, then $e = 1$, but if $\delta \geq \delta^*(\pi)$, then $e = 0$. This indemnity level rises with the premium, so $\delta^*(\pi)$ is an increasing function of $\pi$. In words, if the coverage rate for any given premium is too high, agents will no longer find profitable to undertake effort.

### 10.7.4 Second-Best Contract

The second-best contract maximizes the consumer’s utility subject to the constraint that it must at least break even. The optimization problem describing this can be written as that of maximizing $V(\delta, \pi)$ subject to the constraints that

\begin{align*}
(i) \quad \pi & \geq p(1) \delta \quad \text{for } \delta < \delta^*(\pi), \\
(ii) \quad \pi & \geq p(0) \delta \quad \text{for } \delta^*(\pi) \leq \delta < d.
\end{align*}

The first constraint applies if the consumer chooses to supply effort ($e = 1$) and requires that the contract break even. The second constraint is the break even condition if the consumer chooses to supply no effort ($e = 0$).
The problem is solved by calculating the solution under the first constraint and evaluating the resulting level of utility. Then the solution is found under the second constraint and utility is evaluated again. The two levels of utility are then compared, and the one yielding the highest utility is the optimal second-best contract. This reasoning provides two contracts that are candidates for optimality. These are illustrated in figure 10.13 by $E_0$ and $E_1$ and have the following properties:

- **Contract $E_0$**: No effort and full coverage at high price;
- **Contract $E_1$**: Effort and partial coverage at low price.

Which of these contracts is optimal will depend on the cost, $c$, of effort. When this cost is low, contract $E_1$ will be optimal and partial coverage will be offered to consumers. Conversely, when the cost is high, then it will be optimal to have no effort and contract $E_0$ will be optimal. By this reasoning, it follows that there must be some value of the cost of effort at which the switch is made between $E_0$ and $E_1$. Hence there exists a value of effort, $c_2$, with $c_2 < c_1$, such that $c \leq c_2$ implies that the second-best contract is $E_1$ and $c > c_2$ implies that the second-best contract is $E_0$.

It can now be shown that the second-best contract is inefficient. Since the critical level of cost, $c$, determining when effort is supplied satisfies $c < c_1$, the outcome has to be inefficient relative to the first-best. Furthermore there is too little effort if $c_2 < c < c_1$ and too little coverage if $c < c_2$. These results are summarized in table 10.1.
Table 10.1
Categorization of outcomes

<table>
<thead>
<tr>
<th>Cost of effort</th>
<th>$c_2$</th>
<th>$c_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st best</td>
<td>Effort, full coverage</td>
<td>No effort, full coverage</td>
</tr>
<tr>
<td>2nd best</td>
<td>Effort, partial coverage</td>
<td>No effort, full coverage</td>
</tr>
</tbody>
</table>

10.7.5 Government Intervention

The market failure associated with moral hazard is very profound. The moral hazard problem arises from the nonobservability of the level of care. When individuals are fully insured they tend to exert too little precaution but also over-use insurance. Consider, for instance, a patient who may be either sick with probability 0.09 or very sick with probability 0.01. In the two events his medical expenses will be $1,000 and $10,000. At a fair premium of $190 the patient will not have to pay anything if he gets sick and would buy such insurance if risk averse. But then suppose that when he is a little sick, there is some chance, however small, that he can be very sick. Then he would choose the expensive treatment given that there is no extra cost to the patient and all the extra cost is borne by the insurance company. Each individual ignores the effect of his reckless behavior and overconsumption on the premium, but when they all act like that, the premium increases. The lack of care by each inflates the premium, which generates a negative externality on others. An important implication is that the market cannot be efficient.

Another way to see this generic market inefficiency is that the provision of insurance in the presence of moral hazard causes the insured individual to receive less than the full social benefit of his care. As a result not only will the individual expend less than the socially optimal level of care but also there will be an insurance-induced externality.

This implies that the potential scope for government intervention with moral hazard is substantial. Can the government improve efficiency by intervention when moral hazard is present? In answering this question, it is important to specify what information is available to the government. For a fair evaluation of government intervention, it is natural to assume that the government has the same information as the private sector. In this case it can be argued that efficient government intervention is still possible. The beneficial effects of government intervention stem from the government’s capacity to tax and subsidize. For example, the government cannot monitor smoking, which has an adverse effect on health, any better than an insurance company. But the government can impose taxes, not only on cigarettes but also on commodities that are complements and subsidize substitutes that have a less adverse effect. Also the taxation of insurance
induces firms to offer insurance at less than fair price. As a consequence individuals buy less insurance and expend more effort (as efficiency requires).

10.8 Public Provision of Health Care

10.8.1 Efficiency

Economists do not expect the private market for health care insurance to function well. Our previous discussion suggests that informational problems result in the private provision of health insurance having incomplete and inefficient coverage. The existence of asymmetric information between insurers and insured leads to adverse selection, which can result in the market breaking down, and the nonexistence of certain types of insurance. The moral hazard problem can lead to incomplete insurance in the form of co-payments and deductibles for those who have insurance. Another problem caused by the presence of moral hazard is that the insured who become sick will want to overconsume and doctors will want to oversupply health care, since it is a third party that pays. It is not surprising therefore that the government may usefully intervene in the provision of health care.

There is strong evidence that in the OECD countries the public sector plays an important role in the provision of insurance for health care. From OECD health data, in 1994 the proportion of publicly provided health expenses was 44 percent in the United States, 70 percent in Germany, 73 percent in Italy, 75 percent in France, and 83 percent in Sweden and the United Kingdom. The question is why the government intervenes so extensively in the health care field. In answering the question, one must bear in mind that the government faces many of same informational problems as the private sector. Like a private insurer, it faces the moral hazard of patients who get insurance exerting too little effort in risk-reducing activities and overconsuming health services, and doctors having the incentive to oversupply health services at too high a cost.

One advantage of public provision is to prevent the adverse selection problem by making health coverage compulsory and universal. It is tempting to believe that the actual provision of insurance need not be public to accomplish this effect. Indeed the actual provision of health insurance could remain private and the government mandate that all individuals have to purchase health insurance and private insurers have to insure anyone who applies for insurance. However, mandates may be difficult to enforce at the individual level, and the incentive for private firms to accept only the good risks is a permanent concern. Another advantage of public provision is that as a predominant
insurer it can exert monopsony power with considerable leverage over health suppliers in influencing the prices they set or the amount of services they prescribe.

The fact that private insurance is subject to the problem of moral hazard is less helpful in explaining government provision. Indeed it is questionable whether the government has any advantage in dealing with the problem of moral hazard, since it cannot observe the (hidden) activities of the insured any better than private insurers. One possible form of advantageous government intervention is the taxing and subsidizing of consumption choices that influence the insured’s demand for health care (e.g., a subsidy for health club membership and taxes on smoking). This argument, as noticed by Prescott and Townsend (1984), is based on a presumption that the government can monitor these consumption choices better than the private market; otherwise, private insurers could condition contracts on their clients’ consumption choices and the government would have no advantage over the market. So the potential scope for government provision with moral hazard is seemingly limited.

However, there is a more subtle form of moral hazard that provides a reason for direct government delivery of health care: the time-consistency problem. Imagine that health insurance is provided by the private sector only. Each individual must decide how much insurance to purchase. In a standard insurance situation, risk-averse individuals would fully insure if they could get a fair price. However, in this case they may recognize that if they do not fully insure, a welfaristic government will provide for them should they become ill and uninsured. They have thus an incentive to buy too little insurance and to rely on the government to finance their health care when they become sick. This phenomenon is called the *Samaritan’s dilemma*, and it implies that people will underinvest resources available in the present, knowing that the truly welfaristic government will come to their rescue in the future. The problem is particularly acute for life-threatening diseases where denial of insurance is tantamount to a death sentence for the patient.

A similar time-consistency problem arises on the insurer’s side: insurance companies cannot commit to guaranteeing that the rate charged for insurance will not change as they discover progressively more about the health conditions of their clients. Competition will force insurance companies to update their rate to reflect any new information about an individual’s medical condition. Insurance could then become so expensive for some individuals that they could not afford it. With recent advances in genetic testing and other long-range diagnoses, this problem of the uninsured is likely to grow in the future. With no insurance against unfavorable test results or for the denial of insurance when a policy terminates, those more desperate to get insurance will find it increasingly hard to get it from the private market. The supply- and demand-side time-consistency problems were explicitly recognized in the United States by President Clinton, and
used as a reason to make participation in health insurance compulsory. In response to
the uninsured problem, the government provides a substitute for insurance by directly
funding health care to the poor and long-term sick (Medicaid in the United States).

Another advantage of public provision of insurance is to achieve pooling on a much
larger scale with improved risk-sharing. In including every person in a nationwide
insurance scheme and pooling health insurance with other forms of insurance (unem-
ployment, pension, etc.), public insurance comes closer to the “ideal” optimal insurance
that requires the pooling of all the risks faced by individuals and a single contract
covering them jointly (with a single deductible against all risks).

Both adverse selection and moral hazard have been central in the debates over health
care reform in Europe and North America. Consider, for example, the debate about
medical savings accounts (MSA) in the United States. These were intended to encourage
people to buy insurance with more deductibles and co-payments, thereby reducing the
risk of moral hazard. But critics argued that they will trigger a process of adverse
selection where those less likely to need medical care will avail themselves of MSA.
So those opting for the MSA with larger deductibles might indeed face higher total
medical costs despite the improved incentives (they take more care), simply because
of the self-selection process. Another response to moral hazard problems in the United
States is the mandatory pre-admission referral by peer review organizations before
hospitalization. The increasing popularity of health maintenance organizations can
also be viewed as a response to moral hazard by attracting cost-conscious patients who
wish to lower the cost of insurance. Finally the increasing use of co-payments in many
countries appears to be the effective method of cost containment.

10.8.2 Redistributive Politics

Government provision not only requires mandatory insurance to eliminate the adverse
selection problem, but it also involves socializing insurance. Once insurance is com-
pulsory and financed (at least partly) by taxation, redistributive considerations play
a central role in explaining the extensive public provision of insurance. Government
programs that provide the same amount of public services to all households may still
be redistributive. The amount of redistribution in fact depends on how the programs
are financed and how valuable the services are to individuals with different income
levels.

First, a public health care program offering services that are available to all and
financed by a proportional income tax will redistribute income from the rich to the
poor. If there is not too much diversity of tastes and if consumption of health care
is independent of income, all those with incomes below the average are subsidized
by those above the average. Given the empirical fact that a majority of voters have incomes below the average, a majority of voters would approve of public provision. With diversity of tastes, different individuals prefer different levels of consumption even when incomes are the same and the “one-size-fits-all” public provision may no longer be desirable for the majority. So the trade-off is between income redistribution and preference-matching. However, insofar as consumption of medical care is mostly the responsibility of doctors, reflecting standard medical practices, the preference-matching concern is likely to be negligible.

The second way that redistribution occurs is from the healthy to the sick (or the young to the aged). The tax payments of any particular individual do not depend on that individual’s morbidity. It follows that higher morbidity individuals receive insurance in the public system that is less expensive than the insurance they would get in the private market. So, if a taxpayer has either high morbidity or low income, then his tax price of insurance is lower than the price of private insurance. This taxpayer will vote for public provision. The negative correlation between morbidity and income suggests that the majority below average income are also more likely to be in relatively poor health and so in favor of public insurance.

The third route to redistribution is through opting-out. Universal provision of health care by the government can redistribute welfare from the rich to the poor because the rich refuse the public health care and buy higher quality private health services financed by private insurance. For example, individuals may have to wait to receive treatment in the public system, whereas private treatment is immediate. In opting-out, they lose the value of the taxes they pay toward public insurance, and the resources available for those who remain in the public sector increases as the overall pressure on the system decreases (i.e., the waiting list shortens). So redistribution is taking place because the rich are more likely to use private health care, even though free public health care is available. This redistribution will arise even if everyone contributes the same amount to public health insurance.

Redistribution via health care is also more effective in targeting some needy groups than redistribution in cash. The majority may wish to redistribute from those who inherit good health to those who inherit poor health, which can be thought of as a form of social insurance. If individual health status could be observed, the government would simply redistribute in cash, and there would be no reason for public health insurance. But, because it cannot observe an individual’s poor state of health, providing health care in-kind is a better way to target those individuals. The healthy individuals are less likely to pretend to be unhealthy when health care is provided in-kind than if government were to offer cash compensation to everyone claiming to be in poor health. This is the self-selection benefit of in-kind redistribution.
Information asymmetries have significant implications for the working of competitive markets and the scope for government intervention. Detailed policy recommendations for alleviating these problems also differ depending on whether we face the adverse selection or moral hazard problems. It is crucial to test in different markets the empirical relevance of adverse selection and moral hazard. Such a test is surprisingly simple in the insurance market because both adverse selection and moral hazard predict a positive correlation between the frequency of accident and insurance coverage. This prediction turns out to be very general and to extend to a variety of more general contexts (imperfect competition, multidimensional heterogeneity, etc.).

The key problem is that such correlation can be given two different interpretations depending on the direction of the causality. Under adverse selection high-risk agents, knowing they are more likely to have an accident, self-select by choosing more extensive coverage. Alternatively, under moral hazard agents with more extensive coverage are also less motivated to exert precaution, which may result in higher accident rates. The difference matters a lot for health insurance if we want to assess the impact of co-payments and deductibles on consumption and its welfare implications. Indeed it is a well-documented fact that better coverage is correlated with higher medical expenses. Deductibles and co-payments are likely to be desirable if moral hazard is the main reason, since they reduce overconsumption. But, if adverse selection is the main explanation, then limiting coverage can only reduce the amount of insurance available to risk-averse agents with little welfare gain. Evidence on selection versus incentives can be tested in a number of ways, and we briefly describe some of them.

Manning et al. (1987) separate moral hazard from adverse selection by using a random experiment in which individuals are exogenously allocated to different contracts. Between 1974 and 1977 the Rand Health Insurance Experiment randomly assigned households in the United States to one out of 14 different insurance plans with different co-insurance rates and upper limits on annual out-of-pocket expenses. Compensation was paid in order to guarantee that no household would lose by participating in the experiment. Since individuals were randomly assigned to contracts, any differences in observed behavior can be interpreted as a response to the different incentive structures of the contracts. This experiment has provided some of the most interesting and robust tests of moral hazard and the sensitivity of the consumption of medical services to out-of-pocket expenditures. The demand for medical services was found to respond
significantly to changes in the amount paid by the insuree. The largest decrease in the use of services arises between a free service and a contract involving a 25 percent co-payment rate.

Chiappori et al. (1998) exploit a 1993 change in French regulations to which health insurance companies responded by modifying their coverage rates in a non-uniform way. Some companies increased the level of deductibles, while others did not. They test for moral hazard by using groups of patients belonging to different companies who were confronted with different changes in co-payments and whose use of medical services was observed before and after the change in regulation. They find that the number of home visits by general practitioners significantly decreased for the patients who experienced the increase in co-payments but not for those whose coverage remained constant.

Another interesting study is by Cardon and Hendel (2001) who test for moral hazard versus adverse selection in the US employer-provided health insurance. As argued before, a contract with larger co-payments is likely to involve lower health expenditures, either because of the incentive effect of co-payments or because the high-risk self-select by choosing contracts with lower co-payments. The key identifying argument is that agents do not select their employer on the basis of the health insurance coverage. As a consequence the differences in behavior across employer plans can be attributed to incentive effects. They find strong evidence that incentives matter.

Another way to circumvent the difficulty in empirically distinguishing between adverse selection and moral hazard is to consider the annuity market. The annuity market provides insurance against the risk of outliving accumulated resources. It is more valuable to those who expect to live longer. In this market we can safely expect that individuals will not substantially modify their behavior in response to annuity income (e.g., exerting more effort to extend length of life). It follows that differential mortality rates for annuitants who purchase different types of annuities is convincing evidence that selection occurs. Finkeslein and Poterba (2004) obtain evidence of the following selection patterns: First, those who buy back-loaded annuities (annuities where payments increase over time) are longer-lived (controlling for all observables) than other annuitants, which is consistent with the fact that an annuitant with a longer life expectancy is more likely to be alive in later years when the back-loaded annuity pays out more than the flat annuity. Second, those who buy annuities making payments to the estate are shorter-lived than other annuitants, which is consistent with the fact that the possibility of payments to a short-lived annuitant.
10.10 Conclusions

The efficiency of competitive equilibrium is based on the assumption of symmetric information (or the very strong requirement of perfect information). This chapter has explored some of the consequences of relaxing this assumption. The basic points are that asymmetric information leads to inefficiency and that the inefficiency can take a number of different forms.

Under certain circumstances appropriate government intervention can make everyone better off, even though the government does not have better information than the private sector. The role of the government may also be limited by restrictions on its information. Welfare and public policy implications of the two main forms of information asymmetries are not the same, and it has been an empirical challenge to distinguish between adverse selection and moral hazard. Health insurance is a good illustration of the problems that arise and is characterized by extensive public intervention.

Further Reading

The main contributions on asymmetric information are:


A simple exposition of the moral hazard problem is in:


Applications of the self-selection concept in redistribution programs are:

Chapter 10: Asymmetric Information


Applications to health insurance are:


Empirical testing of adverse selection and moral hazard is in:


Exercises

10.1 What is fair insurance? Why will a risk-averse consumer always buy full insurance when it is fair insurance?

10.2 Should the government allow insurance companies to use genetic testing to better assess the health status of their applicants? Would this genetic testing help or hurt those who are in bad health? Would it exacerbate or mitigate the problem of adverse selection in the health insurance market? Would it increase or decrease the number of people without health insurance? Would it be a good thing?

10.3 Are the following statements true or false?
a. An insurance company must be concerned about the possibility that someone will buy fire insurance on a building and then set fire to it. This is an example of moral hazard.

b. A life insurance company must be concerned about the possibility that the people who buy life insurance may tend to be less healthy than those who do not. This is an example of adverse selection.

c. In a market where there is separating equilibrium, different types of agents make different choices of actions.

d. Moral hazard refers to the effect of an insurance policy on the incentives of individuals to exercise care.

e. Adverse selection refers to how the magnitude of the insurance premium affects the types of individuals that buy insurance.

10.4 Consider each of the following situations involving moral hazard. In each case identify the principal (uninformed party) and the agent (informed party) and explain why there is asymmetric information. How does the action described for each situation mitigate the moral hazard problem?

a. Car insurance companies offer discounts to customers who install anti-theft and speed-monitoring devices in their cars.

b. The International Monetary Fund conditions lending to developing countries upon the adoption of a structural adjustment plan.

c. Firms compensate top executives with options to buy company stock at a given price in the future.

d. Landlords require tenants to pay security deposits.

10.5 Despite the negative stereotype of “women drivers,” women under age of 25 are, on average, noticeably better drivers than men under 25. Consequently insurance companies have been willing to offer young women insurance with a discount of 60 percent over what they charge young men. Similar discrimination applies on the life insurance market given that women are expected to live longer. Sex-based discrimination for auto and life insurance is extremely controversial. Many people have argued that sex-based rates constitute unfair discrimination. After all, some men live longer than some women, and there are some men who are better drivers than some women. In response, several US states have laws mandating “unisex” insurance ratings.

a. What are the likely effects of such interference with the market forces?

b. Should the government allow insurance companies to base life insurance rates on sex? What are the risks for women and for men who were paying very different rates? Who gains and who loses?

c. Should insurance companies be allowed to base automobile insurance rates on sex, age, and marital status? What are the consequences of having some groups paying much less than they would if rates were based on actuarial differences in accident rates across sexes and ages?

10.6 Consider a community of individuals that have different probabilities of falling ill. Individuals of type $H$ have a probability of falling ill of $p^H = 0.6$. These individuals form $\frac{3}{4}$ of the
population. Individual of type \( L \), the remaining \( \frac{1}{4} \) of the population, have a probability of falling ill of \( p^L = 0.2 \). Any individual who is ill suffers an income loss of 200.

a. Assume there is symmetric information. What is the actuarially fair premium for the each group to insure against illness? What level of coverage will each individual request if offered insurance at the actuarially fair premium?

b. Assume that there is asymmetric information and firms cannot distinguish between the types. The same insurance contract must be offered to all individuals. If the contract is to earn an expected profit of zero, what premium must be charged?

c. Is it possible for a firm to offer an alternative contract that is profitable when the contract in part b is available?

10.7 Discuss the argument that paying for human blood has the effect of lowering its average quality because people who are driven by the profit motive to provide blood are more likely to be drug addicts, alcoholics, and have serious infectious diseases than are voluntary donors.

10.8 In California many insurance companies charge different rates depending on what part of the city you live in. Their rationale is that risk factors like theft, vandalism, and traffic congestion vary greatly from one place to the other. The result is that people who live close to each other, but in adjacent zip codes, may end up paying very different insurance premia.

a. What would happen to an insurance company that decided to sell insurance at the same price to all drivers with the same driving records no matter what part of the city they live in?

b. What would happen if the government decides to outlaw geographic rate differentials, given that the government cannot force private insurance companies to provide insurance against their will?

10.9 Georgie has a labor income equal to 100. With probability \( p = 0.25 \), she will have an accident during the year and suffer an income loss equal of 60. With probability \( p = 0.75 \), she will have no accident. Georgie decides to purchase insurance contract to cover the loss of income if she has an accident.

a. What is the actuarially fair premium for the insurance? What level of coverage will Georgie buy?

Georgie loves extreme sports and fast driving. If insured, she will drive faster and try more extreme activities. This increases her risk of accident. Let the risk be given by \( p = 1 - \frac{L}{100} \), where \( L \) is the uninsured income loss.

b. Will the contract in part a break even? If the insurance company knows Georgie’s behavior (the function \( p = 1 - \frac{L}{100} \)), what contract will it offer? If the insurance company does not know her behavior, what contract will break even?

10.10 The European Union has made discrimination between males and females illegal for insurance contracts. Explain the effects of this decision. Will it benefit anyone?

10.11 The local government has hired someone to undertake a public project. If the project fails, it will lose $20,000. If it succeeds, the project will earn $100,000. The employee can choose to “work” or to “shirk.” If she shirks, the project will fail for sure. If she works, the project will succeed half of the time but will still fail half of the time. The employee’s utility is $10,000 lower if she works than if she shirks. In addition the employee could earn $10,000 in another job (where she would shirk). The government is choosing whether to pay the employee a
Part III: Departures from Efficiency

flat wage of $20,000 (no matter how the project turns out) or performance-related pay under which the employee earns 0 if the project fails and $40,000 if it succeeds.

a. Assuming that both parties are risk neutral, which compensation scheme should the government use?

b. Do you see any problem with the performance-related pay scheme when the employee is risk averse?

10.12 Use the signaling model presented in section 10.6 to construct an example in which a government unaware of workers’ productivities can improve the welfare of everyone compared to the (best) separating equilibrium by means of a cross-subsidization policy but not by banning signaling.

10.13 A firm hires two kinds of workers, alphas and betas. One can’t tell a beta from an alpha by looking at her, but an alpha will produce $3,000 worth of output per month and a beta will produce $2,500 worth of output in a month. The firm decides to distinguish alphas from betas by making them pass an examination. For each question that they get right on the exam, alphas have to spend half an hour studying and betas have to spend one hour. A worker will be paid $3,000 if she gets at least 40 answers right and $2,500 otherwise. For either type, an hour of studying is as bad as giving up $20 income. What is the equilibrium of this scheme?

10.14 Consider a loan market to finance investment projects. All projects cost 1. Any project is either good (with probability \( \rho \)) or bad (with probability \( 1 - \rho \)). Only investors know whether their project is good or bad. A good project yields profits of \( \pi > 0 \) with probability \( P_g \) and no profit with probability \( 1 - P_g \). A bad project makes profits of \( \pi \) with a lower probability \( P_b \) (with \( P_b < P_g \)) and no profit with a higher probability \( 1 - P_b \). Banks are competitive and risk neutral, which implies that banks offer loan contracts making expected profits of zero. A loan contract specifies a repayment \( R \) that is supposed to be repaid to the bank only if the project makes profit; otherwise, the investor defaults on her loan contract. The opportunity cost of funds to the bank is \( r > 0 \). Suppose

\[
P_g - (1 + r) > 0 > P_b - (1 + r).
\]

a. Find the equilibrium level of \( R \) and the set of projects financed. How does this depend on \( P_g, P_b, \rho, \pi, \) and \( r \)?

b. Now suppose that the investor can signal the quality of her project by self-financing a fraction of the project. The opportunity cost of funds to the investor is \( s \) (with \( s > r \) implying a costly signal). Describe the investor’s payoff as a function of the type of her project, the loan repayment \( R \) and her self-financing rate. Derive the indifference curve for each type of investor in the \((s, R)\) space. Show that the single-crossing property holds.

c. What is the best separating equilibrium of the signaling game where the investor first chooses \( s \) and banks then respond by a repayment schedule \( R(s) \)? How does the self-financing rate of good projects change with small changes of \( P_g, P_b, \rho, \pi, \) and \( r \)?

d. Compare this (best) separating equilibrium with part a.

10.15 (Akerlof) Consider the following market for used cars. There are many sellers of used cars. Each seller has exactly one used car to sell and is characterized by the quality of the used car he wishes to sell. Let \( \theta, 0 \leq \theta \leq 1 \), index the quality of a used car, and suppose that \( \theta \)
is uniformly distributed on the interval $[0, 1]$. If a seller of type $\theta$ sells his car at price $p$, his utility is $u_s(p, \theta)$. With no sale his utility is 0. Buyers receive utility $\theta - p$ if they buy a car of quality $\theta$ at price $p$, and receive utility 0 if they do not purchase. The quality of the car is only known to sellers, and there are enough cars to supply all potential buyers.

a. Explain why the competitive equilibrium outcome under asymmetric information requires that the average quality of cars that are put for sale conditional on price is just equal to price, $E(\theta | p) = p$. Describe the equilibrium outcome in words. In particular, describe which cars are traded in equilibrium.

b. Show that if $u_s(p, \theta) = p - \theta^2$, then every price $0 < p \leq \frac{1}{2}$ is an equilibrium price.

c. Find the equilibrium price when $u_s(p, \theta) = p - \sqrt{\theta}$.

d. How many equilibrium prices are there when $u_s(p, \theta) = p - \theta^3$?

e. Which (if any) of the preceding outcomes are Pareto-efficient? Describe Pareto improvements whenever possible.

10.16 It is known that some fraction $d$ of all new cars are defective. Defective cars cannot be identified as such except by those who own them. Each consumer is risk neutral and values a nondefective car at $16,000$. New cars sell for $14,000$ each, and used ones for $2,000$. If cars do not depreciate physically with use, what is the proportion $d$ of defective new cars?

10.17 In the preceding question, assume that new cars sell for $18,000$ and used cars sell for $2,000$. If there is no depreciation and risk-neutral consumers know that 20 percent of all new cars are defective, how much do the consumers value a nondefective car?

10.18 There are two types of jobs in the economy, good and bad, and two types of workers, qualified and unqualified. The population consists of 60 percent qualified and 40 percent unqualified. In a bad job, either type of worker produces the same 10 units of output. In a good job, a qualified worker produces 100 and an unqualified worker produces 0. There are numerous job openings of each type, and companies must pay for each type of job what they expect the appointee to produce. The worker’s type is unknown before hiring, but the qualified workers can signal their type (e.g., by getting educated or some other means). The cost of signaling to level $s$ for a qualified worker is $s^2$ and for an unqualified worker is $s^2$. The signaling costs are measured in the same units as output, and $s$ must be an integer (e.g., number of years of education).

a. What is the minimum level of $s$ that will achieve separation?

b. Suppose that the signal is no longer available. Which kinds of job will be filled by which types of workers, and at what wages? Who gains and who loses?

10.19 The government can help those people most in need by either giving them cash or providing free meals. What is the argument for giving cash? What kind of argument based on asymmetric information could support the claim that free meals (an in-kind transfer) are better than the cash handout? Can such an argument apply to free education?

10.20 Explain why an automaker’s willingness to offer a resale guarantee for its cars may serve as a signal of their quality.

10.21 A competitive market for annuity contracts has a number of risk-neutral providers. There are two time periods. There are also two types of individual ($i = L, H$) that vary only in their
probabilities, \( \rho_H \) and \( \rho_L \), of surviving (life expectancy) to live in the second period. Assume \( \rho_H > \rho_L \). Unlike the firms, the individuals know their types ex ante. The distribution of types is commonly known where \( \gamma \in (0, 1) \) denotes the proportion of the \( L \)-type in the population.

a. Let \( a_i \) denote the value of annuities bought by individual \( i \) and \( q_i \) the return on annuities. Moreover the interest rate, \( r \), satisfies \( r = 0 \) (there is no discounting) and the income of the first period is \( w \) for both types. Derive the intertemporal budget of an individual.

b. The utility function is given by

\[
EU_i = -\left(c_{i1}^{\gamma-1} - \rho_i \left(c_{i2}^{\gamma-1}\right)ight), \quad i = L, H.
\]

Show that the indifference curves of both types intersect exactly once. Explain this intuitively.

c. Suppose \( w = 10, \rho_H = \frac{1}{2} \) and \( \rho_L = \frac{1}{4} \). Determine the first-best solution, meaning the solution in the case where the firms are able to distinguish the types. Which type would have an incentive to deviate?

d. Suppose that there exists a separating equilibrium. Derive the optimal solution \((a_H, c_{H1}, c_{H2})\). Give an intuitive explanation for your result.

e. Assume \( w = 10, \rho_H = \frac{1}{2} \) and \( \rho_L = \frac{1}{4} \). Determine \( c_{L1}^{\gamma} \) and \( c_{L2}^{\gamma} \) in the separating equilibrium.

How do \( c_{L1}^{\gamma} \) and \( c_{L2}^{\gamma} \) change if \( \rho_H \) marginally rises?

f. Explain intuitively how the separating equilibrium depends on \( \rho \). Show graphically how you could find the critical \( \gamma \) for which the separating equilibrium of parts d and e does not exist anymore.

10.22 The design of the health care system involves issues of information at several points. The potential users (patients) are better informed about their own state of health and lifestyle than insurance companies. The health providers (doctors and hospitals) know more about what patients need than do either the patients themselves or the insurance companies. Providers also know more about their own skills and efforts. Insurance companies have statistical information about outcomes of treatments and surgical procedures from past records. The drug companies know more about the efficacy of drugs than do others. As is usual, the parties have different interests, so they do not have a natural inclination to share their information fully or accurately with others.

a. From this perspective, consider the relative merits of the following payments schemes:

i. A fee for service versus capitation fees to doctors.

ii. Comprehensive premiums per year versus payment for each visit for patients.

b. Which payments schemes are likely to be most beneficial to the patients and which to the providers?

c. What are the relative merits of private insurance compared to coverage of costs from general tax revenues?
IV POLITICAL ECONOMY
11 Voting

11.1 Introduction

Voting is the most commonly employed method of resolving a diversity of views or eliciting expressions of preference. It is used to determine the outcome of elections from local to supra-national level. Within organizations, voting determines who is elected to committees, and it governs the decision-making of those committees. Voting is a universal tool that is encountered in all spheres of life. The prevalence of voting, its use in electing governments, and its use by those governments elected to reach decisions, is the basis for the considerable interest in the properties of voting.

The natural question to ask of voting is whether it is a good method of making decisions. There are two major properties to look for in a good method. First is the success or failure of the method in achieving a clear-cut decision. Second is the issue of whether voting always produces an outcome that is efficient. Voting would be of limited value if it frequently left the choice of outcome unresolved or led to a choice that was clearly inferior to other alternatives. Whether voting satisfies these properties is shown to be somewhat dependent on the precise method of voting adopted. Ordinary majority voting is very familiar, but it is only one among a number of ways of voting. Several of these methods of voting will be introduced and analyzed alongside the standard form of majority voting.

11.2 Stability

Voting is an example of collective choice—the process by which a group (or collective) reaches a decision. A major issue of collective choice is stability. By stability, we mean the tendency of the decision-making process to eventually reach a settled conclusion, and not to keep jumping around between alternatives. We begin this chapter by a simple illustration of the central fact that when you have a large group of people, with conflicting preferences, stability in matching preferences is not guaranteed.

The example involves three married couples living as neighbors on a remote island. Initially the couples are comprised of Alil and Alice, Bob and Beth, and Carl and Carol. We assume that each husband has his own preference list of the women as potential wives and each wife has a list of preferences among husbands, each ranking partners from best to worst. We also make the assumption that the top preference for
any given wife may or may not be her own husband, and similarly for the men. To avoid untenable frustrations developing, the island society introduces a rule that if two people prefer each other to their existing partners, they can re-form as a new couple. For example, if Alil prefers Beth to his own wife, Alice, and Beth prefers Alil to her own husband, Bob, then Alil can join Beth, leaving Bob and Alice to console each other. (It is forbidden on this island to live alone or to form a couple with someone of the same sex.)

Now consider the lists of preferences for the participants given in table 11.1. It follows from these preferences that Beth will join Alil (she prefers him to Bob, and Alil prefers her to Alice), then she will continue her ascension to Carl (who prefers her to Carol, while he is her first choice). By then Alice has been left with Bob, her worst choice, so she will go to Carl, and finally back to Alil, her favorite. In every case the leaving male is also improving his own position. But now the end result is that this round of spouse trading leaves us back exactly with the initial situation, so the cycle can begin again, and go on forever. The attempt to prevent frustration has lead to an unstable society.

In the example a situation was shown where stability could not be achieved. One argument for wanting stability is that it describes a settled outcome in which a final decision has been reached without further renegotiation. If the process of changing position is costly, as it would be in our example, then stability would be beneficial. It can also be argued that there are circumstances where stability is not desirable. Consider the extreme case where each man is married to his first choice but each husband is at the bottom of his wife’s preference list. This would be a stable outcome because no man would be interested in switching and no wife would be able to switch either because she would find no man who prefers her. So the society is stable but not necessarily at a desirable state, since the stability is forcing some of the participants to remain with unwanted choices.
11.3 Impossibility

Determining the preferences of an individual is just a matter of accepting that an individual’s judgment cannot be open to dispute. In contrast, determining the preferences of a group of people is not a simple matter. And that is what social choice theory (including voting as one particular method) is all about. Social choice takes a given set of individual preferences and tries to aggregate them into a social preference.

The central result of the theory of social choice, Arrow’s Impossibility Theorem, says that there is no way to devise a collective decision-making process that satisfies a few commonsense requirements and works in all circumstances. If there are only two options, majority voting works just fine, but with more than two we can get into trouble. Despite all the talk about the “will of the people,” it is not easy—in fact the theorem proves it impossible—to always determine what that will is. This is the remarkable fact of Arrow’s Impossibility Theorem.

Before presenting the theorem, a taste of it can be obtained with the simplest case of three voters with the (conflicting) rankings over three options shown in table 11.2. Every voter has transitive preferences over the three options. For example, voter 1 prefers $a$ to $b$ to $c$, and therefore $a$ to $c$. As individuals, the voters are entirely self-consistent in their preferences.

Now suppose that we use majority rule to select one of these options. We see that two out of three voters prefer $a$ to $b$, while two out of three prefer $b$ to $c$, and two out of three prefer $c$ to $a$. At the collective level there is a cycle in preference and no decision is possible. We say that such collective preferences are intransitive, meaning that the preference for $a$ over $b$ and for $b$ over $c$ does not imply that $a$ is preferred to $c$. As the example shows, intransitivity of group preferences can arise even when individual preferences are transitive. This generation of social intransitivity from individual transitivity is called the Condorcet paradox.

Table 11.2  
Condorcet paradox

<table>
<thead>
<tr>
<th>Voter 1</th>
<th>Voter 2</th>
<th>Voter 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>$c$</td>
<td>$b$</td>
</tr>
<tr>
<td>$b$</td>
<td>$a$</td>
<td>$c$</td>
</tr>
<tr>
<td>$c$</td>
<td>$b$</td>
<td>$a$</td>
</tr>
</tbody>
</table>
The general problem addressed by Arrow in 1951 was to seek a way of aggregating individual rankings over options into a collective ranking. In doing so, difficulties such as the Condorcet paradox had to be avoided. Arrow’s approach was to start from a set of requirements that a collective ranking must satisfy and then consider if any ranking could be found that met them all. These conditions are now listed and explained.

**Condition I** (Independence of irrelevant alternatives) Adding new options should not affect the initial ranking of the old options, so the collective ranking over the old options should be unchanged.

For example, suppose that a group prefers option A to option C, and the new option B is introduced. Wherever it fits into each individual’s ranking, condition I requires that the group preference should not switch to C over A. They may like or dislike the new option B, but their relative preferences for other options should not change. If this condition were not imposed on collective decision-making, any decision could be invalidated by bringing in new irrelevant (inferior) options. Since it is always possible to add new options, no decision would ever be made.

**Condition N** (Nondictatorship) The collective preference should not be determined by the preferences of one individual.

This is the weakest equity requirement. Having a dictatorship as a collective decision process may solve transitivity problems, but it is manifestly unfair to the other individuals. Any conception of democracy aspires to some form of equity among all the voters.

**Condition P** (Pareto criterion) If everybody agrees on the ranking of all the possible options, so should the group; the collective ranking should coincide with the common individual ranking.

The Pareto condition requires that unanimity should prevail where it arises. It is hardly possible to argue with this condition.

**Condition U** (Unrestricted domain) The collective choice method should accommodate any possible individual ranking of options.
This is the requirement that the collective choice method should work in all circumstances so that the method is not constructed in such a way as to rule out (arbitrarily), or fail to work on, some possible individual rankings of alternatives.

**Condition T** (Transitivity)  If the group prefers $A$ to $B$ and $B$ to $C$, then the group cannot prefer $C$ to $A$.

This is merely a consistency requirement that ensures that a choice can always be made from any set of alternatives. The Condorcet paradox shows that majority voting fails to meet this condition and can lead to cycles in collective preference.

That is it, and one can hardly disagree with any of these requirements. Each one seems highly reasonable taken individually. Yet the remarkable result that Arrow discovered is that there is no way to devise a collective choice method that satisfies them all simultaneously.

**Theorem 11.1** (Arrow’s Impossibility Theorem)  When choosing among more than two options, there exists no collective decision-making process that satisfies the conditions I, N, P, U, T.

The proof is slightly, rather than very, complicated and is quite formal. We will not reproduce it here. The intuition underlying the proof is clear enough and follows this reasoning:

1. The unrestricted domain condition allows for preferences such that no option is unanimously preferred.
2. The independence of irrelevant alternatives forces the social ranking over any two options to be based exclusively on the individual preferences over those two options only.
3. From the Condorcet paradox we know that a cycle can emerge from three successive pairwise comparisons.
4. The transitivity requirement forces a choice among the three options.
5. The only method for deciding must give one individual all the power, thus contradicting the nondictatorship requirement.

The implication of Arrow’s Impossibility Theorem is that any search for a “perfect” method of collective decision-making is doomed to failure. Whatever process is devised, a situation can be constructed in which it will fail to deliver an outcome that
satisfies one or more of the conditions I, N, P, U, T. As a consequence all collective
decision-making must make the most of imperfect decision rules.

11.4 Majority Rule

In any situation involving only two options, majority rule simply requires that the
option with the majority of votes is chosen. Unless unanimity is possible, asking that
the few give way to the many is a very natural alternative to dictatorship. The process
of majority voting is now placed into context and its implications determined.

11.4.1 May’s Theorem

Nondictatorship is a very weak interpretation of the principles of democracy. A widely
held view is that democracy should treat all the voters in the same way. This symmetry
requirement is called Anonymity. It requires that permuting the names of any two
individuals does not change the group preference. Thus Anonymity implies that there
cannot be any dictator. Another natural symmetry requirement is that the collective
decision-making process should treat all possible options alike. No apparent bias in
favor of one option should be introduced. This symmetric treatment of the various
options is called Neutrality.

Now a fundamental result due to May is that majority rule is the obvious way
to implement these principles of democracy (Anonymity and Neutrality) in social
decision-making when only two options are considered at a time. The theorem as-
serts that majority rule is the unique way of doing so if the conditions of Decisiveness
(i.e., the social decision rule must pick a winner) and Positive Responsiveness (i.e.,
increasing the vote for the winning option should not lead to the declaration of another
option the winner) are also imposed.

Theorem 11.2 (May’s theorem) When choosing among only two options, there is
only one collective decision-making process that satisfies the requirements of Ano-
ymity, Neutrality, Decisiveness, and Positive Responsiveness. This process is majority
rule.

Simple majority rule is the best social choice procedure if we consider only two
options at a time. Doing so is not at all unusual in the real world. For instance, when
a vote is called in a legislative assembly, there are usually only two possible options:
to approve or to reject some specific proposal that is on the floor. Also, in a situation of two-party political competition, voters again face a binary choice. Therefore interest in other procedures arises only when there are more than two options to consider.

11.4.2 Condorcet Winner

When there are only two options, majority rule is a simple and compelling method for social choice. When there are more than two options to be considered at any time, we can still apply the principle of majority voting by using binary agendas that allow us to reduce the problem of choosing among many options to a sequence of votes over two alternatives at a time.

For example, one simple binary agenda for choosing among the three options \( \{a, b, c\} \) in the Condorcet paradox is as follows. First, there is a vote on \( a \) against \( b \). Then, the winner of this first vote is opposed to \( c \). The winner of this second vote is the chosen option. The most famous pairwise voting method is the Condorcet method. This consists of a complete round-robin of majority votes, opposing each option against all of the others. The option that defeats all others in pairwise majority voting is called a Condorcet winner, after Condorcet suggested that such an option should be declared the winner. That is, using the symbol \( \succ \) to denote majority preference, a Condorcet winner is an option \( x \) such that \( x \succ y \) for every other option \( y \) in the set of possible options \( X \).

The problem is that the existence of a Condorcet winner requires very special configurations of individual preferences. For instance, with the preferences given in the Condorcet paradox, there is no Condorcet winner. So a natural question to ask is under what conditions a Condorcet winner does exist.

11.4.3 Median Voter Theorems

When the policy space is one-dimensional, sufficient (but not necessary) conditions for the existence of a Condorcet winner are given by the Median Voter Theorems. One version of these theorems refers to single-peaked preferences, while the other version refers to single-crossing preferences. The two conditions of single-peaked and single-crossing preferences are logically independent, but both conditions give the same conclusion that the median position is a Condorcet winner.

For an example of single-peaked preferences, consider figure 11.1 depicting a population of consumers who are located at equally spaced positions along a straight road.
Figure 11.1
Location of households

A bus stop is to be located somewhere on this road. It is assumed that all consumers prefer the stop to be located as close as possible to their own homes. If the location of the bus stop is to be determined by majority voting (taking pairwise comparisons again), which location will be chosen?

When there is an odd number of houseowners, the answer to this question is clear-cut. Given any pair of alternatives, a household will vote for that which is closest to its location. The location that is the closest choice for the largest number of voters will receive a majority of votes.

Now consider a voting process in which votes are taken over every possible pair of alternatives. This is very much in the form of a thought process rather than a practical suggestion, since there must be many rounds of voting and the process will rapidly becomes impractical if there are many alternatives. Putting this difficulty aside, it can easily be seen that this process will lead to the central outcome being the chosen alternative. This location wins all votes and is the Condorcet winner. Expressed differently, the location preferred by the median voter (i.e., the voter in the center) will be chosen. At least half the population will always vote for this.

This result is the basis of the Median Voter Theorem. When there is an even number of voters, there is no median voter but the two locations closest to the center will both beat any other locations in pairwise comparisons. They will tie when they are directly compared. The chosen location must therefore lie somewhere between them.

The essential feature that lies behind the reasoning of the example is that each consumer has single-peaked preferences, and that the decision is one-dimensional. Preferences are termed single-peaked when there is a single preferred option. Figure 11.2b illustrates preferences that satisfy this condition, whereas those in figure 11.2a are not single-peaked. In the bus stop example each consumer most prefers a location close to home and ranks the others according to the how close they are to the ideal. Such preference looks exactly like those in figure 11.2b. The choice variable is one-dimensional because it relates to locations along a line.

The first general form of the Median Voter Theorem can be stated as follows:
Theorem 11.3 (Median Voter Theorem I: Single-peaked version) Suppose that there is an odd number of voters and that the policy space is one-dimensional (so that the options can be put in a transitive order). If the voters have single-peaked preferences, then the median of the distribution of voters’ preferred options is a Condorcet winner.

The idea of median voting has also been applied to the analysis of politics. Instead of considering the line in figure 11.1 as a geographical identity, view it as a representation of the political spectrum running from left to right. The houseowners then becomes voters and their locations represent political preferences. Let there be two parties who can choose their location on the line. A location in this sense represents the manifesto on which they stand. Where will the parties choose to locate? Assume as above that the voters always vote for the party nearest to their location. Now fix the location of one party at any point other than the center and consider the choice of the other. Clearly, if the second party locates next to the first party on the side containing more than half the electorate, it will win a majority of the vote. Realizing this, the first party would not be content with its location. It follows that the only possible equilibrium set of locations for the parties is to be side by side at the center of the political spectrum.

This agglomeration at the center is called *Hotelling’s principle of minimal differentiation* and has been influential in political modeling. The reasoning underlying it can be observed in the move of the Democrats in the United States and the Labor party in the
United Kingdom to the right in order to crowd out the Republicans and Conservatives respectively. The result also shows how ideas developed in economics can have useful applications elsewhere.

Although a powerful result, the Median Voter Theorem does have significant drawbacks. The first is that the literal application of the theorem requires that there be an odd number of voters. This condition ensures that there is a majority in favor of the median. When there is an even number of voters, there will be a tie in voting over all locations between the two central voters. The theorem is then silent on which of these locations will eventually be chosen. In this case, though, there is a median tendency. The second, and most significant drawback, is that the Median Voter Theorem is applicable only when the decision over which voting is taking place has a single dimension. This point will be investigated in the next section. Before doing that, let us consider the single-crossing version of the Median Voter Theorem.

The single-crossing version of the Median Voter Theorem assumes not only that the policy space is transitively ordered, say from left to right (and thus one-dimensional), but also that the voters can be transitively ordered, say from left to right in the political spectrum. The interpretation is that voters at the left prefer left options more than voters at the right. This second assumption is called the single-crossing property of preferences. Formally,

**Definition 11.1** (Single-crossing property) For any two voters $i$ and $j$ such that $i < j$ (voter $i$ is to the left of voter $j$), and for any two options $x$ and $y$ such that $x < y$ ($x$ is to the left of $y$),

(i) if $u^i(x) > u^i(y)$, then $u^j(x) > u^j(y)$, and (ii) if $u^j(y) > u^j(x)$, then $u^i(y) > u^i(x)$.

The median voter is characterized as the median individual on the left to right ordering of voters, so that half the voters are to the left of the median voter and the other half is to the right. Therefore, according to the single-crossing property, for any two options $x$ and $y$, with $x < y$, if the median voter prefers $x$, then all the voters to the left also prefer $x$, and if the median voter prefers $y$, then all the voters to the right also prefer $y$. So there is always a majority of voters who agree with the median voter, and the option preferred by the median voter is a Condorcet winner.

**Theorem 11.4** (Median Voter Theorem II: Single-crossing version) Suppose that there is an odd number of voters and that the policy space is one-dimensional (so that the options can be put in a transitive order). If the preferences of the set of voters satisfy the single-crossing property, then the preferred option of the median voter is a Condorcet winner.
Single-crossing and single-peakedness are different conditions on preferences. But both give us the same result that the median voter’s preferred option is a Condorcet winner. However, there is a subtle difference. With the single-peakedness property, we refer to the median of the voters’ preferred options, but with the single-crossing property, we refer to the preferred option of the median voter. Notice that single-crossing and single-peakedness are logically independent as the example in figure 11.3 illustrates. The options are ranked left to right along the horizontal axis, and the individual 3 is to the left of 2 who is to the left of 1. It can be checked that single-crossing holds for any pair of options but single-peakedness does not hold for individual 2. So one property may fail to hold when the other is satisfied.

An attractive aspect of the Median Voter Theorem is that it does not depend on the intensity of preferences, and thus no voters have an incentive to misrepresent their preferences. This implies that honest, or sincere, voting is the best strategy for everyone. Indeed, for a voter to the left of the median, misrepresenting preference more to the left does not change the median and therefore the final outcome, whereas misrepresenting preferences more to the right either does nothing or moves the final outcome further away from his preferred outcome. Following the same reasoning, a voter to the right of the median has no incentive to misrepresent his preferences either way. Last, the median gets his most-preferred outcome and thus cannot benefit from misrepresenting his preferences.

Having seen how the Median Voter Theorem leads to a clearly predicted outcome, we can now inquire whether this outcome is efficient. The chosen outcome reflects the
preferences of the median voter, so the efficient choice will only be made if this is the most preferred alternative for the median voter. Obviously there is no reason why this should be the case. Therefore the Median Voter Theorem will not, in general, produce an efficient choice. In addition, without knowing the precise details, it is not possible to predict whether majority voting will lead, via the Median Voter Theorem, to a choice that lies to the left or to the right of the efficient choice.

A further problem with the Median Voter Theorem is its limited applicability. It always works when policy choices can be reduced to one dimension but only works in restricted circumstances where there is more than one dimension. We now demonstrate this point.

11.4.4 Multidimensional Voting

The problem of choosing the location of the bus stop was one-dimensional. A second dimension could be introduced into this example by extending the vote to determine both the location of the bus stop and the time at which the bus is to arrive. The important observation for majority voting is that when this extension is made there is no longer any implication that single-peaked preferences will lead to a transitive ranking of alternatives.

This finding can be illustrated by considering the indifference curves of a consumer over the two-dimensional space of location and time. To do this, consider location as the horizontal axis and time as the vertical axis with the origin at the far left of the street and midnight respectively. The meaning of single-peaked preferences in this situation is that a consumer has a most-preferred location and any move in a straight line away from this must lead to a continuous decrease in utility. This is illustrated in figure 11.4 where $x_i$ denotes the most preferred location of $i$, and the oval around this point is one of the consumer’s indifference curves.

Using this machinery, it is now possible to show that the Median Voter Theorem does not apply and majority voting fails to generate a transitive outcome. The three voters, denoted 1, 2, and 3, have preferred locations $x_1$, $x_2$, and $x_3$. Assume that the voting is to decide which of these three locations is to be chosen (this is not necessary for the argument, as will become clear, but it does simplify it). The rankings of the three consumers of these alternatives in table 11.3 are consistent with the preferences represented by the ovals in figure 11.4. Contrasting these to table 11.2, we can see immediately that these are exactly the rankings that generate an intransitive social ordering through majority voting. Consequently, even though preferences are single-peaked,
Figure 11.4
Single-peakedness in multidimensions

Table 11.3
Rankings

<table>
<thead>
<tr>
<th>Voter 1</th>
<th>Voter 2</th>
<th>Voter 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>$x_2$</td>
<td>$x_3$</td>
</tr>
<tr>
<td>$x_2$</td>
<td>$x_3$</td>
<td>$x_1$</td>
</tr>
<tr>
<td>$x_3$</td>
<td>$x_1$</td>
<td>$x_2$</td>
</tr>
</tbody>
</table>

the social ordering is intransitive and the Median Voter Theorem fails. Hence the theorem does not extend beyond one-dimensional choice problems. It is worth noting that if voting were carried out on each dimension separately, then voter 1 would be the median voter on the location dimension and voter 2 would be the median voter on the time dimension. So the time voting outcome will be given by the projection of $x_2$ on the vertical axis and the location voting outcome will be given by the projection of $x_1$ on the horizontal axis. The problem with this item-by-item voting is that it can generate, for some preferences, an inefficient voting outcome. This is the case when the chosen point lies outside the triangle formed by the voters’ blisspoints $x_1$, $x_2$, and $x_3$. 
11.4.5 Agenda Manipulation

In a situation where there is no Condorcet winner, the door is opened to agenda manipulation. This is because changing the agenda, meaning the order in which the votes over pairs of alternatives are taken, can change the voting outcome. Thus the agenda-setter may have substantial power to influence the voting outcome. To determine the degree of the agenda-setter’s power, we must find the set of outcomes that can be achieved through agenda manipulation.

To see how agenda-setting can be effective, suppose that there are three voters with preferences as in the Condorcet paradox (described in table 11.2). Then there is a majority (voters 1 and 2) who prefer $a$ over $b$, there is a majority (voters 2 and 3) who prefer $c$ over $a$, and there is a majority (voters 1 and 3) who prefer $b$ over $c$. Given these voters’ preferences, what will be the outcome of different binary agendas? The answer is that when voters vote sincerely, then it is possible to set the agenda so that any of the three options can be the ultimate winner. For example, to obtain option $a$ as the final outcome, it suffices to first oppose $b$ against $c$ (knowing that $b$ will defeat $c$) and then at the second stage to oppose the winner $b$ against $a$ (knowing that $a$ will defeat $b$). Similarly, to get $b$ as the final outcome, it suffices to oppose $a$ against $c$ at the first stage (given that $c$ will defeat $a$) and then the winner $c$ against $b$ (given that $b$ will defeat $c$). These observations show how the choice of agenda can affect the outcome.

This reasoning is based on the assumption that voters vote sincerely. However, the voters may respond to agenda manipulation by misrepresenting their preferences. That is, they may vote strategically. Voters can choose to vote for options that are not actually their most-preferred options if they believe that such behavior in the earlier ballots can affect the final outcome in their favor. For example, if we first oppose $b$ against $c$, then voter 2 may vote for $c$ rather than $b$. This ensures that $c$ goes on to oppose $a$. Option $c$ will win, an outcome preferred by voter 2 to the victory for $a$ that emerges with sincere voting. So voters may not vote for their preferred option in order to prevent their worst option from winning. The question is then how strategic voting affects the set of options that could be achieved by agenda manipulation. Such outcomes are called sophisticated outcomes of binary agendas because voters anticipate what the ultimate result will be, for a given agenda, and vote optimally in earlier stages.

A remarkable result, due to Miller (1980), is that strategic voting (relative to sincere voting) does not alter the set of outcomes that can be achieved by agenda manipulation when the agenda-setter can design any binary agendas, provided only that every option must be included in the agenda. Miller called the set that can be achieved the top cycle.
When there exists a Condorcet winner, the top cycle reduces to that single option. With preferences as in the Condorcet paradox, the top cycle contains all three options \(\{a, b, c\}\). For example, option \(b\) can be obtained by the following agenda (different from the agenda under sincere voting): at the first stage, \(a\) is opposed to \(b\), then the winner is opposed to \(c\). This binary agenda is represented in figure 11.5.

The agenda begins at the top, and at each stage the voters must vote with the effect of moving down the agenda tree along the branch that will defeat the other with a sophisticated majority vote. To resolve this binary agenda, sophisticated voters must anticipate the outcome of the second stage and vote optimally in the first stage. Either the second stage involves \(c\) against \(a\), and thus \(c\) will beat \(a\), or the second stage involves \(c\) against \(b\), and thus \(b\) will beat \(c\) (as voters will vote sincerely in this last stage). So the voters should anticipate that in the first stage voting for \(a\) will in fact lead to the ultimate outcome \(c\), whereas voting for \(b\) will lead to the ultimate outcome \(b\) (as displayed in parentheses). So, in voting for \(a\) in the first stage, they vote in effect for \(c\), whereas voting for \(b\) in the first stage effectively leads to the choice of \(b\) as the ultimate outcome. Because \(b\) is preferred by a majority to \(c\), it follows that a majority of voters should vote for \(b\) at the first stage (even though a majority prefers \(a\) over \(b\)).

The problem with the top cycle is that it can contain options that are Pareto-dominated. To see this, suppose that the preferences are as in the Condorcet paradox, and add a fourth alternative \(d\) that falls just below \(c\) in every individual’s preference. The resulting rankings are given in table 11.4. Note that there is a cycle: two out of three prefer \(b\) to \(c\), while two out of three prefer \(a\) to \(b\), and two out of three prefer \(d\) to \(a\), and last all prefer \(c\) to \(d\), making it a full circle. So \(d\) is included in the top cycle, even though \(d\) is Pareto-dominated by \(c\).
The situation can be even worse than this. An important theorem, due to McKelvey, says that if there is no Condorcet winner, then the top cycle is very large and can even coincide with the full set of alternatives. There are two implications of this result. First, the agenda-setter can bring about any possible option as the ultimate voting outcome. So the power of the agenda-setter may be substantial. Such dependence implies that the outcome chosen by majority rule cannot be characterized, in general, as the expression of the voters’ will. Second, the existence of a voting cycle makes the voting outcome arbitrary and unpredictable, with very little normative appeal.

We know that the existence of a Condorcet winner requires very special conditions on voters’ preferences. In general, with preferences that do not have the single-peakedness or single-crossing properties on a simple one-dimensional issue space, we should not generally expect that a Condorcet winner exists. For example, Fishburn (1973) has shown that when voters’ preferences are drawn randomly and independently from the set of all possible preferences, the probability of a Condorcet winner existing tends to zero as the number of possible options goes to infinity.

Before embarking on the alternatives to majority rule, let us present some Condorcet-consistent selection procedures; that is, procedures that select the Condorcet winner as the single winner when it exists. The first, due to Miller, is the uncovered set. An option \( x \) is covered if there exists some other option \( y \) such that (1) \( y \) beats \( x \) (with a majority of votes) and (2) \( y \) beats any option \( z \) that \( x \) can beat. If \( x \) is Pareto-dominated by some option, then \( x \) must be covered. The uncovered set is the set of options that are not covered. For the preferences such as in top cycle example above, \( d \) is covered by \( c \) because \( d \) is below \( c \) in everyone’s ranking. Thus the uncovered set is a subset of the top cycle.

If more restrictions are imposed on the agenda, then it is possible to reduce substantially the set of possible voting outcomes. One notable example is the successive-elimination agenda according to which all options are put into an ordered list, and voters are asked to eliminate the first or second option, and thereafter the previous winner or
the next option. The option surviving this successive elimination is the winner, and all eliminations are resolved by sophisticated majority votes. The Bank’s set is the set of options that can be achieved as (sophisticated) outcomes of the successive-elimination agendas. It is a subset of the uncovered set.

11.5 Alternatives to Majority Rule

Even if one considers the principle of majority rule to be attractive, the failure to select the Condorcet winner when one exists may be regarded as a serious weakness of majority rule as a voting procedure. This is especially relevant because many of the most popular alternatives to majority rule also do not always choose the Condorcet winner when one does exist, although they always pick a winner even when a Condorcet winner does not exist. This is the case for all the scoring rule methods, such as plurality voting, approval voting, and Borda voting.

Each scoring rule method selects as a winner the option with the highest aggregate score. The difference is in the score voters can give to each option. Under plurality voting, voters give 1 point to their first choice and 0 points to all other options. Thus only information on each voter’s most preferred option is used. Under approval voting, voters can give 1 point to more than one option, in fact to as many or as few options as they want. Under Borda voting, voters give the highest possible score to their first choice, and then progressively lower scores to worse choices.

11.5.1 Borda Voting

Borda voting (or weighted voting) is a scoring rule. With \( n \) options, each voter’s first choice gets \( n \) points, the second choice gets \( n - 1 \) points, and so forth, down to a minimum of 1 point for the worst choice. Then the scores are added up, and the option with the highest score wins. The procedure is very simple, and almost always picks a winner (even if there is no Condorcet winner). So a fair question is: Which requirements of Arrow’s theorem does it violate?

Suppose that there are seven voters whose preferences over three options \( \{a, b, c\} \) are as shown in table 11.5 (with numbers in parentheses representing the number of voters). Thus three voters have \( a \) as their first choice, \( b \) as their second, and \( c \) as their third.

Clearly, there is no Condorcet winner: five out of the seven voters prefer \( a \) to \( b \), and five out of seven prefer \( b \) to \( c \), and then four out of seven prefer \( c \) to \( a \), which leads to
applying the Borda method as described above, it is easy to see that \( a \) with three first places, two second places, and two third places will be the Borda winner with 15 points (while \( b \) gets 14 points and \( c \) gets 13 points). So we get the Borda ranking
\[
 a \succ b \succ c
\]
where the symbol \( \succ \) denotes strict preference. But now let us introduce a new option \( d \). This becomes the first choice of three voters but a majority prefer \( c \), the worst option under Borda rule, to the new alternative \( d \). The new preference lists are given in table 11.6.

If we compute the scores with the Borda method (now with points from one to four), the election results are different: \( d \) will be the Borda winner with 22 points, \( c \) will be second with 17 points, \( b \) will be third with 16 points, and \( a \) will be fourth with 15 points. So the introduction of the new option \( d \) has reversed the Borda ranking between the original alternatives to \( a \prec b \prec c \). This reversal of the ranking shows that the Borda rule violates the independence of irrelevant alternatives and should be unacceptable as a voting procedure.

This example illustrates the importance of Arrow’s condition I. Without imposing this requirement it would be easy to manipulate the voting outcome by adding or removing irrelevant alternatives without any real chance of them winning the election in order to alter the chance of real contenders winning.
The voters’ preferences over the three options given in table 11.7. Clearly, a majority of voters rate $c$ as worst option, but it also has a dedicated minority who rate it best (four out of nine voters). Under plurality voting $c$ is the winner, with four first-place votes, while $b$ and $a$ have three and two respectively. The example illustrates the problem that the plurality rule fails to select the Condorcet winner, which in this case is $a$ ($a$ beats both $b$ and $c$ with majority votes). The reason for this is that plurality voting dispenses with all information other than about the first choices.

### 11.5.3 Approval Voting

One problem with plurality rule is that voters don’t always have an incentive to vote sincerely. Any rule that limits each voter to cast a vote for only one option forces the voters to consider how likely it is that their first choice will win the election. If the first-choice option is unlikely to win, the voters may instead vote for a second (or even lower) choice to prevent the election of a worse option.

In response to this risk of misrepresentation of preferences (i.e., strategic voting), Brams and Fishburn (1978) have proposed the approval voting procedure. They argue that this procedure allows voters to express their true preferences. Under approval voting voters may each vote (approve) for as many options as they like. Approving one option does not exclude approving any other options. So there is no cost in voting for an option that is unlikely to win. The winning option is the one that gathers the most votes. This procedure is simpler than Borda rule because, instead of giving a score for all the possible options, voters only need to separate the options they approve of from those
Table 11.8
Approval voting

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
<td>c</td>
</tr>
<tr>
<td>b</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>c</td>
<td>c</td>
<td>a</td>
</tr>
</tbody>
</table>

they do not. Approval voting also has the advantage over pairwise voting procedures that voters need only vote once, instead of engaging in a repetition of binary votes (as in the Condorcet method).

The problem with approval voting is that it may fail to pick the Condorcet winner when one exists. Suppose that there are five voters with the preferences shown in table 11.8. With pairwise majority voting, $a$ beats both $b$ and $c$ with a majority of 3 and 4 votes out of 5 respectively, making $a$ a Condorcet winner. Now consider approval voting, and suppose that each voter gives his approval votes to the first and second choices on his list but not the bottom choice. Then $b$ will be the winner with 5 approval votes (everyone gives it an approval vote), $a$ will be second with 4 approval votes (one voter does not approve this option), and $c$ will be third with 1 vote. So approval voting fails to pick the Condorcet winner.

11.5.4 Runoff Voting

The runoff is a very common scheme used in many presidential and parliamentary elections. Under this scheme only first-place votes are counted, and if there is no majority, there is a second runoff election involving only the two strongest candidates. The purpose of a runoff is to eliminate the least-preferred options. Runoff voting seems fair, and it is very widely used. However, runoff has two drawbacks. First, it may fail to select a Condorcet winner when it exists; second, it can violate positive responsiveness, which is a fundamental principle of democracy. Let us consider these two problems in turn.

The failure to select a Condorcet winner is easily seen by considering the same set of voters’ preferences as for the plurality voting example (table 11.7). Recall that in this example, $a$ is the Condorcet winner. In the first round, $c$ has 4 votes, $b$ has 3 votes, and $a$ has 2 votes. So $a$ is eliminated and the second runoff election is between $b$ and $c$. Supporters of the eliminated option, $a$, move to their second choice, $b$; that would give $b$ an additional two votes in the runoff, and a decisive victory over $c$ (with 5 votes against 4). So this runoff voting fails to select the Condorcet winner, $a$. 
Table 11.9
Runoff voting

<table>
<thead>
<tr>
<th></th>
<th>(6)</th>
<th>(5)</th>
<th>(4)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>c</td>
<td>b</td>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>b</td>
<td>a</td>
<td>c</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>c</td>
<td>b</td>
<td>a</td>
<td>c</td>
<td>c</td>
</tr>
</tbody>
</table>

To illustrate the violation of positive responsiveness, consider the example in table 11.9, which is due to Brams (1994), with 4 options and 17 voters. There is no Condorcet winner: a beats b, c beats a, and b beats c. Under runoff voting, the result of the first election is a tie between options a and b, with 6 votes each, while c is eliminated, with only 5 votes. There is no majority, and a runoff is necessary. In the runoff between a and b, the supporters of c move to their second choice, a, giving a an extra 5 votes and a decisive victory for a over b. This seems fair: c is the least-preferred option and there is a majority of voters who prefer a over b.

Now suppose that preferences are changed so that option a attracts extra supporters from the two voters in the last column who switch their first choice from b to a. Then a will lose! Indeed the effect of this switch in preferences is that b is now the option eliminated in the first election, and there is still no majority. Thus a runoff is necessary between a and c. The disappointed supporters of b move to their second choice giving 5 more votes to c and the ultimate victory over a. The upshot is that by attracting more supports, a can lose a runoff election that it would have won without the extra support.

11.6 The Paradox of Voting

The working assumption employed in analyzing voting so far has been that all voters choose to cast their votes. It is natural to question whether this assumption is reasonable. Although in some countries voting is a legal obligation, in others it is not. The observation that many of the latter countries frequently experience low voter turnouts in elections suggests that the assumption is unjustified.

Participation in voting almost always involve costs. There is the direct cost of traveling to the point at which voting takes place, and there is also the cost of the time employed. If the individuals involved in voting are rational utility-maximizers, then they will only choose to vote if the expected benefits of voting exceed the costs.

To understand the interaction of these costs and benefits, consider an election that involves two political parties. Denote the parties by 1 and 2. Party 1 delivers to the voter
an expected benefit of $E^1$ and party 2 a benefit of $E^2$. It is assumed that $E^1 > E^2$, so 
the voter prefers party 1. Let $B = E^1 - E^2 > 0$ be the value of party 1 winning versus 
losing. If the voter knows that party 1 will win the election, then he will choose not 
to vote. This is because the voter gains no benefit from doing so but still bears a cost. 
Similarly voters will also not vote if they expect party 2 to win. In fact the rational voter 
will only ever choose to vote if he expects that he can affect the outcome of the election. 
Denoting the probability of breaking a tie occurring by $P$, then the expected benefit of 
voting is given by $PB$. The voting decision is now based on whether $PB$ exceeds the 
private cost of voting $C$. Intuition suggests that the probability of being pivotal decreases 
with the size of the voting population and increases with the predicted closeness of the 
election. This can be demonstrated formally by considering the following coin-toss 
model of voting.

There is a population of potential voters of size $N$. Each of the voters chooses to cast 
a vote with probability $p$ (so they don’t vote with probability $1 - p$). This randomness 
in the decision to vote is the “coin-toss” aspect of the model. Contesting the election 
are two political parties, which we will call party 1 and party 2. A proportion $\sigma_1$ of the 
population supports party 1, meaning that if this population did vote they would vote 
for party 1. Similarly a proportion $\sigma_2$ of the population supports party 2. It must be the 
case that $0 \leq \sigma_1 + \sigma_2 \leq 1$. If $\sigma_1 + \sigma_2 < 1$, then some of the potential voters do not 
support either political party and abstain from the election. The number of votes cast 
for party 1 is denoted $X_1$ and the number for party 2 by $X_2$.

Now assume that the election is conducted. The question we want to answer is: What 
is the probability that an additional voter can affect the outcome? An additional person 
casting a vote can affect the outcome in two circumstances:

• If the vote had resulted in a tie with $X_1 = X_2$. The additional vote can then break 
the tie in favor of the party they support.
• If the party the additional person supports was 1 vote short of a tie. The additional 
vote will then lead to a tie.

Now assume that the additional voter supports party 1. (The argument is identical 
if they support party 2.) The first case arises when $X_1 = X_2$, so the additional vote 
will break the tie in favor of party 1. The second case occurs if $X_1 = X_2 - 1$, so the 
additional vote will ensure a tie. The action in the event of a tie is now important. We 
assume, as is the case in the United Kingdom, that a tie is broken by the toss of a fair 
coin. Then when a tie occurs each party has a 50/50 chance of winning the vote.

Putting these points together, the probability of being pivotal can be calculated. If 
the original vote resulted in a tie, the additional vote will lead to a clear victory. Without
Chapter 11: Voting

The additional vote the tie would have been broken in favor of party 1 just \( \frac{1}{2} \) of the time so the additional vote leads to a reversal of the outcome with probability \( \frac{1}{2} \). If the original vote had concluded with party 1 having 1 less vote than party 2, the addition of another vote for party 1 will lead from defeat to a tie. The tie is won by party 1 just \( \frac{1}{2} \) of the time. The probability, \( P \), of being pivotal and affecting the outcome can then be calculated as

\[
P = \frac{1}{2} \Pr(X_1 = X_2) + \frac{1}{2} \Pr(X_1 = X_2 - 1).
\]

(11.1)

To see the implication of this formula, take the simple case of \( N = 3, \sigma_1 = \frac{1}{3}, \sigma_2 = \frac{2}{3} \), and \( p = \frac{1}{2} \). The probabilities of the various outcomes of the election are summarized in figure 11.6. These are calculated by observing that with 3 voters and 2 alternatives for each voter (vote or not vote), there are 8 possible outcomes. Since 2 of the 3 voters prefer party 2, the probability of party 2 receiving 1 vote is twice that of party 1 receiving 1 vote.

Using these probabilities, we can calculate the probability of the additional voter affecting the outcome as

\[
P = \frac{1}{2} \left[ \Pr(X_1 = X_2 = 0) + \Pr(X_1 = X_2 = 1) \right] \\
+ \frac{1}{2} \left[ \Pr(X_1 = 0, X_2 = 1) + \Pr(X_1 = 1, X_2 = 2) \right]
\]

(11.2)

\[
= \frac{1}{2} \left[ \frac{1}{8} + \frac{2}{8} \right] + \frac{1}{2} \left[ \frac{2}{8} + \frac{1}{8} \right] \\
= \frac{3}{8}.
\]
With this probability the voter will choose to vote in the election if
\[ V = \frac{3}{8} B - C > 0. \] (11.3)

In an election with a small number of voters the benefit does not have to be much higher than the cost to make it worthwhile to vote.

The calculation of the probability can be generalized to determine the dependence of \( P \) on the values of \( N, \sigma_1, \sigma_2, \) and \( p \). This is illustrated in the following two figures. Figure 11.7 displays the probability of being pivotal against the number of potential voters for three values of \( p \) given that \( \sigma_1 = \sigma_2 = 0.5 \). We can interpret the value of \( p \) as being the willingness to participate in the election. The figures show clearly how an increase in the number of voters reduces the probability of being pivotal. Although the probability tends to zero as \( N \) becomes very large, it is still significantly above zero at \( N = 100 \).

Figure 11.8 confirms the intuition that the probability of being pivotal is highest when the population is evenly divided between the parties. If the population is more in favor of party 2 (the case of \( \sigma_1 = 0.25, \sigma_2 = 0.75 \)), then the probability of the additional voter being pivotal in favor of party 1 falls to 0 very quickly. If the initial population is evenly divided, the probability of a tie remains significant for considerably larger values of \( N \).

The probability of a voter being pivotal can be approximated by a reasonably simple formula if the number of potential voters, \( N \), is large and the probability of each one
voting, \( p \), is small. Assume that this is so, and that the value of \( pN \) tends to the limit of \( n \). The term \( n \) is the number of potential voters that actually choose to vote. The probability of being pivotal is then

\[
P = \frac{e^{n(2\sqrt{\sigma_1\sigma_2} - \sigma_1 - \sigma_2)}}{4\sqrt{\pi n} (\sigma_1\sigma_2)^{3/2}} \left( \frac{\sqrt{\sigma_1} + \sqrt{\sigma_2}}{\sqrt{\sigma_1}} \right),
\]

(11.4)

where \( \pi \) is used in its standard mathematical sense. From this equation can be observed three results:

- The probability is a decreasing function of \( n \). This follows from the facts that \( 2\sqrt{\sigma_1\sigma_2} - \sigma_1 - \sigma_2 \leq 0 \), so the power on the exponential is negative, and that \( n \) is also in the denominator. Hence as the number of voters participating in the election increases, the probability of being pivotal falls.

- For any given value of \( \sigma_1 \), the probability increases the closer is \( \sigma_2 \) to \( \sigma_1 \). Hence the probability of being pivotal is increased the more evenly divided is the support for the parties.

- For a given value of \( n \), the probability of being pivotal is at its maximum when \( \sigma_1 = \sigma_2 = \frac{1}{2} \), and the expression for \( P \) simplifies to \( P = \frac{1}{\sqrt{2\pi n}} \). In this case the effect of increasing \( n \) is clear.

The bottom line of this analysis is that the probability that someone’s vote will change the outcome is essentially zero when the voting population is large enough.
Table 11.10  
Testing the paradox of voting

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.4033</td>
<td>0.0256</td>
</tr>
<tr>
<td>Closeness</td>
<td>0.1656</td>
<td>0.0527</td>
</tr>
<tr>
<td>Voting population</td>
<td>−0.0161</td>
<td>0.0036</td>
</tr>
<tr>
<td>Blacks (%)</td>
<td>−0.4829</td>
<td>0.0357</td>
</tr>
<tr>
<td>Rain on election day</td>
<td>−0.0349</td>
<td>0.0129</td>
</tr>
<tr>
<td>New residents (%)</td>
<td>−0.0127</td>
<td>0.0027</td>
</tr>
</tbody>
</table>

Source: Shachar and Nalebuff (1999, tab. 6).

a. All coefficients are significantly different from zero at the 1 percent level.

So, if voting is costly, the cost–benefit model should imply almost no participation. The small probability of a large change is not enough to cover the cost of voting. Each person’s vote is like a small voice in a very large crowd.

Table 11.10 presents the results of an empirical analysis of voter turnout to test the basic implications of the pivotal-voter theory (i.e., that voting should depend on the probability of a tie). It uses a linear regression over aggregate state-by-state data for 11 US presidential elections (1948–1988) to estimate the empirical correlation between the participation rate and the strategic variables (population size and electoral closeness). The analysis also reveals the other main variables relevant for participation. As the table shows, there is strong empirical support for the pivotal-voter argument: a smaller population and a closer election are correlated with higher participation. It also reveals that black participation is 48 percent lower, that new residents are 1.2 percent less likely to vote, and that rain on the election day decreases participation by 3.4 percent.

The paradox of voting raises serious questions about why so many people actually vote. Potential explanations for voting could include mistaken beliefs about the chance of affecting the outcome or feelings of social obligation. After all, every democratic society encourages its citizens to take civic responsibilities seriously and to participate actively in public decisions. Even if the act of voting is unlikely to promote self-interest, citizens feel they have a duty to vote. And this is exactly the important point made by the cost–benefit model of voting. Indeed economists are suspicious about trying to explain voting only by the civic responsibility argument. This is because the duty model cannot explain what the cost–benefit model can, namely that many people do not vote and that turnout is higher when the election is expected to be close.
11.7 The “Alabama” Paradox

The Alabama paradox is associated with the apportionment problem. Many democratic societies require representatives to the parliament to be apportioned among the states or regions according to their respective population shares. Such a rule for proportional representation apportionment arises in the EU context where representation in European institutions is based on the population shares of member states. At the level of political parties, there is also the proportional representation assignment of seats to different parties based on their respective vote shares. For instance, with the party “list system” in Belgium, electors vote for the list of candidates provided by each party. Then the number of candidates selected from each list is determined by the share of the vote a party receives. The selection is made according to the ordering of the candidates on the list from top to the bottom.

In all these forms of apportionment the solutions may involve fractions, but the number of representatives has to be an integer. How can these fractions be handled? With only two parties, rounding off will do the job. But rounding off loses simplicity once there are more than two parties, and it can produce an unexpected shift in power. To illustrate, suppose that 25 seats are to be allocated among three political parties (or states) based on their voting (population) shares as given in Table 11.11. The exact apportionment for a party is obtained by allocating the 25 seats in proportion of the vote shares. However, such a scheme requires that the three parties should share one seat (hardly feasible!). The obvious solution is to allocate the contested seat to the party with the largest fractional part. This solution seems reasonable and was proposed by the American statesman Alexander Hamilton (despite the strong opposition of Thomas Jefferson). It was then used for a long period of time in the United States. Applying this solution to our problem gives the contested seat to the small Center party (with a fractional part of 0.5 against 0.25 for the two other parties).

Table 11.11
Apportionment of seats

<table>
<thead>
<tr>
<th>Party</th>
<th>Vote share</th>
<th>Exact apportionment</th>
<th>Hamilton apportionment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left</td>
<td>0.45</td>
<td>11.25</td>
<td>11</td>
</tr>
<tr>
<td>Right</td>
<td>0.41</td>
<td>10.25</td>
<td>10</td>
</tr>
<tr>
<td>Center</td>
<td>0.14</td>
<td>3.5</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>1</td>
<td>25</td>
<td>25</td>
</tr>
</tbody>
</table>
Now what is the problem? Recall the runoff voting problem that more support for a candidate can make this candidate lose the election. A similar paradox arises with Hamilton’s apportionment scheme: increasing the number of seats available can remove seats from some parties. And it did happen in practice: when the size of the US House of Representatives grew, some states lost representation. The first to lose seats was Alabama (hence the name of Alabama paradox). To see this paradox with our simple example, suppose that one extra seat has to be allocated bringing the total number of seats to 26. Recalculating the Hamilton apportionment accordingly, it follows that the small party loses out by one seat, which implies a 25 percent loss of its representation. The large parties have benefited from this expansion in the number of seats. It is unfair that one party loses one seat when more seats become available. The explanation for this paradox is that larger parties have their fractional part quickly jumping to the top of the list when extra seat becomes available.

11.8 Political Competition

The theory of political competition addresses the question of how political parties choose the policies they present to the electorate. It might be thought that parties choose policies solely on the basis of political philosophy. This view ignores the incentives that parties have to announce policies that make them electable, even if the policies are not exactly in accord with stated philosophy. The models we now describe show how these trade-off between electability and political philosophy are resolved.

11.8.1 Downsian Model

The seminal model of political competition is the Downsian model. Anthony Downs (1957) assumed two-party competition without uncertainty about voter preferences and with the central assumption that “parties formulate policies in order to win elections, rather than win elections in order to formulate policies” (Downs 1957, p. 28). The Downsian model suggests that political parties choose and create their policies to secure the most votes as opposed to following party ideology. When there are only two parties the Downsian model, in which parties are only interested in winning elections, predicts that parties will converge to the Condorcet winner (the option preferred by a majority to any other feasible option) if it exists.

In an earlier contribution Harold Hotelling (1929) had already shown how the median voter preference would emerge as the outcome of two-party competition. In
the Hotelling model voters had political opinions represented by a point on a one-dimensional ideology line measured from left to right. Voters had a preference for proximity in the sense that they would vote for the party proposing the policy (also represented by a point on the ideology line) closest to their own position. The Hotelling–Downs model follows this tradition and assumes single-peaked preferences. It also makes numerous other assumptions such as a unidimensional policy, a symmetric and unimodal distribution of voter preferences, no voting abstention, and no uncertainty over the outcome of voting.

As already noted, the existence of a Condorcet winner requires unidimensional policy and well-behaved preferences (single-peaked or single crossing). But, more important, real world observations suggest that parties do not always converge to a common policy. The model is therefore not in keeping with policy divergence (or political polarization) as may sometimes be observed in electoral competition. To fit this observation, the model has been adapted to account for candidates with policy preferences. When candidates have policy preferences, their payoff depends on the policy they implement, not just on being elected. Indeed we could reasonably argue that most parties are not only interested in winning elections but are also policy-motivated (if not, what would be the meaning of party ideology, and the cleavage between “left” and “right”?). However, making the change to candidates with policy preferences does not suffice to affect the main result: in a two-party election both candidates will still converge to the middle in an effort to attract the median voter, no matter the extent of disagreement on the preferred policy between candidates. The reason is simply that party ideology is useless if the candidate cannot secure election. So, by deviating from the median vote, the candidate loses the election and thus the opportunity to implement policy. A further change to the model, to include parties that are uncertain about the result of the election, can obtain policy divergence. We now add these two features into the standard Hotelling–Downs model to demonstrate policy divergence.

### 11.8.2 Policy Divergence

Consider a two-party model with two candidates, $A$ and $B$, displaying their policy preference. Candidate $j$’s ideal policy is $\tau_j$, where $0 \leq \tau_j \leq 1$, for $j = A, B$. Suppose that $\tau_A = t$ (with $t < \frac{1}{2}$) and that $\tau_B = 1 - t$. Candidate $j$ has preferences over policy choices, $q$, given by

$$u(q; \tau_j) = -(q - \tau_j)^2. \tag{11.5}$$
This utility function reflects preferences for proximity, so a candidate prefers to choose a policy that is closer to their own preferred position. Voters have similar preferences over policy but with the ideal policy denoted by $\theta$, $0 \leq \theta \leq 1$. That is, the voter with preference $\theta$ has the (single-peaked) utility over policy $q$ given by

$$u(q; \theta) = -(q - \theta)^2.$$ \hfill (11.6)

We also include uncertainty of candidates over voter preferences. Candidates believe that the ideal policy of the median voter, $\theta_m$, is uniformly distributed on the interval $\frac{1}{2} \pm \Delta$ with $0 \leq \Delta \leq \frac{1}{2} - t$. If the two candidates choose policies $q_A$ and $q_B$, with $q_A < q_B$, then the voter who is indifferent between the parties is defined by $u(q_A; \theta^*) = u(q_B; \theta^*)$. Solving this equation for $\theta^*$ gives

$$\theta^*(q_A, q_B) = \frac{q_A + q_B}{2}.$$ \hfill (11.7)

Candidate A wins the election if the preference of the median voter is less than the preference value for the indifferent voter. This occurs when $\theta_m < \theta^*(q_A, q_B)$. Assume that $\frac{1}{2} - \Delta \leq \frac{q_A + q_B}{2} \leq \frac{1}{2} + \Delta$, the probability that A wins then is

$$\pi(q_A, q_B) = \text{prob} \left( \theta_m < \frac{q_A + q_B}{2} \right) = \frac{1}{2} + \frac{q_A + q_B - 1}{4\Delta}.$$ \hfill (11.8)

Candidates obtain a rent from office of $r > 0$ if they win the election and 0 otherwise. Assume that $r < 2\Delta(1-2t)$ to ensure that the desire to win the election is not so strong that it dominates the preference for policy.

The timing of events is as follows. Each candidate proposes their policy. Then each voter votes for the candidate whose policy he prefers (i.e., that is closest to his ideal policy). The candidate with the most votes is elected and implements the announced policy. A pair of policies $(q_A^*, q_B^*)$ is an equilibrium of this game if $q_A^*$ maximizes the payoff

$$\pi(q_A, q_B^*) \left[ r - (q_A - t)^2 \right] + (1 - \pi(q_A, q_B^*)) \left[ -(q_B - t)^2 \right],$$ \hfill (11.9)

of candidate A given $q_B^*$, and $q_B^*$ maximizes the payoff

$$\pi(q_A^*, q_B) \left[ -(1 - t - q_A)^2 \right] + (1 - \pi(q_A^*, q_B)) \left[ r - (1 - t - q_B)^2 \right]$$ \hfill (11.10)

of candidate B given $q_A^*$. 
We can solve the game for a symmetric equilibrium \( q^*_A = 1 - q^*_B \). This equilibrium need not be the median outcome \( q^*_A = 1 - q^*_B = \frac{1}{2} \), but it can be divergent. The first-order condition for the choice of candidate \( A \) is

\[
\frac{\partial \pi(q^*_A, q^*_B)}{\partial q_A} \left[ r - (q^*_A - t)^2 + (q^*_B - t)^2 \right] - \pi(q^*_A, q^*_B)2(q^*_A - t) = 0. \quad (11.11)
\]

The symmetry of the equilibrium implies \( \pi(q^*_A, q^*_B) = \frac{1}{2} \) and \( \frac{\partial \pi}{\partial q_A} = \frac{1}{4} \Delta \), so the first-order condition becomes

\[
\frac{1}{4\Delta} \left[ r - (q^*_A - t)^2 + (q^*_B - t)^2 \right] - (q^*_A - t) = 0. \quad (11.12)
\]

Using \( q^*_A = 1 - q^*_B \) obtains

\[
\frac{1}{4\Delta} \left[ r - (q^*_A - t)^2 + (1 - q^*_A - t)^2 \right] - (q^*_A - t) = 0. \quad (11.13)
\]

or equivalently,

\[
\frac{1}{4\Delta} \left[ r + 4r q^*_A - 2q^*_A + 1 + 2t \right] - (q^*_A - t) = 0. \quad (11.14)
\]

Solving for \( q^*_A \) gives

\[
q^*_A = \frac{\frac{r}{2} + \frac{1}{2} - t(1 - 2\Delta)}{2\Delta + 1 - 2t}. \quad (11.15)
\]

From this equation three results follow.

**Result 1** The candidates moves toward the median preference (since \( q^*_A > t \)) but not entirely (since \( q^*_A < \frac{1}{2} \) for \( r < 2\Delta(1 - 2t) \)). So \( t < q^*_A < \frac{1}{2} \). The policy choices do not converge, since \( q^*_A < \frac{1}{2} < q^*_B \).

**Result 2** The extent to which candidates depart from the median preference is increasing in voter uncertainty \( \Delta \) and decreasing in the rent for office \( r \).

**Result 3** Without uncertainty the condition \( r < 2\Delta(1 - 2t) \) cannot be satisfied and the Downsian model predicts policy convergence to the median \( q^*_A = 1 - q^*_B = \frac{1}{2} \).

These results show that policies can remain divergent, though there is a tendency toward the median. This is a consequence of the uncertainty. If this is removed, the result will be the simple Hotelling outcome of policy location at the median.
11.8.3 Multidimensional Competition

The model of political competition can be extended to permit a multidimensional policy space. This is a much more compelling setting since it allows candidates to state their positions on a range of issues.

There are two candidates \( j = \{A, B\} \) who are only motivated by winning the election (i.e., there is no policy motivation). The policy space, \( Q \), is multidimensional (\( Q \subset \mathbb{R}^m \)), and a chosen policy is a vector \( q \) located in \( Q \). There are \( n \) voters indexed \( i = 1, \ldots, n \), with voter \( i \) having utility function \( u_i \). The candidates simultaneously choose their policies, \( q_j \), and each voter votes for the candidate he prefers given the chosen policies. The candidate with the most votes wins the election and implements his announced policy.

Let \( s(q, q') \), where \( 0 \leq s(q, q') \leq 1 \), be the share of voters who prefer policy \( q \) to policy \( q' \). Using this notation, we obtain the probability that candidate \( A \) wins as

\[
\pi(q_A, q_B) = \begin{cases} 
0 & \text{if } s(q_A, q_B) < \frac{1}{2}, \\
\frac{1}{2} & \text{if } s(q_A, q_B) = \frac{1}{2}, \\
1 & \text{if } s(q_A, q_B) > \frac{1}{2}.
\end{cases}
\]

A policy pair \( (q_A^*, q_B^*) \) is an equilibrium if (1) \( q_A^* \) maximizes \( \pi(q_A, q_B) \) given \( q_B^* \), and (2) \( q_B^* \) maximizes \( 1 - \pi(q_A^*, q_B) \) given \( q_A^* \). A policy \( q^* \) is a Condorcet winner if \( s(q^*, q) > \frac{1}{2} \) for all \( q \neq q^* \). The central proposition for the model is:

**Proposition 11.5** A policy pair \( (q_A^*, q_B^*) \) is an equilibrium if and only it coincides with the Condorcet winner: \( q_A^* = q_B^* = q^* \)

The implication of the proposition is that there can be no equilibrium if there is no Condorcet winner. We have already seen that there are situations where a Condorcet winner does not exist. The nonexistence of a Condorcet winner is in fact a general feature unless we restrict our attention to single-peaked preferences (or single-crossing preferences) and one-dimensional policies.

The classical example without a Condorcet winner is the “divide the dollar game” in which \( N = 3 \), \( Q = \{q_i : \sum q_i = 1\} \), and \( u_i(q) = q_i \). In this game a single dollar is divided among 3 players. Observe that no matter how the dollar is divided among the three participants it is always possible to gain a majority in favor of taxing one participant and transferring that income to the other two. In other words, no transfer policy \( q = (q_1, q_2, q_3) \) can be a Condorcet winner. For instance, the policy \( q' = (0, q_2 + \frac{q_1}{2}, q_3 + \frac{q_1}{2}) \) is preferred by participants 2 and 3 to policy \( q \). But then policy
Chapter 11: Voting

$q'' = (q_1', q_2' + q_3', 0)$ is preferred to policy $q'$ by participants 1 and 2. And so it is possible to cycle forever by taking income from one participant and giving it to the other two.

11.8.4 “Swing Voter” Politics

The standard solution to fix the problem of there being no Condorcet winner is to consider voter uncertainty so that voting becomes probabilistic. We now provide an illustration with a model of distributive politics.

There are $n$ voters indexed $i = 1, \ldots, n$ with voter $i$ having an endowment of income $\omega_i$. A policy is a list of transfers $t = (t_1, \ldots, t_n)$, where $t_i$ is the transfer given to voter $i$. The set of feasible policies is $T = \{ t : \sum t_i = 0 \text{ and } \omega_i + t_i \geq 0 \forall i \}$. That is, the transfers must sum to 0 and the transfer from $i$ cannot exceed the income of $i$. There are two candidates $A$ and $B$ who are only motivated by winning the election (i.e., this is a pure Downsian model with no policy motivation for the candidates).

Each candidate $j$ chooses a policy $t^j \in T$. The utility of voter $i$ if candidate $j$ wins is $u_i(\omega_i + t^j_i) + \delta^j_i$, where $u_i$ is the utility of income (with $u_i'' < 0 < u_i'$) and $\delta^j_i$ is the nonpecuniary benefit derived from candidate $j$ being in office. Voter $i$ votes for candidate $A$ if

$$u_i(\omega_i + t^A_i) - u_i(\omega_i + t^B_i) > \delta^B_i - \delta^A_i,$$

(11.17)

where $\delta^B_i - \delta^A_i$ is a random variable which is not observable by the candidates. The value of $\delta^B_i - \delta^A_i$ is distributed according to the cumulative distribution function $F_i : \mathbb{R} \rightarrow [0, 1]$ with density $f_i > 0$. Using the cumulative distribution function, we write the probability that voter $i$ votes for candidate $A$ as

$$\pi_i(t^A, t^B) = F_i(u_i(\omega_i + t^A_i) - u_i(\omega_i + t^B_i)).$$

(11.18)

A pair of policies $(t^A, t^B)$ is an equilibrium if (1) $t^A$ maximizes $\sum_i \pi_i(t, t^B)$ given $t^B$, and (2) $t^B$ maximizes $\sum_i \left[ 1 - \pi_i(t^A, t) \right]$ given $t^A$. The Lagrangian describing the optimization problem for the choice of $t^A$ facing candidate $A$ is

$$L_A = \sum_i F_i(u_i(\omega_i + t_i) - u_i(\omega_i + t^B_i)) - \lambda_A \sum_i t_i,$$

(11.19)

where $\lambda_A$ is the Lagrange multiplier. The first-order conditions are, for all $i$,

$$u_i'(\omega_i + t^A_i)f_i(u_i(\omega_i + t^A_i) - u_i(\omega_i + t^B_i)) = \lambda_A,$$

(11.20)
and similarly the first-order conditions for candidate $B$ are

$$u'_i(\omega_i + t^B_i) f_i(u_i(\omega_i + t^A_i) - u_i(\omega_i + t^B_i)) = \lambda_B.$$  \hfill (11.21)

This implies that for all $i$,

$$\frac{u'_i(\omega_i + t^A_i) f_i(u_i(\omega_i + t^A_i) - u_i(\omega_i + t^B_i))}{u'_i(\omega_i + t^B_i) f_i(u_i(\omega_i + t^A_i) - u_i(\omega_i + t^B_i))} = \frac{\lambda_A}{\lambda_B}. \hfill (11.22)$$

Thus $t^A_i = t^B_i$ for all $i$ because, otherwise, the constant ratio condition cannot be satisfied. Therefore the equilibrium policy choice is $t^A = t^B = t^*$ with

$$f_i(0)u'_i(w_i + t^*_i) = \lambda.$$  \hfill (11.23)

Two results follow from these calculations.

**Result 1** Suppose that voters have different utility functions $u_i$ but the same distribution function: $f_i(0) = f$ for all $i$. Then

$$u'_i(w_i + t^*_i) = \frac{\lambda}{f(0)}, \hfill (11.24)$$

so the equilibrium policy is equivalent to the utilitarian policy of maximizing $\sum_i u_i(\omega_i + t_i)$. In this case electoral competition is equivalent to imposing a utilitarian social welfare function.

**Result 2** Suppose that voters have identical utility functions, so that $u_i = u$, but different distribution functions: $f_i(0) \neq f_k(0)$ for $i \neq k$. Then

$$u'(w_i + t^*_i) = \frac{\lambda}{f_i(0)}, \hfill (11.25)$$

and in equilibrium the marginal utility of income decreases as $f_i$ increases. The interpretation is that the more volatile voters (high $f_i(0)$) will be given greater transfers because they have a lower marginal utility of income in equilibrium. This leads to the observation that swing voters receive greater transfers!

The outcome of political competition is that candidates bid to buy the votes of swing voters. The existence of an equilibrium requires sufficient concavity of the utility of income (a decreasing marginal utility of income) and sufficient uncertainty in voters preferences so that when trading off one dollar from one voter to another, the change in the support from the involved voters will be smooth enough.
11.8.5 Citizen-Candidate Model

In the Downsian model, candidates have no policy preferences, and thus they can credibly adopt whatever policy will win the election. The Downsian model was then extended to include policy preference. Candidates could trade-off policy preferences against the necessity of winning the election. When there is some voter uncertainty, the extended model can generate policy divergence, which is the relevant outcome from an empirical perspective. However, the model still makes two strong assumptions. First, it assumes that candidates can credibly commit to implement policies that differ from their most preferred policy. Second, the number of candidates is assumed to be fixed. The model in fact assumes just two candidates competing for election, and this number is exogenously given with no theoretical justification. The only possible argument for this assumption is the empirical prevalence of two candidates in electoral competition (the so-called Duverger law). The citizen-candidate approach attempts to fill this gap by determining the number of candidates as part of the process of political competition.

In the citizen-candidate approach the number of candidates is endogenous. Every citizen in the society is both a voter and a potential candidate for office. Running for office is costly, but the elected candidate can choose the policy that is implemented. Each candidate is restricted to make only a credible policy announcement, so they cannot announce that they plan to implement a policy that differs from the candidate’s preferred policy. This restriction can be justified by the assumption that all citizens know the preferred policies of all other citizens.

The citizen-candidate model is analyzed as a three-stage game. In stage 1, each citizen decides whether or not to run for election. When running for election, there is a sunk cost $c > 0$. In stage 2, citizens vote among the self-declared candidates. They may vote sincerely for the candidate they prefer (no matter how likely this candidate is to win), or they may vote strategically for the candidate more likely to win among the candidates they prefer. In stage 3, the candidate with the most votes (adopting the plurality voting rule) is elected and implements his preferred policy. In the event of a tie, the winner is decided among the tied candidates by the toss of a (fair) coin.

The game is solved by backward induction. In stage 3, the elected candidate chooses his preferred policy. In stage 2, citizens rationally anticipates that the winner will adopt his preferred policy, and so voters can vote for candidates according to their own preferences. With sincere voting (the Osborne–Slivinski model), they will vote for the self-declared candidate closest to their own preference. With strategic voting
(the Besley–Coate model), they vote for the candidate more likely to win that is closest to their own preferences. In stage 1, the candidates anticipate perfectly how citizens will vote in the second stage given any set of candidates, and so they decide whether or not to run for election.

**Definition 11.6** A political equilibrium is a set of candidates such that after perfectly anticipating voting behavior, each candidate is better off running for election, given that the other self-declared candidates have chosen to stand, and every citizen who is not a candidate is better off not running for election.

With this political competition model, the existence of equilibrium is not an issue (even if there is no Condorcet winner). Rather, the model generates a multiplicity of equilibria, which poses a problem for prediction. To illustrate, multiplicity of equilibria is what happens when the cost of running is sufficiently small. The utility of candidate $i$ with preferred policy $x_i$ is $-c$ if elected and $-|x_i - x_j| - c$ if candidate $j$ is elected and implements policy $x_j$. Each candidate maximizes his expected utility where randomization is over the set of self-declared candidates to determine the winner. Then, it can be shown, a one-candidate equilibrium exists. Indeed, if the cost of running is sufficiently small, then the entering candidate must be a Condorcet winner. Conversely, if a Condorcet winner exists, then there is a $c$ sufficiently low that there is an equilibrium in which the Condorcet winner runs unopposed. At the same time the existence of a political equilibrium with two candidates with preferred policies $x^*_1$ and $x^*_2$ is quite general. In this case, $x^*_1 \neq x^*_2$, but the position of the two candidates is symmetric around the median, so both candidates attract the same number of votes and are equally likely to win. They cannot be too close to the median, for otherwise a successful entry is possible to either side. Their positions cannot be too far away from the median either, for otherwise a candidate between the two candidates could enter and win.

There may also exist equilibrium with three or more candidates. In a three-candidate equilibrium there are two possibilities. Either, each candidate gets exactly $\frac{1}{3}$ of the vote or the two extreme candidates receive the same number of votes, and a moderate candidate enters, though he surely loses. The motivation of the moderate candidate is that if he does not enter the race, the extremist whom he likes less wins with certainty. By entering, the moderate steals more votes from the extremist who is on the other side of the median than from the extremist who is one the same side. In other words, the moderate candidate is a spoiler.
11.9 Conclusions

Voting is one of the most common methods used to make collective decisions. Despite its practical popularity, it is not without its shortcomings. The theory of voting that we have described carefully catalogs the strengths and weaknesses of voting procedures. The major result is due to Arrow who pointed out the impossibility of finding the perfect voting system. Although there are many alternative systems of voting, none can always deliver in every circumstance.

Voting is important, but we should never forget its limitations. When discussing the various alternative voting schemes (Borda rule, approval voting, runoff voting, and plurality voting), we mentioned their respective drawbacks in terms of the violation of some of the conditions of Arrow’s theorem. However, such violations are inevitable given the content of the Impossibility Theorem. Thus violation of one condition does not rule out the use of a particular voting scheme. Whatever scheme we choose will have some problem associated with it.

The initial models of political competition assumed that the number of candidates is given and focused on the choice of policy platform by candidates. The policies proposed converge to the median of voters’ preferences if candidates care only about winning. Limited divergence of policies occurs when candidates trade some probability of winning for being able to implement a policy closer to their own preference if they win. The citizen-candidate model takes the analysis one step further by endogenizing the set of candidates. The consequence is that the model has a multiplicity of equilibria so that its predictive power is limited.

Further Reading

Some of the fundamental work on collective choice can be found in:


The two fundamental papers on the inevitable manipulability of voting schemes are:


Two excellent books providing comprehensive surveys of the theory of voting are:


Two very original presentations of voting theory are:


A quite simple and striking proof of the impossibility result is in:


There is also a nice proof of the impossibility theorem in:


The voting paradox is based on:


Some Condorcet-consistent alternatives to majority rule are presented and discussed in:


Chapter 11: Voting


A good review on political competition is:


The citizen-candidate approach is in:


Exercises

11.1 Show that unidimensional median voting with single-peaked preferences satisfies the conditions of theorem 11.2.

11.2 Suppose that to overthrow the status quo, an alternative requires 70 percent or more of the vote. Which property of voting is violated? In many committees the chairman has the casting vote. Which property of voting is violated?

11.3 With sincere voting can an example be given in which an agenda is constructed so that a Condorcet winner is defeated? Is the same true with strategic voting?

11.4 Consider five people with the preference rankings over four projects $a$, $b$, $c$, and $d$ as follows:

<table>
<thead>
<tr>
<th></th>
<th>b</th>
<th>a</th>
<th>c</th>
<th>a</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>d</td>
<td>d</td>
<td>b</td>
<td>c</td>
<td>b</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>c</td>
<td>d</td>
<td>b</td>
<td>c</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>b</td>
<td>a</td>
<td>d</td>
<td>a</td>
<td></td>
</tr>
</tbody>
</table>

a. Draw the preferences by ranking the projects by alphabetical order from left to right.
b. Who has single-peaked preferences and who has not?
c. Which project will be selected by majority voting? If none is selected, explain why.

11.5 Is condition $U$ acceptable when some voters hold extreme political preferences?

11.6 Let $G$ be the number of hours of television broadcast each day. Consider three individuals with preferences:

$$U^A = \frac{G}{4}, \quad U^B = 2 - G^{3/4}, \quad U^C = G - \frac{G^2}{2}.$$  

a. Show that the three consumers have single-peaked preferences.
b. If the government is choosing $G$ from the range $0 \leq G \leq 2$, what is the majority voting outcome?
Part IV: Political Economy

c. Does this outcome maximize the sum of utilities $W = U^A + U^B + U^C$?
d. How are the answers to parts a through c altered if the preferences of $C$ become $U^C = \frac{G^2}{T} - G$?

11.7 Consider the Cobb–Douglas utility function $U = \left[ Y^i - T^i \right]^{1-\alpha} G^\alpha$ with $0 < \alpha < 1$. Suppose that a poll tax $T = T^i$ for all $i$ is levied on each of $N$ members of society. Tax revenues are used to finance a public good $G$.

a. Show that the majority voting outcome involves the amount of public good $G = \alpha NY^m$, where $Y$ denotes the before-tax income of the median voter.
b. Now suppose that a proportional income tax $T^i = tY^i$ is levied. Show that the majority voting outcome involves the amount of public good $G = \alpha N\overline{Y}$, where $\overline{Y}$ is the mean income level.
c. When income is uniformly distributed, which outcome is closest to the efficient outcome?

11.8 Construct an example of preferences for which the majority voting outcome is not the median. Given these preferences, what is the median voting outcome? Is there a Pareto-preferred outcome?

11.9 If preferences are not single-peaked, explain why the Median Voter Theorem fails.

11.10 Show that the preferences used in section 5.3.4 are single-peaked.

11.11 Consider a scoring rule in which the preferred option is given one point and all others none.

a. Show that this need not select the Condorcet winner.
b. Demonstrate the scope for false voting.

11.12 Which of Arrow’s conditions does approval voting violate?

11.13 Discuss the individual benefits that may arise from a preferred party winning. How large are these likely to be relative to the cost of voting?

11.14 Assume that all voters have an hourly wage of $10 and that it takes half an hour to vote. They stand to gain $50 if their party wins the election (in a two-party system where support is equal). What is the number of voters at which voting no longer becomes worthwhile?

11.15 Has there been any national election where a single vote affected the outcome?

11.16 Which of Arrow’s conditions is removed to prove the Median Voter Theorem? Which condition does the Borda rule violate? Which condition does the Condorcet method fail? Why do we wish to exclude dictatorship?

11.17 In a transferable voting system each voter provides a ranking of the candidates. The candidate with the lowest number of first-choice votes is eliminated, and the votes are transferred to the second-choice candidates. This process proceeds until a candidate achieves a majority.

a. Can the Condorcet winner lose under a transferable vote system?
b. Is it possible for a candidate that is no one’s first choice to win?
c. Show how strategic voting can affect the outcome.
11.18 Consider four people with preference rankings over three projects $a$, $b$, and $c$ as follows:

\[
\begin{array}{ccc}
\text{a} & \text{b} & \text{c} \\
\text{b} & \text{b} & \text{c} \\
c & c & \text{a} \\
\end{array}
\]

a. Assume that voters cast their votes sincerely. Find a Borda rule system (scores to be given to first, second, and third choices) where project $a$ wins.
b. Find a Borda weighting system where $b$ wins.
c. Under plurality voting, which proposal wins?

11.19 The Hare procedure was introduced by Thomas Hare in 1861. It is also called the “single transferable vote system.” The Hare system is used to elect public officials in Australia, Malta, and the Republic of Ireland. The system selects the Condorcet winner if it exists. If not, then it will proceed to the successive deletion of the least-desirable alternative or alternatives until a Condorcet winner is found among the remaining alternatives. Consider the following preference profile of five voters on five alternatives:

\[
\begin{array}{cccccc}
\text{a} & \text{b} & \text{c} & \text{d} & \text{e} \\
\text{b} & \text{c} & \text{b} & \text{c} & \text{d} \\
\text{e} & \text{a} & \text{e} & \text{a} & \text{c} \\
\text{d} & \text{d} & \text{d} & \text{e} & \text{a} \\
c & \text{e} & \text{a} & \text{b} & \text{b} \\
\end{array}
\]

a. What social choice emerges from this profile under the Hare procedure? Explain in detail the successive deletions.
b. Repeat the exercise for the opposite procedure proposed by Clyde Coombs. The Coombs system operates exactly as the Hare system does, but instead of deleting alternatives with the fewest first places, it deletes alternatives with the most last places.
c. Which of Arrow’s conditions are violated in the Coombs and Hare procedures?

11.20 Define a collective choice procedure as satisfying the “top condition” if an alternative is never among the social choices unless it is on top of at least one individual preference list. Prove or disprove each of the following:

a. Plurality voting satisfies the top condition.
b. The Condorcet method satisfies the top condition.
c. Sequential pairwise voting satisfies the top condition.
d. A dictatorship satisfies the top condition.
e. Approval voting satisfies the top condition.
f. Runoff voting satisfies the top condition.
g. If a procedure satisfies the top condition, then it satisfies the Pareto condition.
h. If a procedure satisfies the top condition, then it selects the Condorcet winner (if any).
11.21 Consider the following preference profile of three voters and four alternatives:

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>c</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td></td>
<td></td>
<td>b</td>
</tr>
<tr>
<td>b</td>
<td>a</td>
<td></td>
<td>d</td>
</tr>
<tr>
<td>d</td>
<td>b</td>
<td>c</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>b</td>
<td>a</td>
<td></td>
</tr>
</tbody>
</table>

a. Show that if the social choice method used is sequential pairwise voting with a fixed agenda, and if you have agenda-setting power, then you can arrange the order to ensure whichever alternative you want to be chosen.

b. Define an alternative as a “Condorcet loser” if it is defeated by every other alternative in pairwise voting. Prove that there is no Condorcet loser in this preference profile.

c. Modify the preference profile for one voter to ensure that there exists a Condorcet loser.

d. Modify the preference profile for one voter to ensure that there exists a Condorcet winner.

11.22 Show that for an odd number of voters and a given preference profile over a fixed number of alternatives, an alternative is a Condorcet winner if and only if it emerges as the social choice in sequential pairwise voting with a fixed agenda, no matter what the order of the agenda.

11.23 (Ferejohn) There are an infinite number of time periods indexed by \( t = 0, 1, \ldots, \infty \). There is a set of identical politicians and a single representative voter. At the end of each period there is an election opposing the incumbent to a challenger. The challenger is randomly drawn. An incumbent who is voted out cannot run again for election. In any period, \( t \), the incumbent observes (privately) the realisation of a random productivity shock \( \theta_t \in [\underline{\theta}, \bar{\theta}] \) and exerts effort \( e_t \in \mathbb{R}^+ \). The random productivity shock is distributed according to a smooth and increasing CDF \( F(\theta) \) with density \( f(\theta) > 0 \) for all \( \theta \). When in office the politician receives a per-period payoff of

\[
 r - c(e_t),
\]

where \( r > 0 \) and the cost of effort \( c(e_t) \) is increasing and strictly convex with \( c(0) \). When not in office, a politician receives payoff of 0. The voter’s payoff is increasing in the incumbent performance as

\[
y_t = e_t \theta_t.
\]

The representative voter observes only the outcome \( y_t \) but does not observe the effort or the productivity. Both voter and politicians have common per-period discount rate \( 0 < \delta < 1 \).

a. Define the equilibrium of the voting model assuming the voter uses a cutoff strategy of re-electing the incumbent only if performance is above a fixed threshold.

b. Describe the optimization problem of the politicians.

c. Describe the optimization problem of the voter.

d. Show that if there are no elections, the incumbent will exert no effort in equilibrium.

e. Show that if there are elections, by using retrospective voting, the voter can induce politicians to exert effort.
12 Rent-Seeking

12.1 Introduction

The *United States National Lobbyist Directory* records there to be over 40,000 state-registered lobbyists and a further 4,000 federal government lobbyists registered in Washington. Some estimates put the total number, including those who are on other registers or are unregistered, as high as 100,000. Although the number of lobbyists in the United States dwarfs those elsewhere, there are large numbers of lobbyists in all major capitals. These lobbyists are not engaged in productive activity. Instead, their role is to seek favorable government treatment for the organizations that employ them. Viewed from the US perspective, the country has at least 40,000 (presumably skilled) individuals who are contributing no net value to the economy but are merely attempting to influence government policy and shift the direction of income flow.

The behavior that the lobbyists are engaged in has been given the name of *rent-seeking* in the economic literature. Loosely speaking, rent-seeking is the act of trying to seize an income flow rather than create an income flow. What troubles economists about rent-seeking is that it uses valuable resources unproductively and can push the government into inefficient decisions. This places the economy inside its production possibility frontier and implies that efficiency improvements will be possible. As such, rent-seeking can be viewed as a potential cause of economic inefficiency.

The chapter will first consider the nature and definition of rent-seeking. It will then proceed to the analysis of a simple game that demonstrates the essence of rent-seeking. This game generates the fundamental results on the consequences of rent-seeking and forms the basis on which the later analysis is developed. The insights from the game are then applied to rent-seeking in the context of monopoly. The basic point made there is that the standard measure of monopoly welfare loss understates the true loss to society if rent-seeking behavior is present. This partial equilibrium analysis of monopoly is then extended to a general equilibrium setting. Following this, the emphasis turns to how and why rents are created. Government policy is analyzed and the relationship between lobbying and economic welfare is characterized in detail. The reasons why a government might allow itself to be swayed by lobbyists are then discussed. Finally, possible policies for containing rent-seeking are considered.
12.2 Definitions

Rent-seeking has received a number of different definitions in the literature. These vary only in detail, particularly, in whether the resources used in rent-seeking are directly wasted and in whether the term can be applied only to rents created by government. It is not the purpose here to catalog these definitions but instead to motivate the concept of rent-seeking by example and to draw out the common strands of the definitions.

The ideas that lie behind rent-seeking can be seen by considering the following two situations:

- A firm is engaged in research intended to develop a new product. If the research is successful, the product will be unique, and the firm will have a monopoly position, and extract some rent from this, until rival products are introduced.
- A firm has introduced a new product to the home market. A similar product is manufactured overseas. The firm hires lawyers to lobby the government to prevent imports of the overseas product. If it is successful, it will enjoy a monopoly position from which it will earn rents.

What is the difference between these two situations? Both will give the firm a monopoly position, at least in the short run, from which it can earn monopoly rents. The first, though, would be seen by many economists as something to be praised, but the second as something to be condemned. In fact the fundamental difference is that the first case, with the firm expending resources to develop a new product, will lead to monopoly rent only if the product is successful and valued by consumers. Hence the resources used in research may ultimately lead to an increase in economic welfare. In contrast, the resources used in the second case are reducing economic welfare. If the lawyers are successful, consumers will be denied a choice between products, and the lack of competition will mean that they face higher prices. Their welfare is reduced and some of their income, via the higher prices, is diverted to the monopolist. There is also (implicitly) a transfer from the overseas producers to the monopolist. Some of the monopoly rents are transferred to the lawyers via their fees (we will clarify how much in section 12.3). In short, although the research and the lawyers are both directed to attaining a monopoly position, in the first case research potentially increases economic welfare, but in the second the lawyers reduce it.

These comments now allow us to distinguish between two concepts:

- *Profit-seeking* is the expenditure of resources to create a profitable position that is ultimately beneficial to society. Profit-seeking, as exemplified by the example of
research, is what drives progress in the economy and is the motivating force behind competition.

- Rent-seeking is the expenditure of resources to create a profitable opportunity that is ultimately damaging to society. Rent-seeking, as exemplified by the use of lawyers, hinders the economy and limits competition.

There are some other points that can be drawn out of these definitions. Notice that the scientists and engineers employed in research are being productive. If their work is successful, then new products will emerge that raise the economy’s output. In contrast, the lawyers engaged in lobbying the government are doing nothing productive. Their activity does not raise output. At best, it simply redistributes what there already is, and generally it reduces it. Furthermore output would be higher if they were usefully employed in a productive capacity rather than working as lawyers. In this respect rent-seeking always reduces total output, since the resources engaged in rent-seeking can be expected to have alternative productive uses.

It can be inferred from this discussion that rent-seeking can take many forms. All lobbying of government for beneficial treatment, be it protection from competition or the payment of subsidies, is rent-seeking. Expenditure on advertising is rent-seeking and so is arguing for tariffs to protect infant industries. These activities are rife in most economies, so rent-seeking is a widespread and important issue.

One of the factors that will feature strongly in the discussion below is the level of resources wasted in the lobbying process. At first sight there appears to be a clear distinction between the time a lobbyist uses talking to a politician and a bribe passed to a politician. The time is simply lost to the economy—it could have been used in some productive capacity but has not. This is a resource wasted. In contrast, the bribe is just a transfer of resources. Beyond the minimal costs needed to deliver the bribe, there appear to be no other resource costs. Hence it is tempting to conclude that lobbying time has a resource cost whereas bribes do not. Thus, if rent-seeking is undertaken entirely by bribes, it appears to have no resource cost.

Before reaching this conclusion, it is necessary to take a further step back. Consider the position of politicians receiving bribes. How did they achieve their position of authority? Clearly, resources would have been expended to obtain election. If potential politicians believed they would receive bribes once elected, they would be willing to expend more resources to become politicians—they are in fact rent-seeking themselves. Much of the resources used in seeking election will simply be a cost to the economy with no net output resulting from them. Through this process a bribe, which is just a transfer, actually becomes transformed further down the line into a resource loss caused
by rent-seeking. These arguments suggest that caution is required in judging between
lobby costs that seem to be transfers and those that are clearly resource costs.

So far the discussion has concentrated on rent-seeking. The economic literature has
also dealt with the very closely related concept of directly unproductive activities.
The distinction between the two is not always that clear, and many economists use
them interchangeably. If there is a precise distinction, it is in the fact that directly
unproductive activities are by definition a waste of resources whereas the activity of
rent-seeking may not always involve activities that waste resources. The focus below
will be placed on rent-seeking, though almost all of what is said could be rephrased in
terms of directly unproductive activities.

12.3 Rent-Seeking Games

This section considers several variants of a simple game that is designed to capture
the essential aspects of rent-seeking. From the analysis emerge several important con-
clusions that will form the basis of more directly economic applications in the follow-
ing sections. The game may appear at first sight to be extreme, but on reflection, its
interpretation in terms of rent-seeking will become clear.

The basic structure of the game is as follows: Consider the offer of a prize of $10,000.
Competitors enter the game by simultaneously placing a sum of money on a table and
setting it alight. The prize is awarded to the competitor that burns the most money.
Assuming that the competitors are all identical and risk-neutral, how much money will
each one burn? This question will be answered when there is either a fixed number of
competitors or the number of competitors is endogenously determined through free-
entry into the competition.

Before conducting the analysis, it is worth detailing how this game relates to rent-
seeking. The prize to be won is the rent—think of this as the profit that will accrue if
awarded a monopoly in the supply of a product. The money that is burned represents the
resources used in lobbying for the award of the monopoly. Instead of burning money,
if could be fees paid to a lobby company for the provision of their services. The game
can then be seen as representing a number of companies each wishing to be granted
the monopoly and employing lobbyists to make their case. We consider two different
games. In the deterministic game, the prize is awarded to the firm that spends most on
lobbying. In the probabilistic game, the chance of obtaining the prize is an increasing
function of one’s share in the total spending on lobbying, so spending the most does
not necessarily secure a win.
Chapter 12: Rent-Seeking

12.3.1 Deterministic Game

A game of this form is solved by constructing its equilibrium. In this case we look for the *Nash equilibrium*, which occurs when each competitor’s action is optimal given the actions of all other competitors. Consequently at a Nash equilibrium no variation in one competitor’s choice can be beneficial for that competitor. It is this latter property that allows potential equilibria to be tested.

Say initially that there are two competitors for the prize. To apply the Nash equilibrium argument, the method is to fix the strategy choice of one competitor and to consider what the remaining competitor will do. Strategies for the game can be of two types. There are *pure strategies* that involve the choice of a single quantity of money to burn. There are also *mixed strategies* where the competitor uses a randomizing device to select its optimal strategy. The benefit of randomizing is that if one player engages in any determinate behavior, the rival can take advantage of it. The only sensible thing for each to do is to mix its action randomly to act in an unpredictable way for its rival. For instance, labeling six possible strategies from 1 to 6 and then using the roll of a die to choose which one to play is a mixed strategy. The central component of finding a mixed strategy equilibrium is to determine the mixing rule described by the probabilities assigned to each pure strategy. The argument will first show that there can be no pure strategy equilibrium for the game. The mixed strategy equilibrium will then be constructed.

To show that there can be no pure strategy equilibrium, let competitor 1 burn an amount $B^*$. Then, if competitor 2 burns $B^* + \epsilon$, this competitor will win the contest and receive the prize of value $V$. The same argument applies for any value of $B^* < V$ and any positive value of $\epsilon$, no matter how small. Since competitor 1 has lost the contest, burning $B^*$ cannot be an equilibrium choice: competitor 1 will wish to burn slightly more than $B^* + \epsilon$. By this reasoning, no amount of burning less than $V$ can be an equilibrium. The only way a competitor can prevent this “leapfrogging” argument is by burning exactly $V$. The other competitor must then also burn $V$.

However, each competitor burning $V$ is still not an equilibrium. If both competitors burn $V$, then each has an equal chance of winning. This chance of winning is $\frac{1}{2}$, so their expected payoff, $EP$, is equal to the expected value of the prize minus the money burned,

$$EP = \frac{1}{2}V - V = -\frac{1}{2}V < 0.$$  \hspace{1cm} (12.1)

Clearly, given that the other burns $V$, a competitor would be better off to burn 0 and make an expected payoff of 0 rather than burn $V$ and make an expected loss of $-\frac{V}{2}$. 
So the strategies of both burning $V$ are not an equilibrium. The conclusion of this reasoning is that the game has no equilibrium in pure strategies. Therefore, to find an equilibrium, it becomes necessary to look for one in mixed strategies.

The calculation of the mixed strategy for the game is easily motivated. It is first noted that each player can obtain a payoff of at least 0 by burning nothing. Therefore the equilibrium strategy must yield a payoff of at least 0. No player can ever burn a negative amount of money, nor is there any point in burning more than $V$. Hence the strategy must assign positive probability only to amounts in the range 0 to $V$.

It turns out that the equilibrium strategy is to assign the same probability to all amounts in the range 0 to $V$. This probability, denoted $f(B)$, must then be given by $f(B) = \frac{1}{V}$. Given that the other competitor also plays this mixed strategy, the probability of winning when burning an amount $B$ is the probability that the other competitor burns less than $B$. This can be calculated as $F(B) = \frac{B}{V}$. Burning $B$ then gives an expected payoff of

$$EP = \left[\frac{B}{V}\right]V - B = 0.$$  \hfill (12.2)

Therefore, whatever amount the random device suggests should be played, the expected payoff from that choice will be zero. In total, the mixed strategies used in this equilibrium give both players an expected payoff of zero.

In the context of rent-seeking, the important quantity is the total sum of money burned, since this can be interpreted as the value wasted. The mixed strategy makes each value between 0 and $V$ equally likely so the expected burning for each player is $\frac{V}{2}$. Adding these together, the total amount burnt is $V$—which is exactly equal to the value of the prize. This conclusion forms the basis of the important result that the effort put into rent-seeking will be exactly equal to the rent to be won.

The argument can now be extended to any number of players. With three players the strategy of giving the same probability to each value between 0 and $V$ is not the equilibrium. To see this, observe that with this mixed strategy the average amount burned remains at $\frac{V}{2}$, but the probability of winning with three players is reduced to $\frac{1}{3}$. The expected payoff is therefore

$$EP = \left[\frac{1}{3}\right]V - \frac{V}{2} = -\frac{V}{6},$$  \hfill (12.3)

so an expected loss is made. This strategy gives too much weight to higher levels of burning now that there are three players. Consequently the optimal strategy must give
less weight to higher values of burning so that the level of expected burning must match the expected winnings.

The probability distribution for the mixed strategy equilibrium when there are \( n \) players can be found as follows. Let the probability of beating one of the other competitors when \( B \) is burned be \( F(B) \). There are \( n - 1 \) other competitors, so the probability of beating them all is \([F(B)]^{n-1}\). The expected payoff in equilibrium must be zero, so \([F(B)]^{n-1} V = B\) for any value of \( B \) between 0 and \( V \). Solving this equation for \( F(B) \) gives the equilibrium probability distribution as

\[
F(B) = \left( \frac{B}{V} \right)^{1/(n-1)}. \tag{12.4}
\]

This distribution has the property that the probability applied to higher levels of \( B \) falls relative to that for lower levels as \( n \) increases. It can also be seen that when \( n = 2 \), it gives the solution found earlier.

What is important for the issue of rent-seeking is the expected amount burned by each competitor. Given that the expected payoff in equilibrium is zero and that everyone is equally likely to win \( V \) with probability \( \frac{1}{n} \), the expected amount burnt by each competitor is

\[
EB = \frac{V}{n}. \tag{12.5}
\]

By this result the expected amount burned by all the competitors is \( nEB = V \), which again is exactly equal to the prize being competed for. This finding is summarized as a theorem.

**Theorem 12.1** (Complete Dissipation Theorem) If there are two or more competitors in a deterministic rent-seeking game, the total expected value of resources expended by the competitors in seeking a prize of \( V \) is exactly \( V \).

The interpretation of this theorem is that between them the set of competitors will burn (in expected terms) a sum of money exactly equal to the value of the prize. The theorem is just a restatement of the fact that the expected payoff from the game is zero. The theorem has been very influential in the analysis of rent-seeking. Originally demonstrated in the context of monopoly (we will look at its application in this context later), the theorem provides the conclusion that from a social perspective there is nothing gained from the existence of the prize. Instead, all the possible benefits of the prize
are wasted through the burning of money. In the circumstances where it is applicable, the finding of complete dissipation provides an exact answer to the question of what quantity of resources is expended in rent-seeking.

It is important to note before proceeding that the theorem holds whatever the value of $n$ (provided it is at least 2). Early analyses of rent-seeking concluded that rents would be completely dissipated only if there were large numbers of competitors for the rent. This conclusion was founded on standard arguments that competition between many would drive the return down to zero. Prior to the proof of the Complete Dissipation Theorem it had been suspected that this would not be the case with only a small number of competitors and that some rent would be undissipated. However, the theorem proves that this reasoning is false and that even with only two competitors attempting to win the prize, rents are completely dissipated.

12.3.2 Probabilistic Game

The key feature of the Complete Dissipation Theorem is that it takes only a slight advantage over one’s competitors to obtain a sure win. This is the situation where the rent-seeking contest takes the form of a race or an auction with maximal competition. However, in many cases there is inevitably uncertainty in rent-seeking, so higher effort increases the probability of obtaining the prize but does not ensure a win. A natural application is political lobbying where lobbying expenditures involve real resources that seek to influence public decisions. Even if a lobby can increase its chance of success by spending more, it cannot obtain a sure win by simply spending more than its competitors. We now show that such uncertainty will reduce the equilibrium rent-seeking efforts, preventing full dissipation of the rent.

Consider modifying the payoff function to let the probability of anyone obtaining the prize be equal to their share of the total rent-seeking expenditures of all contestants,

$$EP_i = \frac{B_i}{B_i + [n - 1]B_{-i}} V - B_i,$$

(12.6)

where $[n - 1]B_{-i}$ is the total effort of the other contestants. So the expected payoff of contestant $i$ is the probability of obtaining the prize, which is their spending as a proportion of the total amount spent in the competition, times the value of the rent, $V$, minus their own spending.

A Nash equilibrium in this game is an expenditure level for each contestant such that no one would want to alter his expenditure given that of the other contestants.
Because all contestants are identical, we should expect a symmetric Nash equilibrium in which rent-seeking activities are the same for all and everyone is equally likely to win the prize. To find this Nash equilibrium, we proceed in two steps. First, we derive the optimal response of contestant \( i \) as a function of the total efforts of the other contestants. Second, we use the symmetry property to obtain the Nash equilibrium.

To find player \( i \)’s best response when the others are choosing \( B_{-i} \), we must take the derivative of player \( i \)’s expected payoff and set it equal to zero (this is the first-order condition). To facilitate the derivative, express the probability of winning as a power function in the expected payoff,

\[
EP_i = B_i[B_i + (n - 1)B_{-i}]^{-1} V - B_i.
\]  

(12.7)

Using the product rule for the derivative of the first term (the derivative of the first function times the second, plus the first function times the derivative of the second), we have the first-order condition

\[
[B_i + (n - 1)B_{-i}]^{-1} V - B_i[B_i + (n - 1)B_{-i}]^{-2} V - 1 = 0.
\]  

(12.8)

Next we use the fact that in a symmetric equilibrium \( B_i = B_{-i} = B \). Making this substitution in the first-order condition gives

\[
[B + (n - 1)B]^{-1} V - B[B + (n - 1)B]^{-2} V - 1 = 0,
\]  

(12.9)

or

\[
\]  

(12.10)

Finally, multiplying both sides by \( n^2 B \), we obtain

\[
nV - V = n^2 B.
\]  

(12.11)

Hence the equilibrium level of rent-seeking expenditure is

\[
B = \frac{n - 1}{n^2} V,
\]  

(12.12)

and the total expenditure of all contestants in equilibrium is

\[
nB = \frac{n - 1}{n} V.
\]  

(12.13)

Thus the fraction of the rent that is dissipated is \( \frac{n - 1}{n} < 1 \), which is an increasing function of the number of contestants. With two contestants only one-half of the rent
is dissipated in a Nash equilibrium, and the fraction increases to one as the number of contestants gets large. In equilibrium each contestant is equally likely to obtain the prize (with probability $\frac{1}{n}$), and using the equilibrium value of $B$, their expected payoff is $EP = \left[\frac{1}{n}\right] V - B = \frac{V}{n^2}$.

**Theorem 12.2** (Partial Dissipation Theorem) If there are two or more competitors in a probabilistic rent-seeking game, the total expected value of resources expended by the competitors in seeking a prize of $V$ is a fraction $\frac{n-1}{n}$ of the prize value $V$, and is increasing with the number of competitors.

It follows that the total costs of rent-seeking activity are significant, and are greater than one-half of the rent value in all cases. Notice that the rate of rent dissipation is independent of the value of the rent. It is also worth mentioning that in the Nash equilibrium contestants play a pure strategy and do not randomize as in the previous deterministic rent-seeking game. This is because the probability of obtaining the rent is a continuous function of the person’s own rent-seeking activity. Finally, in equilibrium, no single person spends more on rent-seeking than the prize is worth, but the total expenditure on rent-seeking activities may dissipate a substantial fraction of the prize value. This destruction of value is often innocuous because the contestants participate willingly expecting to gain. However, as in any competition where the winner takes all, there is only one winner who may earn large profits but many losers who bear the full cost of the destruction of value.

### 12.3.3 Free-Entry

Beginning with a fixed number of competitors does not capture the idea of a potential pool of competitors who may opt to enter the competition if there is a rent to be obtained. It is therefore of interest to consider what the equilibrium will be if there is free-entry into the competition. In the context of the game, free-entry means that competitors enter to bid for the prize until there is no expected benefit from further entry. This has the immediate implication that the expected payoff has to be driven to zero in any free-entry equilibrium.

How can the game be solved with free-entry? The analysis of the deterministic game showed that the expected payoff of each competitor is zero in the mixed strategy equilibrium. From this it follows that once at least two players have entered the competition, the expected payoff is zero. The free-entry equilibrium concept is therefore compatible
with any number of competitors greater than or equal to two, and all competitors who enter play the mixed strategy.

There is an important distinction between this equilibrium and the one considered for fixed numbers. In the former case it was assumed (but without being explicitly stated) that all competitors played the same strategy and only such symmetric equilibria were considered. If this is applied to the free-entry case, it means that the entire (unlimited) set of potential competitors must enter the game and play the mixed strategy given by (12.4) as \( n \to \infty \). An alternative to this cumbersome equilibrium is to consider an asymmetric equilibrium in which different competitors play different strategies. An asymmetric equilibrium of the game is for some competitors to choose not to enter while some (at least 2) enter and play the mixed strategy in (12.4). All competitors (both those who enter and those who do not) have an expected payoff of 0. The other important feature of the both the symmetric and asymmetric free-entry equilibria is that there is again complete dissipation of the rent. This finding is less surprising in this case than it is with no-entry, since the entry could be expected to reduce the net social value of the competition to zero.

In the probabilistic game contestants get a positive expected payoff from their rent-seeking activities of \( \frac{V}{n} \). Such a gain from rent-seeking will attract new contestants until the rent value is fully dissipated, that is, \( n \to \infty \) and \( \frac{V}{n} \to 0 \). So free-entry will make the two games equivalent with full dissipation of the rent.

### 12.3.4 Risk Aversion

The analysis so far has relied on the assumption that competitors for the prize care only about the expected amount of money with which they will leave the contest. This is a consequence of the assumption that they are risk neutral and hence indifferent about accepting a fair gamble. Although risk neutrality may be appropriate in some circumstances, such as for governments and large firms that can diversify risk, it is not usually felt to correctly describe the behavior of individual consumers. It is therefore worth reflecting on how the results are modified by the incorporation of risk aversion.

The first effect of risk aversion is that the expected monetary gain from entering the contest must be positive in order for a competitor to take part—this is the compensation required to induce the risk-averse competitors to take on risk. In terms of the deterministic game with a mixed strategy equilibrium, for a given number of competitors this means that less probability must be given to high levels of money burning and more to lower levels. However, the expected utility gain of the contest will be zero,
since competition will bid away any excess utility. In contrast to the outcome with risk neutrality, there will not be complete dissipation of the rent. This is a consequence of the expected monetary gain being positive, which implies that something must be left to be captured. With risk aversion the resources expended on rent-seeking will be strictly less than the value of the rent. But note carefully that this does not say that society has benefited. Since the expected utility gain of each competitor is zero, the availability of the rent still does not raise society’s welfare.

The same reasoning applies to the probabilistic game with more risk-averse individuals tending to expend less on rent-seeking activities. As a result a lower fraction of the rent will be dissipated. The effect of free-entry will be to drive the expected utility gain of each contender to zero.

### 12.3.5 Conclusions

This section has analyzed a simple game that can be interpreted as modeling the most basic of rent-seeking situations. The burning of money captures the use of resources in lobbying and the fact that these resources are not used productively. The fundamental conclusion is that when competitors are risk neutral, competition leads to the complete dissipation of the rent. This applies no matter how many competitors there are (provided there are at least two) and whether or not the number of competitors is fixed or variable. This fundamental conclusion of the rent-seeking literature shows that the existence of a rent does not benefit society, since resources (possibly equal in value to that rent) will be exhausted in capturing it. This conclusion has to be slightly modified with risk aversion. In this case there is less expenditure on rent-seeking and thus less rent dissipation. However, the expected utility gain of the competition is zero. In welfare terms, society does not benefit from the rent.

### 12.4 Social Cost of Monopoly

Monopoly is one of the causes of economic inefficiency. A monopolist restricts output below the competitive level in order to raise price and earn monopoly profits. This causes some consumer surplus to be turned into profit and some to become deadweight loss. Standard economic analysis views this deadweight loss to be the cost of monopoly power. The application of rent-seeking concepts suggests that the cost may actually be much greater.
Consider figure 12.1. This depicts a monopoly producing with constant marginal cost $c$ and no fixed costs. Its average revenue is denoted $AR$ and marginal revenue $MR$. The monopoly price and output are $p^m$ and $y^m$ respectively, while the competitive output would be $y^c$. Monopoly profit is the rectangle $\pi$, and deadweight loss the triangle $d$. In a static situation the deadweight loss $d$ is the standard measure of the cost of monopoly. (The emphasis on “static” is necessary here because there may be dynamic gains through innovation from the monopoly that offset the deadweight loss.)

How can the introduction of rent-seeking change this view of the cost of monopoly? There are two scenarios in which it can do so. First, the monopoly position could have been created by the government. An example would be the government deciding that an airline route can be served only by a single carrier. If airlines must then compete in lobbying for the right to fly this route the situation is just like the money-burning competition of section 12.3. The rent-seeking here comes from the bidders for the monopoly position. Another example is the allocation in the late 1980s by the US Federal Communications Commission of regional cell phone licenses. The lure of extremely high potential profits was strong enough to attract many contenders. There were about 320,000 contestants competing for 643 licenses. Hazlett and Michaels (1993) estimated the total cost of all applications (due to the technical expertise required) to be about $400$ million. Each winner earned very large profits well in excess of their application costs. However, the costs incurred by others were lost, and the total cost of
the allocation of licenses was estimated to be about 40 percent of the market value of
the license. Second, the monopoly may be already in existence but in a position where
it has to defend itself from potential competitors. Such defense could involve lawyers
or an effective lobbying presence attempting to prevent the production of similar goods
using copyright or patent law, or it could mean advertising to stifle competition. It could
even mean direct action to intimidate potential competitors.

Whichever case applies, the implications are the same. The value of having the
monopoly position is given by the area $\pi$. If there are a number of potential monopolists
bidding for the monopoly, then the analysis of money-burning can be applied to show
that if they are risk neutral, the entire value will be dissipated in lobbying. Alternatively,
if an incumbent monopolist is defending their position, they will expend resources up
to value $\pi$ to do so. In both cases the costs of rent-seeking will be $\pi$.

Combining these rent-seeking costs with the standard deadweight loss of monopoly,
the conclusion of the rent-seeking approach is that the total cost of the monopoly to
society is at least $d$ and may be as high as $\pi + d$. What determines the total cost is the
nature of the rent-seeking activity. We can conclude that resources of value $\pi$ will be
expended but not how much is actually wasted. As the discussion of section 12.2 noted,
some of the costs may be transfer payments (or, more simply, bribes) to officials. These
are not directly social costs but, again referring to section 12.2, may become so if they
induce rent-seeking in obtaining official positions. In contrast, if all the rent-seeking
costs are expended on unproductive activities, such as time spent lobbying, then the
total social cost of the monopoly is exactly $\pi + d$.

These results demonstrate one of the most basic insights of the rent-seeking literature:
the social costs of monopoly may be very much greater than measurement through
deadweight loss would suggest. To see the extent of the difference that this can make,
reconsider the measurements of welfare loss given in chapter 9. Harberger, using just the
deadweight loss $d$, calculated the cost of monopolization in US manufacturing industry
for the period 1924 to 1928 as equal to 0.08 percent of national income. In contrast,
the 1978 calculations by Cowling and Mueller followed the rent-seeking approach and
included the cost of advertising in the measure of welfare loss. Their analysis of US
industry concluded that welfare loss was between 4 and 13 percent of gross corporate
product. The difference between these measures reflects the additional loss through
rent-seeking.

This discussion of monopoly has shown that rent-seeking does have important impli-
cations. In particular, it strongly alters our assessment of the social costs of monopoly
and shows that the standard deadweight loss measure seriously understates the true
loss. This conclusion does not apply just to monopoly. Rent-seeking has the same
effect when applied to any distortionary government policy. This includes regulation, tariffs, taxes, and spending. It also shows that the net costs of a distortionary policy may be much higher than an analysis of benevolent government suggests. Attempts at quantifying the size of these effects show that they can be very dramatic.

12.5 Equilibrium Effects

The discussion of monopoly welfare loss in the previous section is an example of partial equilibrium analysis. It considered the monopolist in isolation and did not consider any potential spillovers into related markets nor the consequences of rent-seeking for the economy as a whole. This section will go some way toward remedying these omissions. The analysis here will be graphical; an algebraic development of similar arguments will be given in section 12.6.1.

Consider an economy that produces two goods and has a fixed supply of labor. The production possibility frontier depicting the possible combinations of output of the two goods is denoted by \(ppf\) in figure 12.2. The competitive equilibrium prices ratio \(p^c = \frac{p_1}{p_2}\) determines the gradient of the line tangent to the \(ppf\) at point \(a\). This will be the equilibrium for the competitive economy in the absence of lobbying.

The form of lobbying that we consider is for the monopolization of industry 1. If this lobbying is successful, it will have two effects. The first effect will be to change the relative prices in the economy. The second will be to use some labor in the lobbying process that could be used productively elsewhere. The consequences of these effects will now be traced on the production possibility diagram. Let the monopoly price for good 1 be given by \(p^m_1\) and the monopoly price ratio be denoted by \(p^m = \frac{p^m_1}{p_2}\). Since \(p^m > p^c\), the monopoly price line will be steeper than the competitive price line. This change in the relative prices will move the economy from point \(a\) to point \(b\) around the initial production possibility frontier (see figure 12.2). Evaluated at the competitive prices, the value of output has fallen (point \(b\) lies below the extension of the competitive price line).

The consequence of accounting for the labor used in lobbying is derived by observing that the labor of lobbyists produces neither good 1 nor good 2 but is effectively lost to the economy. This loss of labor reduces the potential output of the economy. Hence the production possibility frontier with lobbying must lie inside that without lobbying. This is shown in figure 12.3 where the production possibility frontier with lobbying is denoted \(ppf^L\). With the monopoly price line the equilibrium with both monopoly and lobbying will be at point \(c\) in figure 12.3. The outcome in figure 12.3 shows that the move to monopoly pricing shifts the equilibrium around the frontier.
Figure 12.2
Competitive and monopoly equilibria

Figure 12.3
Monopoly and lobbying
and lobbying shifts the frontier inward. The value at competitive prices of output at $a$ is higher than at $b$, and the value at $b$ is higher than at $c$. Hence successful lobbying has reduced the value of output by altering the price ratio and by causing an inward move of the production possibility frontier. At the aggregate level this is damaging for the economy. At the micro level there will be a transfer of income to the owners of the monopoly and the lobbyists, and away from the consumers, so the outcome is not necessarily bad for all individuals.

A further comparison that can be made is between the equilibrium with unsuccessful lobbying where the resource cost of lobbying is incurred but the prices remain at the competitive level (point $d$ in figure 12.4) and monopoly with no lobbying (point $b$). As panels a and b of figure 12.4 show, either outcome $b$ or $d$ could have the highest value of output when computed using the competitive prices. From this it can be concluded that there may be situations (as shown in figure 12.4b) where it is better to concede to the threat of lobbying and allow the monopoly (without the lobbying taking place) rather than refuse to concede to the lobby.

This section has extended the partial equilibrium analysis of lobbying to a general equilibrium setting to illustrate the combined effects of the distortions generated by successful lobbying and the waste of the resources used in the lobbying process. The switch from the competitive to the monopoly price reduces the value of output. Including lobbying moves the production possibility frontier inward. Moving the equilibrium onto this new frontier can lower the value of output even further.
12.6 Government Policy

Rent-seeking may be important for the study of private-sector monopoly, but most proponents of rent-seeking would see its application to government as being far more significant. Much analysis of policy choice views the government as benevolent and trying to make the best choices it can. The rent-seeking model of government is very different. This takes the view of the government as a creator of rents and those involved in government as seeking rent wherever possible. Chapter 5 touched on some of these issues in the discussion of bureaucracy, but that discussion can be extended much further.

There are two channels through which the government is connected with rent-seeking:

- **Lobbying**  
  We began this chapter by noting that the United States may have up to 100,000 professional lobbyists. These lobbyists attempt to change government policy in favor of the interests that employ them. If the lobbyists are successful, rents are created.

- **Bureaucrats and politicians**  
  Bureaucrats and politicians in government are able to create rents through their policy choices. These rents can be “sold” to the parties that benefit. Selling rents generates income for the seller and gives an incentive for careers to be made in politics and bureaucracy.

These two channels of rent-seeking are now discussed in turn.

12.6.1 Lobbying

The discussion so far has frequently referred to lobbying but without going into great detail about its economic effects. Section 12.4 showed graphically how the use of labor in lobbying shifted the production possibility frontier inward, but a graphical analysis of that kind could not provide an insight into the size of the effects. The purpose now is to analyze an example that can quantify the potential size of the economic loss resulting from the use of labor in lobbying.

Many of the implications of lobbying can be found by analyzing the use of productive labor to lobby for a tariff. The effect of a tariff is to make imports more expensive, so allowing the home firm to charge a higher price and earn greater profits. The potentially higher profit gives an incentive for lobbying. For example, the owners of textile firms will benefit from a lobby-induced tariff on imported clothing. Also the US steel industry
Chapter 12: Rent-Seeking

is a well-organized group and has long been active in encouraging tariffs on competing imports. The resources used for lobbying have a social value (equal to their productivity elsewhere in the economy), so the lobbying is not without cost. The calculations below will reveal the extent of this cost.

Consider a small economy in which two consumption goods are produced. In the absence of tariffs, the world prices of these commodities are both equal to 1, and the assumption that the economy is small means that it treats these prices as fixed. Some output is consumed and some is exported. A quantity $\bar{\ell}$ of labor is supplied inelastically by consumers. This is divided between production of the two goods and lobbying. Good 2 is produced with constant returns to scale, and one unit of labor produces one unit of output. This implies that the wage rate, $w$, must equal 1 (if it were higher, the firms would make a loss producing good 2; if it were lower, their profit would be unlimited since the price is fixed at the world level).

The cost function for the firm producing good 1 is assumed to be $C(y_1) = \frac{1}{2} y_1^2$, where $y_1$ is output. With a tariff $\tau$, which may be zero, the price of good 1 on the domestic market becomes $1 + \tau$. Assume that all of the output of the firm is sold on the domestic market. Then the profit level of the firm is

$$\pi_1(\tau) = y_1[1 + \tau] - \frac{1}{2} y_1^2. \quad (12.14)$$

Profit is maximized at output level

$$y_1 = 1 + \tau. \quad (12.15)$$

It can be seen from (12.15) that the output of good 1 is increasing in the value of the tariff. The monopolist therefore produces a higher output if it succeeds in obtaining tariff protection. The level of profit that results from this output is given by

$$\pi_1(\tau) = \frac{1}{2} [1 + \tau]^2. \quad (12.16)$$

so profit increases as the square of the tariff. This indicates the benefits that are obtained from protection.

Equilibrium on the labor market requires that labor supply must equal the use of labor in the production of good 1, $\ell_1$, plus that used in the production of good 2, $\ell_2$, plus that used for lobbying, $\ell_L$. Hence $\bar{\ell} = \ell_1 + \ell_2 + \ell_L$. The labor demand from firm 1 is $\ell_1 = \frac{1}{2} [1 + \tau]^2$. To determine the labor used in lobbying, the Complete Dissipation Theorem of section 12.3 is applied. That is, it is assumed that resources are used in lobbying up to the point at which the extra profit they generate is exactly equal to the
resource cost. Without lobbying, profit is $\pi_1(0)$. After a successful lobby with a tariff implemented, it becomes $\pi_1(\tau)$. The value of labor that the firm will devote to lobbying is therefore

$$\ell_L = \pi_1(\tau) - \pi_1(0) = \frac{1}{2} \left[ 2\tau + \tau^2 \right].$$

(12.17)

Hence the value of labor wasted in lobbying is increasing as the square of the tariff. Finally, since the production of each unit of good 2 requires one unit of labor, the output of good 2 equals the labor input into the production of that good, so $\ell_2 = y_2$ or

$$y_2 = \ell - \frac{1}{2} \left[ 1 + 4\tau + 2\tau^2 \right].$$

(12.18)

This shows that the output of good 2 is decreasing as the square of the tariff.

From these observations it can be judged that the rent-seeking is damaging the economy, since the production of good 2 falls at a faster rate than the production of good 1 increases. One method for quantifying the effect of this process is to determine the value of national output at world prices. World prices are used since these are the true measure of value. Doing this gives

$$y_1 + y_2 = \ell - \frac{1}{2} \left[ 2\tau^2 + 2\tau - 1 \right].$$

(12.19)

Hence national income is reduced at the rate of the tariff squared. The conclusion in (12.19) shows just how damaging rent-seeking can be. The possible availability of a tariff causes resources to be devoted to lobbying. These resources are withdrawn from the production of good 2 and national income, evaluated at world prices, declines.

12.6.2 Rent Creation

The analysis so far has focused primarily on the effects of rent-seeking in the presence of preexisting rents. We now turn to study the other side of the issue: the motives for the deliberate creation of rents. Such rent-creation is important because the existence of a rent implies a distortion in the economy. Hence the economic cost of a created rent is the total of the rent-seeking costs plus the cost of the economic distortion. This is the sum of deadweight loss plus profit identified in the section 12.4.

To be in a position to create rents requires the power to make policy decisions. In most political systems this authority is formally vested in politicians. Assuming that they are solely responsible for decision-making would, though, be shortsighted.
Politicians are advised and informed by bureaucrats. Many of the responsibilities for formulating policy options and for clarifying the vague policy notions of politicians are undertaken by bureaucrats. By carefully limiting the policies suggested or by choosing their advice carefully, a bureaucrat may well be able to wield implicit political power. It therefore cannot be judged in advance whether it is the politicians or the bureaucrats who actually make policy decisions. This does not matter unduly. For the purpose of the analysis, all that is necessary is that there is someone in a position to make decisions that can create rents, be it a politician or a bureaucrat. When the arguments apply to both politicians and bureaucrats, the generic term “policy maker” will be adopted.

How are rents created? To see this most clearly, consider an initial position where there is a uniform rate of corporation tax applicable to all industries. A rent can then be created by making it known that sufficient lobbying will be met by a reduction in the rate of tax. For instance, if the oil sector were to expend resources on lobbying, then it would be made a special case and permitted a lower corporation tax. The arguments already applied several times show that the oil sector will be willing to lobby up to a value equal to the benefit of the tax reduction. The creation of a monopoly airline route mentioned in section 12.4 is another example of rent-creation.

The reason for the rent-creation can now be made clear. By ensuring that the nature of the lobbying is in a form that they find beneficial, the policy maker will personally gain. Such benefits could take many forms, ranging from meals to gifts through to actual bribes in the form of cash payments. Contributions to campaign funds are an especially helpful form of lobbying for politicians, as are lucrative appointments after a term of office is completed. All of these forms of lobbying are observed to greater or lesser degrees in political systems across the world.

It has already been noted that this rent-creation leads to the economic costs of the associated rent-seeking. There are also further costs. Since there are rents to be gained from being a politician or a bureaucrat (the returns from the lobbying), there will be excessive resources devoted to securing these positions. Political office will be highly sought after with too many candidates spending too much money in seeking election. Bureaucratic positions will be valued far in excess of the contribution that bureaucrats make to economic welfare. Basically, if politicians or bureaucrats can earn rents, then this will generate its own rent-seeking as these positions are competed for. In short, the ability to create rents has cumulative effects throughout the system. The Complete Dissipation Theorem can again be applied here: in expected terms these rents are dissipated. It is important to bear in mind that the winner of the rent does gain: the politician who is elected or the bureaucrat who secures their position will personally benefit from the rent. Losses arise for those who competed but failed to win.
Two further effects arise. First, there will be an excessive number of distortions introduced into the economy. Distortions will be created until there is no further potential for the decision maker to extract additional benefits from lobbyists. Second, there will be an excessive number of changes in policy. Decision makers will constantly seek new methods of creating rents and this will involve policy being continually revised. One simple way for a new policy maker to obtain rents is to make tax rates uniform with a broad base on appointment, and then gradually auction off exemptions throughout the term of office. The broader the chosen base, the greater the number of exemptions that can be sold.

12.6.3 Conclusions

The discussion of this section has presented a very negative view of government and economic policy-making. The rent-seeking perspective argues that decisions are not made for reasons of economic efficiency but on the basis of how much can be earned for making them. As a result the economy is damaged by inefficient and distortionary policies. In addition resources are wasted in the process of rent-seeking. Both lobbying and attempting to obtain decision-making positions waste resources. When these are combined, the damage to the economy is significant. It suggests that political power is sought after not as an end in itself but simply as a means to access rent.

12.7 Informative Lobbying

The discussion so far has presented a picture of lobbyists as a group who contribute nothing to the economy and are just a source of welfare loss. To provide some balance, it is important to note that circumstances can arise where lobbyists do make a positive contribution. Lobbyists may be able to benefit the economy if they, or the interest groups they represent, have superior information about the policy environment than the policy maker. By transmitting this information to the policy maker, they can assist in the choice of a better policy.

Several issues arise in this process of information transmission. To provide a simple description of these, assume that a policy has to be chosen for the next economic period. At the time the policy has to be chosen, the policy maker is uncertain about the future economic environment. This uncertainty is modeled by assuming that the environment can be described by one of several alternative “states of the world.” Here a state of the world is a summary of all relevant economic information. The policy maker knows
that different states of the world require different policy choices to be made, but they
do not know the future state of the world. Without additional information, the policy
maker would have to base policy choice upon some prior beliefs about the probability
of alternative states of the world. Unfortunately, if the chosen policy is not correct for
the state of the world that is realized, welfare will not be maximized.

Now assume that there is a lobbyist who knows which state of the world will occur.
It seems that if the lobbyist were just to report this information to the policy maker, then
the correct policy would be chosen and welfare maximized. But this misses the most
important point: the objectives of the lobbyist. If the lobbyist had the same preferences
as the policy maker, there would be no problem. The policy maker would accept the
information that was offered knowing that the lobbyist was pursuing the same ends.
In contrast, if lobbyists have different preferences, then they may have an incentive
to reveal false information about the future state of the world with the intention of
distorting policy in a direction that they find beneficial. Therefore the policy maker
faces the problem of determining when the information they receive from lobbyists is
credible and correct, and when it represents a distortion of the truth.

To see how these issues are resolved, consider a model where there are only two
possible values for the future state of the world. Let these values be denoted $\theta_h$ and $\theta_l$
with $\theta_l < \theta_h$, where we term $\theta_h$ the “high state” and $\theta_l$ is the “low state.” The policy
maker seeks to maximize a social welfare function that depends on the state of the
world and the policy choice, $\pi$. Suppose that this objective function takes the form

$$ W(\pi, \theta) = - [\pi - \theta]^2, \quad (12.20) $$

which implies that welfare loss is minimized when the policy choice is adapted to the
state of the world. If the policy maker had perfect information, then when the state was
known to be high, a high policy level $\pi_h = \theta_H$ would be chosen. In contrast, when the
state was known to be low, a low policy level $\pi_l = \theta_l$ would be chosen. Now assume
that the policy maker is uninformed about the state of the world and initially regards
the two states as equally likely. In this case the policy maker will choose a policy based
on the expected state of the world, so

$$ \pi_e = \frac{\theta_l + \theta_h}{2}. \quad (12.21) $$

That is, the policy maker sets the policy equal to the expected value of $\theta$.

Now we introduce a lobbyist who knows what the state of the world will be. The
welfare of the lobbyist also depends on the policy level and the state of the world.
However, the lobbyist does not share the same view as the policy maker about the ideal
policy level in each state. We assume that the ideal policy for the lobbyist exceeds the ideal policy of the policy maker by an amount $\Delta$ in both states of the world. We can refer to such a difference in the ideal policy as the *extent of the disagreement* between the policy maker and the lobbyist. Such a lack of agreement can be obtained by adopting preferences for the lobbyist given by

$$U(\pi, \theta) = -[\pi - \theta - \Delta]^2.$$  \hspace{1cm} (12.22)

To find the conditions under which the lobbyist can credibly transmit information about the state of the world, we must investigate the incentives the lobbyist has to truthfully report the state of the world. The lobbyist can only report either $\theta_h$ or $\theta_\ell$, and if he is trusted by the policy maker, the policy choice will be, respectively, $\pi_h$ or $\pi_\ell$. If the true state is $\theta_h$ the lobbyist has no incentive to misreport the information. Indeed the lobbyist has a bias toward a high policy level; misreporting the state as being low would lead to a policy $\pi_\ell$, which is worse than the lobbyist’s ideal policy of $\pi_h + \Delta$ when the state is high. In contrast, if the state is $\theta_\ell$, the lobbyist has a potential incentive to misreport because a truthful report, if trusted by the policy maker, leads to a policy level $\pi_\ell$ that is below the ideal policy of the lobbyist $\pi_\ell + \Delta$. The lobbyist may prefer to claim that the state is high to exploit the trust and obtain policy $\pi_h$ instead of $\pi_\ell$. However, it may be that $\pi_h$ is too large for the lobbyist when the state is $\theta_\ell$, in which case the lobbyist will report truthfully. The latter is the case where $\pi_\ell$ is closer to the ideal policy of the lobbyist in the low state than $\pi_h$, which occurs if the following inequality is satisfied:

$$[\theta_\ell + \Delta] - \theta_\ell \leq \theta_h - [\theta_\ell + \Delta].$$  \hspace{1cm} (12.23)

This inequality reduces to

$$\Delta \leq \frac{\theta_h - \theta_\ell}{2}.$$  \hspace{1cm} (12.24)

This condition says that policy maker can expect the lobbyist to truthfully report the state of the world if the extent of the disagreement is not too large. The equilibrium that results is fully revealing because the lobbyist can credibly transmit information about the state of the world. Lobbying is then informative and desirable for the society. If, in contrast, the inequality above is not satisfied, the extent of the disagreement is too large for the policy maker to expect truthful reporting when the state is low. The lobbyist’s report lacks credibility because the policy maker knows that the lobbyist prefers reporting the high state no matter what the true state happens to be. The report
is then uninformative, and the policy maker will rightly ignore it. So the policy maker sets the policy equal to the expected value \( \pi^e = \frac{\theta_i + \theta_{i+1}}{2} \). This policy choice is suboptimal for society because it is too large when the state is low and too small when the state is high. Note that the lobbyist is also worse off with the uninformative outcome because the policy choice \( \pi^e \) is smaller than their ideal policy \( \pi_h + \Delta \) when the state is high.

The problem of securing credibility is magnified when there are more than two states of the world. As the number of possible states increases, honest information revelation becomes ever more difficult to obtain. This is easily demonstrated. For a lobbyist to credibly report the true state, \( \Delta \) must be smaller than one-half of the distance between any two adjacent states—this is the content of (12.24). With \( n \) states, \( \theta_1 < \ldots < \theta_i < \ldots < \theta_n \), the conditions for truth-telling are for all \( i = 1, \ldots, n-1, \)

\[
\Delta \leq \frac{\theta_{i+1} - \theta_i}{2}.
\]

Evidently, as the number of states grows, intermediate states are added, and this reduces the distance between any two states. Eventually the states become too close to each other for the lobbyist to be able to credibly communicate the true state, even if \( \Delta \) is small. Full revelation is then impossible. What can the lobbyist do in such a situation? The answer is to reveal partial information, as pointed out by Crawford and Sobel (1982).

Suppose that the states are partitioned into two intervals, \( L = (\theta_1, \ldots, \theta_i) \) and \( H = (\theta_{i+1}, \ldots, \theta_n) \). Then the lobbyist can report the interval in which the true state falls, instead of reporting the precise state—we term this partial revelation. If he reports \( \Theta_L \), then it means that \( \theta_1 \leq \theta \leq \theta_i \). If all states are equally likely and equally spaced, then a trusty policy maker sets the policy equal to the expected value on this interval \( \pi(L) = \frac{\theta_1 + \theta_i}{2} \). Similarly the report \( \Theta_H \) induces a policy choice \( \pi(H) = \frac{\theta_{i+1} + \theta_n}{2} \). The question is whether the lobbyist has any incentive to lie.

Among the states in the interval \( L \), the greatest temptation to lie (by reporting \( \Theta_H \)) is when the true state is close to \( \theta_i \); if the lobbyist does not want to claim \( H \) when \( \theta = \theta_i \), then he will not wish to do so when \( \theta < \theta_i \), since this would push the policy choice further away from his ideal policy. Hence we can restrict attention to the incentive to report truthfully \( L \) when the true state is \( \theta_i \). Truthful reporting induces policy \( \pi(L) \) and misreporting induces policy \( \pi(H) \). The lobbyist will report truthfully if the former policy is closer than the latter to his ideal policy \( \theta_i + \Delta \) given the true state \( \theta_i \). This is the case if
[θ_i + Δ] − π(L) ≤ π(H) − [θ_i + Δ], \hspace{1cm} (12.26)

which reduces to

\[ θ_i + Δ ≤ \frac{π(H) + π(L)}{2}. \] \hspace{1cm} (12.27)

Now suppose that θ actually is in H. We must check the incentive of the lobbyist to truthfully report H instead of L. The temptation to misreport is highest when the true state is close to θ_i + 1. In such a case, to sustain truthful reporting, it is required that the lobbyist induce a policy π(H) that is closer to the ideal policy θ_i + 1 + Δ than the policy that would be induced by misreporting π(L). That is,

[θ_i + 1 + Δ] − π(L) ≥ π(H) − [θ_i + 1 + Δ], \hspace{1cm} (12.28)

which reduces to

\[ θ_i + 1 + Δ ≥ \frac{π(H) + π(L)}{2}. \] \hspace{1cm} (12.29)

Combining the two incentive constraints (12.27) and (12.29), truth-telling requires that

\[ \frac{π(H) + π(L)}{2} − θ_{i+1} ≤ Δ ≤ \frac{π(H) + π(L)}{2} − θ_i. \] \hspace{1cm} (12.30)

This condition puts both a lower bound and an upper bound on the extent of the disagreement for the lobbyist to be able to communicate credibly partial information about the state to the policy maker.

The outcome of this analysis is that lobbyists can raise welfare if they are able to credibly report information to the policy maker. Unfortunately, this argument is limited by the potential incentive to report false information when there is divergence between the preferences of the lobbyist and the preferences of the policy maker. With a limited number of states, credible correct transmission can be sustained if the divergence is not too great. However, as the number of states of the world increases, credible transmission cannot be sustained if there is any divergence at all in preferences. In this latter case, though, it is possible to have partial information credibly released—again, provided that the divergence is limited. In conclusion, informed lobbyists can be beneficial through the advice they can offer a policy maker, but this can be undermined by their incentives to reveal false information.
12.8 Controlling Rent-Seeking

Much has been made of the economic cost of rent-seeking. These insights are interesting (and also depressing for those who may believe in benevolent government) but are of little value unless they suggest methods of controlling the phenomenon. This section gathers together a number of proposals that have been made in this respect. There are two channels through which rent-seeking can be controlled. The first channel is to limit the efforts that can be put into rent-seeking. The second is to restrict the process of rent-creation.

Beginning with the latter, rent-creation relies on the unequal treatment of economic agents. For instance, the creation of a monopoly is based on one economic agent being given the right to operate in the market and all other agents being denied. Equally, offering a tax concession for one industry treats the agents in that industry more favorably than those outside. Consequently a first step in controlling rent-seeking is to disallow policies that discriminate among economic agents. Restricting the policy maker to the implementation of nondiscriminatory policies would eliminate the creation of tax breaks for special interests or the imposition of tariffs on particular imports. If restricted in this way, the decision maker cannot auction off rents—if all parties gain, none has the incentive to pay.

The drawback of a rule preventing discrimination is that it is sometimes economically efficient to discriminate. For example, the theory of optimal commodity taxation (see chapter 15) describes circumstances where it is efficient for necessities to be taxed more heavily than luxuries. This would not be possible with nondiscrimination because the industries producing necessities would have grounds for complaint. Similarly the theory of income taxation (see chapter 16) finds that, in general, it is optimal to have a marginal rate of income tax that is not uniform. If implemented, the taxpayers facing a higher marginal rate would have ground for alleging discrimination. Hence a nondiscrimination ruling would result in uniform commodity and income taxes. These would not usually be efficient, so there would be a trade-off between economic losses through restrictions on feasible policy choices and losses through rent-seeking. It is not unlikely that the latter will outweigh the former.

There are other ways in which the process of rent-seeking can be lessened, but all of these are weaker than a nondiscrimination rule. These primarily focus on ensuring that the policy-making process is as transparent as possible. Among them would be policies such as limiting campaign budgets, insisting on the revelation of names of donors, requiring registration of lobbyists, regulating and limiting gifts, and reviewing
bureaucratic decisions. Policing can be improved to lessen the use of bribes. Unlike nondiscrimination, none of these policies has any economic implications other than their direct effect on rent-seeking.

12.9 Conclusions

Lobbyists are very numerous in number; they are also engaged in an activity that is not productive. The theory of rent-seeking provides an explanation for this apparent paradox and looks at the consequences for the economy. The fundamental insight of the literature is the Complete Dissipation Theorem: competition for a rent will result in resources being expended up until the expected gain of society from the existence of the rent is zero. If competitors for the rent are risk-neutral, this implies that the resources used in rent-seeking are exactly equal in value to the size of the rent. The application of these rent-seeking ideas show that the losses caused by distortions are potentially much larger than the standard measure of deadweight loss.

The other aspect of rent-seeking is that economic policy makers have an incentive to create distortions. They do this in order to receive benefits from the resulting rent-seeking. This leads to a perspective of policy driven not by what is good for the economy but by what the policy maker can get out of it and of a politics corrupted by self-interest. If this view is the correct description of the policy-making process, the response should be to limit the discretion for policy makers. Last, lobbying can be desirable when the lobbyists have better information about a policy-relevant variable than the policy maker. The question is then how the lobbyists can credibly communicate this information when there is some disagreement about the ideal policy choice.

Further Reading

The classic analysis of rent-seeking is in:

The second article is reprinted in:

This book also contains other interesting reading.
Chapter 12: Rent-Seeking

For more discussion of the definition of rent-seeking and a survey of the literature see:
The complete analysis of the rent-seeking game is in:
An alternative and very simple treatment of the rent-seeking game as an aggregative game is in:
Estimates of the social costs of monopoly are taken from:
Another important paper in this area is:
More on interest groups and lobbying can be found in:
A debate about the relative merits of rent-seeking and the traditional public finance approach is found in:

Exercises

12.1 One country invades another to create a demand for its construction industry. Is this rent-seeking?
12.2 IBM assembled its first personal computers from standard components to lower the cost. Was this rent-seeking?
12.3 A computer software company refuses to release its code to other developers. Is this rent-seeking?
12.4 Explain why “rent-seeking” by special interest groups is considered wasteful by most economists, while “profit-seeking” in the marketplace is not considered wasteful. What’s the difference? How does this relate to Tullock’s definition of rent seeking?
12.5 Construct a variation of the rent-seeking game without the discontinuity in winning/losing. Find the pure strategy equilibrium.

12.6 If demand is linear, show that profit and monopoly deadweight loss are related by $\pi = 2d$. Hence contrast the total loss with rent-seeking to the deadweight loss.

12.7 Should advertising be banned?

12.8 Does the observation that profit is positive show that the rent-seeking argument does not apply?

12.9 Using figure 12.3, display monopoly welfare loss (measured in units of good 2) in the diagram. Show that this increases as the monopoly price relative to the competitive price gets higher.

12.10 You are competing for a rent with one rival. Your valuation and your competitor’s valuation are private information. You believe that the other bidder’s valuation is equally likely to lie anywhere in the interval between $0$ and $5,000$. Your own valuation is $2,000$. Suppose that you expect that your rival will submit a bid that is exactly one-half of his valuation. Thus you believe that your rival is equally likely to bid anywhere between $0$ and $2,500$ (depending on the realized valuation between $0$ and $5,000$).

a. Show that if you submit a bid of $B$, the probability that you win the contest is the probability that your bid $B$ will exceed your rival’s bid, and that this probability of winning is $B/250$.

b. Your expected profit from bidding $B$ is $[200 - B] \times \text{Probability of winning}$. Show that the profit-maximizing strategy consists of bidding half your valuation.

12.11 Three firms have applied for the franchise to operate the cable TV system during the coming year. The annual cost of operating the system is $250$ and the demand curve for its services is $P = 500 - Q$, where $P$ is the price per subscriber per year and $Q$ is the expected number of subscribers. The franchise is assigned for only one year, and it allows the firm with the franchise to charge whatever price it chooses. The government will choose the applicant that spends the most money lobbying the government members. If the applicants cannot collude, how much will each spend on lobbying? (Hint: The winner will set the monopoly price for the service.)

12.12 (Rent-seeking contest with nonlinear technology.) There are $n$ players $i = 1, \ldots, n$, ranked by increasing order of initial income $M_i$. Each player $i$ spends $x_i$ to compete for a rent of size $R$. Every player is risk-neutral and maximizes his expected income. The probability to obtain the rent is $p_i = \frac{f(x_i)}{\sum_j f(x_j)} = \frac{x_i^\alpha}{\sum_j x_j^\alpha}$.

a. Write down the payoff function of player $i$.

b. Use the following transformation $z_i = x_i^\alpha$ to compute the best response of player $i$. How does a change in $\alpha$ and $M_i$ affect the best-response function? Discuss.

c. Calculate the Nash equilibrium. How does it depend on $\alpha$ and the distribution of income $M_1 < \ldots < M_i < \ldots < M_n$?

12.13 In exercise 12.11 the rent goes to the firm with the highest lobbying activity, and it takes only a small advantage to obtain a sure win. Now suppose that a higher lobbying activity increases
the probability of getting the rent but does not ensure a win. If firm $i$ spends the amount $x_i$ on lobbying activity, it will get the franchise with probability $p_i = \frac{x_i}{\sum_{j=1}^{3} x_j}$.

a. What is the optimal spending of firm $i$ in response to the total spending of the two other firms $x_{-i} = \sum_{j \neq i} x_j$? Draw the best-response function of firm $i$ to $x_{-i}$.

b. Suppose a symmetric equilibrium in lobbying where $x_i = x^*$ for all $i$. How much will each firm will spend on lobbying?

c. How does your answer change if there are $N$ extra firms competing for the franchise (assuming again all firms are identical)?

12.14 Consider a rent-seeking game with $N \geq 2$ contestants. The effort for person $i$ is denoted by $x_i$ for $i = 1, \ldots, N$. The cost per unit of effort is $C$. All contestants are identical. They value the rent at $V$, and each contestant can win the prize with a probability equal to their effort relative to the total effort of all contestants. Thus the payoff function of person $i$ exerts effort $x_i$ is given by $U_i = \left( \frac{x_i}{X} \right) V - Cx_i$, where $X = \sum x_j$. Note that the cost must be paid whether or not the prize is obtained. A Nash equilibrium is a lobbying effort for each contestant such that no one would want to alter their expenditure given that of the other competitors.

a. Find the derivative of player $i$’s expected utility and then set it equal to zero. Draw the resulting best-response function for player $i$ when the $N - 1$ each of the others chooses $x$.

b. In a symmetric equilibrium it must be the case that all contestants choose the same effort $x_i = x^*$. Using this symmetry condition and the best-response function in part a, show that the equilibrium outcome is $x^* = \frac{N-1}{N^2} \frac{V}{C}$ and that the fraction of the rent that is dissipated (i.e., total cost relative to the value of the prize) is an increasing function of $N$.

c. Suppose that there are four contenders ($N = 4$), the value of the prize is $V = 20,000$, and the cost of effort is $C = 5,000$. What is the equilibrium level of rent-seeking activity $x^*$? What is the fraction of rent dissipation?

d. Suppose that the cost of rent-seeking effort reduces from 5,000 to 2,500 with four competitors and a prize of 20,000. What is the impact on the common equilibrium level of rent-seeking activity? Does it affect the fraction of rent dissipation? Why or why not?

12.15 There is a given rent of $R$. Each of two players spends resources competing for the rent. If player 1 spends $x_1$, the probability that he wins the rent is $p_1 = \frac{\phi x_1}{x_1 + x_2}$, where $\phi > 1$.

a. What is the optimal spending of player 1 in response to a given spending level of player 2? What is the optimal response of player 2 to player 1? Draw the best-response functions. Discuss the effect of changing $\phi$ on the function.

b. How much will each player spend on lobbying? Which player is more likely to win the rent in equilibrium?

c. Compare the total equilibrium spending for $\phi > 1$ and $\phi = 1$. Should we expect more spending in rent-seeking activities when players are identical? Why or why not?

12.16 Consider the following situation. There are $N < 2$ players competing for a chance to win benefits from the government of $R$. The rent is given to the highest bidder. The second-highest
bidder gets nothing but must also spend the amount he bids. What is the likely outcome of such a situation? Where will the process stop? Is it possible that the first- and second-highest bidder could together bid more than the value of the rent? Could each of them spend more than the value of the rent? Why or why not?

12.17 (Coate and Morris 1995) This exercise is about rent-seeking in politics. The Tullock model develops the common idea that rent-seeking leads to inefficiency. In terms of policy-making, it is less clear that rent-seeking will lead to inefficient policy simply because the policy maker is maximizing his own utility (and the one of some special interest group). Surely any policy change will reduce the welfare of the policy maker (and that of the interest group), so the policy cannot be inefficient in the Pareto sense. However, Tullock argued that politicians would use disguised transfer mechanisms that are inefficient relative to cash transfers. Could you elaborate on this idea to demonstrate that excessive pork-barrel spending could reflect inefficient forms of redistribution to special interest groups?

12.18 Consider the informative lobbying model with only two states of nature \( \theta_1 < \theta_2 \). It was shown in the chapter that truth-telling by the lobbyist requires some upper bound on the extent of the disagreement \( \Delta > 0 \) (positive bias) between the lobbyist and the policy maker. The upper bound was given by half the distance between the two possible states, that is, \( \Delta \leq \frac{\theta_2 - \theta_1}{2} \). Now suppose that a second lobbyist can be consulted by the policy maker. This second lobbyist knows the true state (i.e., both lobbyists have the same information). But she has opposite preference to the other lobbyist, namely a negative bias in policy choice \( -\Delta \). So under state \( \theta_1 \) she prefers policy \( \theta_1 - \Delta \), whereas the other lobbyist prefers \( \theta_1 + \Delta \). Each lobbyist reports independently either the low state or the high state.

a. What are policy choices induced by the lobbyists’ reports?

b. Given the policy response to different pair of reports, what is the optimal report for each lobbyist?

c. What is the condition for each lobbyists to credibly report the true state?

d. Compare with the condition for truthful report with a single lobbyist. Explain why the problem of truthful report is exacerbated with two opposite lobbyists. Discuss.

12.19 (Baland and Ray) Suppose that a group of \( n \) agents are engaged in the joint production of a particular output. The production function is \( Y = F(e) \), where \( e = (e_1, \ldots, e_n) \) is a (nonnegative) vector of efforts. Suppose that the output of the joint production is distributed according to a fixed sharing vector ranked by increasing order \( \lambda = (\lambda_1, \ldots, \lambda_n) \). The payoff function of agent \( i \) is

\[
U_i(e_i, e_{-i}) = \lambda_i F(e_i, e_{-i}) - e_i,
\]

where \( e_{-i} \) stands for the vector \( e \) without the element \( e_i \). The efficient vector of effort levels maximizes

\[
F(e) = \sum_{i=1}^{n} e_i.
\]

Assume that \( F(e) \) is an increasing concave function satisfying the Inada end-point conditions \( F'(0) = \infty \) and \( F'(|\infty|) = 0 \) so that the maximization problem is well defined.
a. (Perfect substitutes) Suppose that $F$ is an increasing concave function of the sum of efforts $F(\sum_i e_i)$. Show that the equilibrium outcome is inefficient with too little effort.

b. Define the index of inequality as the highest share $\lambda_n \in \left[\frac{1}{n}, 1\right]$ and show that increasing inequality raises efficiency with the limit result of full efficiency for $\lambda_n = 1$.

c. (Perfect complements) Suppose that $F$ is an increasing concave function of the lowest effort in the group $F(\min_i \{e_i\})$. Show that the equilibrium outcome is inefficient with too little effort, and show that reducing inequality (defined as before by $\lambda_n \in \left[\frac{1}{n}, 1\right]$) raises efficiency, but that full efficiency cannot be achieved in the limit with $\lambda_n = \frac{1}{n}$. 
EQUITY AND DISTRIBUTION
13.1 Introduction

On April 17, 1975, the Khmer Rouge seized power in Cambodia. Pol Pot began to implement his vision of Year Zero in which all inequalities—of class, money, education, and religion—would be eliminated. Driven by their desire to achieve what they perceived as the social optimum, the Khmer Rouge attempted to engineer a return to a peasant economy. In the process they slaughtered an estimated two million people, approximately one-quarter of Cambodia’s population. The actions of the Khmer Rouge are an extreme example of the pursuit of equality and the willingness to accept an immense loss in order to achieve it. In normal circumstances governments impose a limit on the cost they are willing to pay for an improvement in equality.

When it comes to the efficiency/equity trade-off the Second Theorem of Welfare Economics has very strong policy implications. These were touched on in chapter 2 but were not developed in detail at that point. This was because the primary value of the theorem is what it says about issues of distribution. To fully appreciate the Second Theorem, it is necessary to view it from an equity perspective and to assess it in the light of its distributional implications.

This chapter will begin by investigating the implications of the Second Theorem for economic policy. This is undertaken on the premise that a social planner is able to make judgments between different allocations of utility. The concept of an optimal allocation is developed and the Second Theorem is employed to show how this can be achieved. Once this analysis has been accomplished, questions are raised about the applicability of lump-sum taxes and the value of Pareto-efficiency as a criterion for social decision-making. This provides a basis for re-assessing the interpretation of the First Theorem of Welfare Economics.

The major deficiency of Pareto-efficiency is identified as its inability to trade utility gains for one consumer against losses for another. This is important since most policy changes will involve some people gaining while other people lose. To proceed further, the informational basis for making welfare comparisons has to be addressed. We describe different forms of utility and different degrees of comparability of utility among consumers. These concepts are then related to Arrow’s Impossibility Theorem and the potential for constructing a social welfare function.
13.2 Social Optimality

The importance of the Second Theorem of Welfare Economics for policy analysis is very easily explained. In designing economic policy, a policy maker will always aim to achieve a Pareto-efficient allocation. If an allocation that was not Pareto-efficient was selected, then it would be possible to raise the welfare of at least one consumer without harming any other. It is hard to imagine why any policy maker would want to leave such gains unexploited. If it is presumed that this argument is correct, the set of allocations from which a policy maker will choose reduces to the Pareto-efficient allocations.

Suppose that a particular Pareto-efficient allocation has been selected as the policy maker’s preferred outcome. The Second Theorem shows that this allocation can be achieved by making the economy competitive and providing each consumer with the level of income needed to purchase the consumption bundle assigned to him in the chosen allocation. The consumers will then trade, and the chosen equilibrium will emerge as the competitive equilibrium. This is the process of *decentralization*. In achieving the decentralization of the allocation, only two policy tools are employed: the encouragement of competition and a set of lump-sum taxes to ensure that each consumer has the required income. If this approach could be applied in practice, then economic policy analysis would reduce to the formulation of a set of rules that guarantee competition and the calculation and redistribution of the lump-sum taxes. The subject matter of public economics, and economic policy, in general, would then be closed.

Looking at this process in detail, the first point that arises is the question of selecting the most preferred allocation. There are a number of ways to imagine this being done. An obvious one would be to consider voting, either over the alternative allocations directly or else for the election of a body (a “government”), to make the choice. Alternatively, the consumers could agree for it to be chosen at random or else they might hold unanimous views, perhaps via conceptions of fairness, about what the outcome should be. The method that is considered here is to assume that there is a social planner (which could be the elected government). This planner forms social preferences over the alternative allocations by taking into account the utility levels of the consumers. The most preferred allocation according to the social preferences is the one that is chosen.

To see how this method functions, consider the set of Pareto-efficient allocations described by the contract curve in the left-hand part of figure 13.1. Each point on the contract curve is associated with an indifference curve for consumer 1 and an indifference curve for consumer 2. These indifference curves correspond to a pair of
utility levels $\{U^1, U^2\}$ for the two consumers. As the move is made from the southwest corner of the Edgeworth box to the northeast corner, the utility of consumer 1 rises and that of 2 falls. These utility levels can be plotted by observing that each pair of utility levels on the contract curve can be represented as a point in utility space. The loci formed by these points is usually called the utility possibility frontier. This is shown in the right-hand panel of figure 13.1 where the utility values corresponding to the points $a$, $b$, and $c$ are plotted. Points such as $a$ and $b$ lie on the frontier: they are Pareto-efficient, so it is not possible to raise both consumers’ utilities simultaneously. Point $c$ is off the contract curve and is inefficient according to the Pareto criterion. It therefore lies inside the utility possibility frontier.

The utility possibility frontier describes the Pareto-efficient options from which the social planner will choose. It is now necessary to describe how the choice is made. To do this, it is assumed that the social planner measures the welfare of society by aggregating the individual consumers’ welfare levels. Given the pair of welfare levels $\{U^1, U^2\}$, the function determining the aggregate level of welfare is denoted by $W(U^1, U^2)$. This is termed a Bergson–Samuelson social welfare function. Basically, given individual levels of happiness, it imputes a social level of happiness. Embodied within it are the equity considerations of the planner. Two examples of social welfare functions are the utilitarian $W = U^1 + U^2$ and the Rawlsian (or maxi-min) $W = \min\{U^1, U^2\}$. The social indifference curves for these welfare functions are illustrated in figure 13.2, alongside those for an “intermediate” social welfare function. These curves show combinations of the two consumers’ utilities that give a constant level of social welfare. The view on equity taken by the social planner translates into their willingness to trade off the
utility of one consumer against the utility of the other. This determines the shape of the indifference curves. From the shape of the indifference curves it can be seen that the utilitarian and Rawlsian social welfare functions represent two extremes. The utility of one consumer can be substituted perfectly for that of another with the utilitarian social welfare function, but no substitution is possible for the Rawlsian. The intermediate case allows imperfect substitution.

Given the welfare function, the social planner considers the attainable allocations of utility described by the contract curve and chooses the one that provides the highest level of social welfare. Indifference curves of the welfare function can be drawn as in figure 13.3. The social planner then selects the outcome that achieves the highest indifference curve. This optimal point on the utility possibility locus, denoted by point \( o \), can then be traced back to an allocation in the Edgeworth box. This allocation represents the socially optimal division of resources for the economy given the preferences captured by the social welfare function. If these preferences were to change, so would the optimal allocation.

Having chosen the socially optimal allocation, the reasoning of the Second Theorem is applied. Lump-sum taxes are imposed to ensure that the incomes of the consumers are sufficient to allow them to purchase their allocation conforming to point \( o \). Competitive economic trading then takes place. The chosen socially optimal allocation is achieved through trade as the equilibrium of the competitive economy. This process is called decentralization because the allocation is achieved as a consequence of individuals making optimizing decisions rather than the social planner imposing the allocation.
Chapter 13: Optimality and Comparability

The decentralization argument shows that the use of the Second Theorem allows the economy to achieve the outcome most preferred by its social planner. Given the economy’s limited initial stock of resources, the socially optimal allocation reaches the best trade-off between efficiency and equity as measured by the social welfare function. In this way the application of the Second Theorem can be said to solve the economic problem, since the issues of both efficiency and equity are resolved to the greatest extent possible and there is no better outcome attainable. Clearly, if this reasoning is applicable, all that a policy maker has to do is choose the allocation, implement the required lump-sum taxes, and ensure that the economy is competitive. No further policy or action is required. Once the incomes are set, the economy will take itself to the optimal outcome.

13.3 Lump-Sum Taxes

The role of lump-sum taxes has been made very explicit in describing the application of the Second Theorem. In the economic environment envisaged, lump-sum taxes are the only tool of policy that is required beyond an active competition policy. To justify the use of policies other than lump-sum taxes, it must be established that such taxes are either not feasible or else are restricted in the way in which they can be employed. This is the purpose of the next two sections. The results described are important in their own right, but they also provide important insights into the design of other forms of taxation.
In order for a tax to be lump sum, the consumer on whom the tax is levied must not be able to affect the size of the tax by changing their behavior. Most tax instruments encountered in practice are not lump sum. Income taxes cannot be lump sum by this definition because a consumer can work more or less hard and vary income in response to the tax. Similarly commodity taxes cannot be lump sum because consumption patterns can be changed. Estate duties are lump sum at the point at which they are levied (since, by definition, the person on which they are levied is dead and unable to choose any other action) but can be affected by changes in behavior prior to death (e.g., by making gifts earlier in life).

There are some taxes, though, that are close to being lump sum. For example, taxing every consumer some fixed amount imposes a lump-sum tax. Setting aside minor details, this was effectively the case of the UK Poll Tax levied in the late 1980s as a source of finance for local government. This tax was unsuccessful for two reasons. First, taxpayers could avoid paying the tax by ensuring that their names did not appear on any official registers. Usually this was achieved by moving house and not making any official declaration of the new address. It appears large numbers of taxpayers did this (unofficial figures put the number as high as 1 million). This “disappearance” is a change in behavior that reduces the tax burden. Second, the theoretical efficiency of lump-sum taxes rests partly on the fact that their imposition is costless, though this was far from the case with the Poll Tax. As it turned out, the difficulty of actually collecting and maintaining information on the residential addresses of all households made the imposition of a uniform lump-sum tax prohibitively expensive. The mobility of taxpayers proved to be much greater than had been expected. Therefore, although the structure of lump-sum taxes makes them appear deceptively simple to collect, this may not be the case in practice, since the tax base (people) is highly mobile and keen to evade. Consequently, in practice, even a uniform lump-sum tax has proved difficult and costly to administer.

However, the costs of collection are only part of the issue. The primary policy concern is the possibility of employing optimal lump-sum taxes. Optimal here means a tax that is chosen, via application of the Second Theorem, to achieve the income distribution necessary to decentralize the chosen allocation of the social planner. The optimal lump-sum tax system is unlikely to be a uniform tax on each consumer. This is because the role of the lump-sum taxes is fundamentally redistributive, so the taxes will be highly differentiated across consumers. Since even uniform lump-sum taxes are implemented with difficulty, the use of differentiated taxes presents even greater problems.

The extent of these problems can be seen by considering the information needed to calculate the taxes. First, the social planner must be able to construct the contract curve...
Chapter 13: Optimality and Comparability

of Pareto-efficient allocations so that the social optimum can be selected. Second, the planner needs to predict the equilibrium that will emerge for all possible income levels so that the incomes needed to decentralize the chosen allocation can be determined. Both of these steps require knowledge of the consumers’ preferences. Finally the social planner must also know the value of each consumer’s endowment in order to calculate their incomes before taxes and hence the lump-sum taxes that must be imposed. The fundamental difficulty is that these economic characteristics, preferences and endowments, are private information. As such they are known only to the individual consumers and are not directly observed by the social planner. The characteristics may be partly revealed through market choices, but these choices can be changed if the consumers perceive any link with taxation. The fact that lump-sum taxes are levied on private information is the fundamental difficulty that hinders their use.

Some characteristics of the consumers are public information, or at least can be directly observed. Lump-sum taxes can then be levied on these characteristics. For example, it may be possible to differentiate lump-sum taxes according to characteristics of the consumers such as sex, age, or eye color. However, these characteristics are not those that are directly economically relevant as they convey neither preference information nor relate to the value of the endowment. Although we could differentiate taxes on this basis, there is no reason why we should want to do so.

This returns us to the problem of private information. Since the relevant characteristics such as ability are not observable, the social planner must either rely on consumers honestly reporting their characteristics or infer them from the observed economic choices of consumers. If the planner relies on the observation of choices, there is invariably scope for consumers to change their market behavior, which then implies that the taxes cannot be lump sum. When reports are the sole source of information, unobserved characteristics cannot form a basis for taxation unless the tax scheme is such that individuals are faced with incentives to report truthfully.

As an example of the interaction between taxes and reporting, consider the following. Let the quality of a consumer’s endowment of labor be determined by their IQ level. Given a competitive market for labor, the value of the endowment is then related to IQ. Assume that there are no economically relevant variables other than IQ, so that any set of optimal lump-sum taxes must be levied on IQ. If the level of lump-sum tax was inversely related to IQ and if all households had to complete IQ tests, then the tax system would not be cheated because the incentive would always be to maximize the score on the test. In this case the lump-sum taxes are said to be incentive compatible, meaning that they give incentives to behave honestly. In contrast, if the taxes were positively related to IQ, a testing procedure could easily be manipulated by the high-IQ
consumers who would intentionally choose to perform poorly. If such a system were put into place, the mean level of tested IQ would be expected to fall considerably. This indicates the potential for misrevelation of characteristics, and the system would not be incentive compatible. Clearly, if a high-IQ results in higher earnings and, ultimately, greater utility, a redistributive policy would require the use of lump-sum taxes that increased with IQ. The tax policy would not be incentive compatible. As the next section shows, such problems will always be present in any attempt to base lump-sum taxes on unobservable characteristics.

13.4 Impossibility of Lump-Sum Taxes

Imagine that each individual in a society can be described by a list of personal attributes upon which the society wishes to condition taxes and transfers (e.g. tastes, needs, talents, and endowments). Individuals are also identified by their names and possibly other publicly observable attributes (e.g., eye color), which are not judged to be relevant attributes for taxation. The list of personal attributes associated to every agent is not publicly known but is the private information of each individual. This implies that the lump-sum taxes the government would like to implement must rely on information about personal attributes which individuals must either report or reveal indirectly through their actions.

Lump-sum taxes are incentive incompatible when at least one individual, who understands how the information that is reported will be used, chooses to report falsely. We have already argued that there can be incentive problems in implementing optimal lump-sum taxes. What we now wish to demonstrate is that these problems are fundamental ones and will always afflict any attempt to implement optimal lump-sum taxes. In brief, the argument will show that optimal lump-sum taxes cannot be incentive compatible. This does not mean that lump-sum taxes cannot be used—for instance, all individuals could be taxed the same amount—but only that the existence of private information places limits on the extent to which taxes can be differentiated before incentives for the false revelation of information come into play. These issues are first illustrated for a particular example and then a general result is provided.

Before describing the general result a good illustration of the failure of incentive compatibility is provided in the following example due to Mirrlees. Assume that individuals can have one of two levels of ability: either low or high. The low ability level is denoted by \( s_l \) and the high ability level by \( s_h \) with \( s_l < s_h \). For simplicity, suppose the number with high ability is equal to the number with low. The two types have
the same preferences over consumption, $x$, and labor, $\ell$ as represented by the utility function $U(x, \ell) = u(x) - v(\ell)$. It is assumed that the marginal utility of consumption is decreasing in $x$ and the marginal disutility of labor is increasing in $\ell$.

To determine the optimal lump-sum taxes, suppose that the government can observe the ability of each individual and impose taxes that are conditioned upon ability. Let the tax on an individual of ability level $i$ be $T_i > 0$ (or a subsidy if $T_i < 0$). The budget constraint of a type $i$ is

$$x_i = s_i \ell_i - T_i,$$

where earnings are $s_i \ell_i$. Given the lump-sum taxes, each type chooses labor supply to maximize utility subject to this budget constraint. The choice of labor supply equates the marginal utility of additional consumption to the disutility of labor

$$\frac{s_i}{\partial x_i} u - \frac{\partial v}{\partial \ell_i} = 0.$$  \hspace{1cm} (13.2)

This provides a labor-supply function $\ell_i = \ell_i(T_i)$.

Now suppose that the government is utilitarian and chooses the lump-sum taxes to maximize the sum of utilities. Then the optimal lump-sum taxes solve

$$\max \left\{ T_l, T_h \right\} \sum_{l,h} u(s_i \ell_i(T_i) - T_i) - v(\ell_i(T_i)),$$  \hspace{1cm} (13.3)

subject to government budget balance, which requires

$$T_h + T_l = 0,$$  \hspace{1cm} (13.4)

since there are an equal number of the two types. This budget constraint can be used to substitute for $T_l$ in (13.3). Differentiating the resulting expression with respect to the tax $T_h$ and using the first-order condition (13.2) for the choice of labor supply, the optimal lump-sum taxes are characterized by the condition

$$\frac{\partial u}{\partial x_h} = \frac{\partial u}{\partial x_l}.$$  \hspace{1cm} (13.5)

Since the marginal utility of consumption is decreasing in $x_i$, the optimality condition (13.5) implies that there is equality of consumption for the two types, $x_h = x_l$. When this conclusion is combined with (13.2) and the fact that $s_l < s_h$, it follows that

$$\frac{\partial v}{\partial \ell_l} = s_l \frac{\partial u}{\partial x_l} < s_h \frac{\partial u}{\partial x_h} = \frac{\partial v}{\partial \ell_h}.$$  \hspace{1cm} (13.6)
Under the assumption of an increasing marginal disutility of labor, this inequality shows that the optimal lump-sum taxes should induce the outcome $\ell_h > \ell_l$, so the more able work harder than the less able. The motivation for this outcome is that working the high-ability type harder is the most efficient way to raise the level of total income for the society which can then be redistributed using the lump-sum taxes. Thus the high-ability type works harder than the low-ability type but only gets to consume the same. Therefore, the high-ability type is left with a lower utility level than the low-ability type after redistribution.

Now suppose that the government can observe incomes but cannot observe the ability of each individual. Assume that it still attempts to implement the optimal lump-sum taxes. The taxes are obviously not incentive compatible because, if the high-ability type understand the outcome, they can always choose to earn as little as the low-ability type. Doing so then qualifies the high-ability type for the redistribution aimed at the low-ability type. This will provide them with a higher utility level than if they did not act strategically. The optimal lump-sum taxes cannot then be implemented with private information.

Who would work hard if the government stood ready to tax away the resulting income? Optimal (utilitarian) lump-sum redistribution makes the more able individuals worse off because it requires them to work harder but does not reward them with additional consumption. In this context it is profitable for the more able individuals to make themselves seem incapable. Many people believe there is something unfair about inequality that arises from the fact that some people are born with superior innate ability or similar advantage over others. But many people also think it morally right that one should be able to keep some of the fruits of one’s own effort. This example may have been simple but its message is far-reaching. The Soviet Union and other communist economies have shown us that it is impossible to generate wealth without simultaneously offering adequate material incentives. Incentive constraints inevitably limit the scope for redistribution.

This example is now shown to reflect a general principle concerning the incentive compatibility of optimal lump-sum taxes. We state the formal version of this result for a “large economy,” which is an economy where the actions of an individual are insignificant relative to the economy as a whole. In other words, there is a continuum of different agents, which is the mathematical form of the idealized competitive economy with a very large number of small agents with no market power. The theorem shows that optimal lump-sum taxation is never incentive compatible.
Theorem 13.1 (Hammond) In a large economy, redistribution through optimal lump-sum taxes is always incentive incompatible.

The logic behind this theorem is surprisingly simple. A system of optimal lump-sum taxes is used to engineer a distribution of endowments that will decentralize the first-best allocation. The endowments after redistribution must be based on the agents’ characteristics (recall that in the analysis of the Second Theorem the taxes were based on knowledge of endowments and preferences), so assume the endowment of an agent with characteristics $\theta_i$ is given by $e_i = e(\theta_i)$. For those characteristics that are not publicly observable, the government must rely on an announcement of the values by the agents. Assume, for simplicity, that none of the characteristics can be observed. Then the incentive exists for each agent to announce the set of characteristics that maximize the value of the endowment at the equilibrium prices $p$. This is illustrated in figure 13.4 where $\theta_1$ and $\theta_2$ are two potential announcements, with related endowments $e(\theta_1)$ and $e(\theta_2)$, and $\theta^*$ is the announcement that maximizes $pe(\theta)$. The announcement of $\theta^*$ leads to the highest budget constraint from among the set of possible announcements and, by giving the agent maximum choice, allows the highest level of utility to be attained. Consequently all agents will announce $\theta^*$ and the optimal lump-sum taxes are not incentive compatible.

The main points of the argument can now be summarized. To implement the Second Theorem as a practical policy tool, it is necessary to employ optimal lump-sum taxes. Such taxes are unlikely to be available in practice or to satisfy all the criteria required

![Figure 13.4](image-url)

Optimal lump-sum taxes and incentive compatibility
of them. The taxes may be costly to collect and the characteristics on which they need to be based may not be observable. When characteristics are not observable, the relationship between taxes and characteristics can give consumers the incentive to make false revelations. It is therefore best to treat the Second Theorem as being of considerable theoretical interest but of very limited practical relevance. The theorem shows us what could be possible, not what is possible.

Lump-sum taxes can achieve the optimal allocation of resources provided all information is public. If some of the characteristics that are relevant for taxation are private information, then the optimal lump-sum taxes are not incentive compatible. Information limitations therefore place a limit on the extent to which redistribution can be undertaken using lump-sum taxation. It is the impracticality of lump-sum taxation that provides the motive for studying the properties of other tax instruments. The income taxes and commodity taxes that are analyzed in chapters 16 and 15 are second-best solutions and are used because the first-best solution, lump-sum taxation, is not available. Lump-sum taxes are used as a benchmark from which to judge the relative success of these alternative instruments. Lump-sum taxes also help clarify what it is that we are really trying to tax.

13.5 Redistribution In-Kind

The lump-sum taxes we have been discussing are a very immediate form of redistribution. In practice, there are numerous widely used methods of redistribution that do not directly involve taxation. Governments frequently provide goods such as education or health services at less than their cost, which may be viewed as a redistributitional policy. One may expect that a cash transfer of the same value would have more redistributitional power than such in-kind transfer programs. This is mistaken. There are three reasons why transfers in-kind may be superior to the cash transfers achieved through standard tax-transfer programs.

One reason is political. Political considerations dictate that many governments ensure that the provision of programs like education, pension, and basic health insurance is universal. Without this feature the programs would not have the political support required to be adopted or continued. For instance, public pensions and health care would be far more vulnerable politically if they were targeted to the poor and not available to others. Redistribution through cash would be even more vulnerable. It should be noted that because a government program is universal, it does not follow that there is no redistribution. First, if the program is financed by proportional income
taxation, the rich will contribute more to its finance than the poor. Second, even if everyone contributes the same to the program, it is possible that the rich will not use the publicly provided good to the same extent as the poor. Consider, for example, a program of public provision of basic health care that is available to everyone for free and financed by a uniform tax on all households. Assume that there exists a private health care alternative with higher quality than the public system but only available at a cost. Since the rich can afford the higher quality, they will use the private health care, even though free public health care is available. These rich households still pay their contribution to the public program, and thus the poor households derive a net benefit from this cross-subsidization.

Another reason for preferring in-kind redistribution is *self-selection*. What ultimately limits redistribution is that it will eventually become advantageous for higher ability people to earn lower incomes by expending less effort and thereby paying the level of taxes (or receiving the transfers) intended for the lower ability groups. The self-selection argument is that anything that makes it less attractive for people to mimic those with lesser ability will extend the limit to redistribution. The use of in-kind transfers can obtain a given degree of redistribution more efficiently because of differences in preferences among different income groups. Consider two individuals who differ not only in their ability but also in their health status. Suppose that lesser ability means also poorer health, so the less able spend relatively more on health. Then both income and health expenditures act as a signal of ability. It follows that the limits to redistribution can be relaxed if transfers are made partly in the form of provision of health care (or equivalently with full subsidization of health expenditures). The reason is simply that the more able individual (with less tendency to become ill) is less likely to claim in-kind benefits in the form of health care provision than he would be to claim cash benefits. To take another example, suppose that the government is considering redistribution either in cash or in the form of low-quality housing. All households, needy or not, would like the cash transfer. However, few non-needy households would want to live in low-quality housing as they can afford better housing. Thus self-selection occurs, and the non-needy drop out of the housing program, which is taken up only by the needy. In short, transfers in-kind invite people to self-select in a way that reveals their neediness. When need is correlated with income-earning ability, then in-kind transfers can relax incentive and selection constraints, thereby improving the government’s ability to redistribute income.

A third reason is the idea of *time consistency* that we introduced in chapter 3. Here the argument for in-kind transfers relies on the inability of government to commit to its future actions. Unlike the argument of Strotz (1956) on government time inconsistency,
this does not arise from a change in government objective over time (e.g., because of elections) nor from the fact that the government is not welfaristic or rational. The time-consistency problem arises from a perfectly rational government that fully respects individual preferences but that does not have the power to commit to its policy in the long run. The time-consistency problem is obvious with regard to pensions. To the extent that households expect governments to provide some basic pension to those with too little savings, their incentive to save for retirement consumption and provide for themselves is reduced. Anticipating this, the government may prefer to provide public pensions. A related time-consistency problem can explain why transfer programs, such as social security, education, and job training are in-kind. If a welfaristic government cannot commit not to come to the rescue of those in need in the future, potential recipients will have little reason to invest in their education or to undertake job training because the government will help them out anyway. Again, the government can improve both economic efficiency and redistribution by making education and job training available at less than their cost, rather than making cash transfers of equivalent value.

13.6 Aspects of Pareto-Efficiency

The analysis of lump-sum taxation has raised questions about the practical value of the Second Theorem of Welfare Economics. Although the theorem shows how an optimal allocation can be decentralized, the means to achieve the decentralization may be absent. If the use of lump-sum taxes is restricted, the government must resort to alternative policy instruments. All alternative instruments will be distortionary and will not achieve the first-best.

These criticisms do not extend to the First Theorem of Welfare Economics, which states only that a competitive equilibrium is Pareto-efficient. Consequently the First Theorem implies no policy intervention, so it is safe from the restrictions on lump-sum taxes. However, at the heart of the First Theorem is the use of Pareto-efficiency as a method for judging the success of an economic allocation. The value of the First Theorem can only be judged once a deeper understanding of Pareto-efficiency has been developed.

The Pareto criterion was introduced into economics by the Italian economist Vilfredo Pareto at the beginning of the twentieth century. This was a period of reassessment in economics during which the concept of utility as a measurable entity was rejected. Alongside this rejection of measurability, the ability to compare utility levels between consumers also had to be rejected. Pareto-efficiency was therefore constructed
explicitly to allow comparisons of allocations without the need to make any interpersonal comparisons of utility. As will be seen, this avoidance of interpersonal comparisons is both its strength and its main weakness.

To assess Pareto-efficiency, it is helpful to develop the concept in three stages. The first stage defines the idea of making a Pareto improvement when moving from one allocation to another. From this can be constructed the Pareto preference order that judges whether one allocation is preferred to another. The final stage is to use Pareto preference to find the most preferred states, which are then defined as Pareto-efficient. Reviewing each of these steps allows us to assess the meaning and value of the concept.

Consider a move from economic state $s_1$ to state $s_2$. This is defined as a Pareto improvement if it makes some consumers strictly better off and none worse off. If there are $H$ consumers, this definition can be stated formally by saying a Pareto improvement is made in going from $s_1$ to $s_2$ if

$$U^h(s_2) > U^h(s_1) \quad \text{for at least one consumer, } h, \quad (13.7)$$

and

$$U^h(s_2) \geq U^h(s_1) \quad \text{for all consumers } h = 1, \ldots, H. \quad (13.8)$$

The idea of a Pareto improvement can be used to construct a preference order over economic states. If a Pareto improvement is made in moving from $s_1$ to $s_2$, then state $s_2$ is defined as being Pareto-preferred to state $s_1$. This concept of Pareto preference defines one state as preferred to another if all consumers are at least as well off in that state and some are strictly better off. It is important to note that this stage of the construction has converted the set of individual preferences of the consumers into social preferences over the states.

The final stage is to define Pareto-efficiency. The earlier definition can be re-phrased as saying that an economic state is Pareto-efficient if there is no state that is Pareto-preferred to it. That is, no move can be made from that state to another that achieves a Pareto improvement. From this perspective, we can view Pareto-efficient states as being the “best” relative to the Pareto preference order. The discussion now turns to assessing the usefulness of Pareto preference in selecting an optimal state from a set of alternatives. By analyzing a number of examples, several deficiencies of the concepts will become apparent.

The simplest allocation problem is to divide a fixed quantity of a single commodity between two consumers. Let the commodity be a cake, and assume that both consumers prefer more cake to less. The first observation is that no cake should be wasted—it is
always a Pareto improvement to move from a state where some is wasted to one with the wasted cake given to one, or both, of the consumers. The second observation is that any allocation in which no cake is wasted is Pareto-efficient. To see this, start with any division of the cake between the two consumers. Any alternative allocation must give more to one consumer and less to the other; therefore, since one must lose some cake, no change can be a Pareto improvement.

From this simple example two deficiencies of Pareto-efficiency can be inferred. First, since no improvement can be made on an allocation where none is wasted, extreme allocations such as giving all of the cake to one consumer are Pareto-efficient. This shows that even though an allocation is Pareto-efficient, there is no implication that it need be good in terms of equity. This illustrates quite clearly that Pareto-efficiency is not concerned with equity. The cake example also illustrates a second point: there can be a multiplicity of Pareto-efficient allocations. This was shown in the cake example by the fact that every nonwasteful allocation is Pareto-efficient. This multiplicity of efficient allocations limits the value of Pareto-efficiency as a tool for making allocative decisions. For the cake example, Pareto-efficiency gives no guidance whatsoever in deciding how the cake should be shared, other than showing that none should be thrown away. In brief, Pareto-efficiency fails to solve even this simplest of allocation problems.

The points made in the cake division example are also relevant to allocations within a two-consumer exchange economy. The contract curve in figure 13.5 shows the set of Pareto-efficient allocations, and there is generally an infinite number of these. Once again the Pareto preference ordering does not select a unique optimal outcome. In addition the competitive equilibrium may be as the one illustrated in the bottom left corner of the box. This has the property of being Pareto-efficient, but it is highly inequitable and may not find much favor using other criteria for judging optimality.
Another failing of the Pareto preference ordering is that it is not always able to compare alternative states. In formal terms, it does not provide a complete ordering of states. This is illustrated in figure 13.6 where the allocations $s_1$ and $s_2$ cannot be compared, although both can be compared to $s_3$ ($s_3$ is Pareto-preferred to both $s_1$ and $s_2$). When faced with a choice between $s_1$ and $s_2$, the Pareto preference order is silent about which should be chosen. It should be noted that this incomparability is not the same as indifference. If the preference order were indifferent between two states, then they are judged as equally good. Incomparability means the pair of states simply cannot be ranked.

The basic mechanism at work behind this example is that the Pareto preference order can only rank alternative states if there are only gainers or only losers as the move is made between the states. If some gain and some lose, as in the choice between $s_1$ and $s_2$ in figure 13.6, then the preference order is of no value. Such gains and losses are invariably a feature of policy choices and much of policy analysis consists of weighing up the gains and losses. In this respect Pareto-efficiency is insufficient as a basis for policy choice.

To summarize these arguments, Pareto-efficiency does not embody any concept of justice, and highly inequitable allocations can be efficient under the criterion. In many situations there are very many Pareto-efficient allocations, in which case the criterion provides little guidance for policy choice. Finally Pareto-efficiency may not provide a complete ordering of states, so some states will be incomparable under the criterion. The source of all these failing is that the Pareto criterion avoids weighing gains against losses, but it is just such judgments that have to be made in most allocation decisions. To make a choice of allocation, the evaluation of the gains and losses has to be faced directly.
13.7 Social Welfare Functions

The social welfare function was employed in section 13.2 to introduce the concept of a socially optimal allocation. At that point it was simply described as a means by which different allocations of utility between consumers could be socially ranked. What was not done was to provide a convincing description of where such a ranking could come from or of how it could be constructed. Three alternative interpretations will now be given, each of which provides a different perspective on the social welfare function.

The first possibility is that the social welfare function captures the distributive preferences of a central planner or dictator. Under this interpretation there can be two meanings of the individual utilities that enter the function. One is that they are the planner’s perception of the utility achieved by each consumer at their level of consumption. This provides a consistent interpretation of the social welfare function, but problems arise in its relation to the underlying model. To see why this is so, recall that the Edgeworth box and the contract curve within it were based on the actual preferences of the consumers. There is then a potential inconsistency between this construction and the evaluation using the planner’s preferences. For example, what is Pareto-efficient under the true preferences may not be one under the planner’s (it need not even be an equilibrium).

The alternative meaning of the utilities is that they are the actual utilities of the consumers. This leads directly into the central difficulty faced in the concept of social welfare. In order to evaluate all allocations of utility it must be possible to determine the social value of an increase in one consumer’s utility against the loss in another’s. This is only possible if the utilities are comparable across the consumers. More will be said about this below.

The second interpretation of the social welfare function is that it captures some ethical objective that society should be pursuing. Here the social welfare function is determined by what is viewed as the just objective of society. There are two major examples of this. The utilitarian philosophy of aiming to achieve the greatest good for society as a whole translates into a social welfare function that is the sum of individual utilities. In this formulation only the total sum of utilities counts, so it does not matter how utility is distributed among consumers in the society. Alternatively, the Rawlsian philosophy of caring only for the worst-off member of society leads to a level of social welfare determined entirely by the minimum level of utility in that society. With this objective the distribution of utility is of paramount importance. Gains in utility achieved by anyone other than the worst-off consumer do not improve social welfare.
Although this approach to the social welfare function is internally consistent, it is still not entirely satisfactory. The utilitarian approach requires that the utilities of the consumers be added in order to arrive at the total sum of social welfare. The Rawlsian approach necessitates the utility levels being compared in order to find the lowest. The nature of the utility comparability is different for the two approaches (being able to add utilities is different to being able to compare), but both rely on some form of comparability. This again leads directly into the issue of utility comparisons.

The final view that can be taken of the social welfare function is that it takes the preferences of the individual consumers (represented by their utilities) and aggregates these into a social preference. This aggregation process would be expected to obey certain rules; for instance, if all consumers prefer one state to another, it should be the case that the social preference also prefers the same state. The structure of the social welfare function then emerges as a consequence of the rules the aggregation must obey.

Although this arrives at the same outcome as the other two interpretations, it does so by a distinctly different process. In this case it is the set of rules for aggregation that are foremost rather than the form of social welfare. That is, the philosophy here would be that if the aggregation rules are judged as satisfactory, then society should accept the social welfare function that emerges from their application, whatever its form. An example of aggregating preferences is the rule of majority voting (despite the failings already identified in chapter 11), since the minority must accept what the majority chooses.

The consequences of constructing a social welfare function by following this argument are of fundamental importance in the theory of welfare economics. In fact doing so leads straight back into Arrow’s Impossibility Theorem, which was described in chapter 11. The next section is dedicated to interpreting the theorem and its implications in this new setting.

### 13.8 Arrow’s Theorem

Although they appear very distinct in nature, both majority voting and the Pareto criterion are examples of procedures for aggregating individual preferences into a social preference. It has been shown that neither is perfect. The Pareto preference order can be incomplete and unable to rank some of the alternatives. Majority voting always leads to a complete social preference order, but this may not be transitive. What Arrow’s Impossibility Theorem has shown is that such failings are not specific to these
aggregation procedures. All methods of aggregation will fail to meet one or more of its conditions, so the Impossibility Theorem identifies a fundamental problem at the heart of generating social preferences from individual preferences.

The conditions of Arrow’s theorem were stated in terms of the rankings induced by individual preferences. However, since individual preferences can usually be represented by a utility function, the theorem also applies to the aggregation of individual utility functions into a social welfare function. The implication behind applying the theorem is that a social welfare function does not exist that can aggregate individual utilities without conflicting with one, or more, of the conditions I.N.P.U.T. This means that whatever social welfare function is proposed, there will be some set of utility functions for which it conflicts with at least one of the conditions. In other words, no ideal social welfare function can be found. No matter how sophisticated the aggregation mechanism is, it cannot overcome this theorem.

Since the publication of Arrow’s theorem there has been a great deal of research attempting to find a way out of the dead end into which it leads. One approach that has been tried is to consider alternative sets of aggregation rules. For instance, transitivity of the social preference ordering can be relaxed to quasi-transitivity (only strict preference is transitive) or weaker versions of condition I and condition P can be used. Most such changes just lead to further impossibility theorems for these different sets of rules. Modifying the rules does not therefore really seem to be the way forward out of the impossibility.

What is at the heart of the impossibility is the limited information contained in individual utility functions. Effectively all that is known is the individuals’ rankings of the alternatives—which is best, which is worst, and how they line up in between. What the rankings do not give is any strength of feeling either between alternatives for a given individual or across individuals for a given option. Such strength of feeling is an essential art of any attempt to make social decisions. Consider, for instance, a group of people choosing where to dine. In this situation a strong preference in one direction (“I really don’t want to eat fish”) usually counts for more than a mild preference (“I don’t really mind, but I would prefer fish”). Arrow’s theorem rules out any information of this kind.

Using information on how strongly individuals feel about the alternatives can be successful in choosing where to dine. It is interesting that the strength of preference comparisons can be used in informal situations, but this does not demonstrate that it can be incorporated within a scientific theory of social preferences. This issue is now addressed in detail.
13.9 Interpersonal Comparability

Earlier in this chapter it was noted that Pareto-efficiency was originally proposed because it provided a means by which it was possible to compare alternative allocations without requiring interpersonal comparisons of welfare. It is also from this avoidance of comparability that the failures of Pareto-efficiency emerge. This point is at the core of the Impossibility Theorem. To proceed further, this section first reviews the development of utility theory in order to provide a context and then describes alternative degrees of utility comparability.

Nineteenth-century economists viewed utility, the level of happiness of an individual, as something that was potentially measurable. Advances in psychology were expected to deliver the machinery for conducting the actual measurement. If utility were measurable, it follows naturally that it would be comparable among individuals. This ability to measure utility, combined with the philosophy that society should aim for the greatest good, came to provide the underpinnings of utilitarianism. The measurability of utility permitted social welfare to be expressed by the sum of individual utilities. Ranking states by the value of this sum then gave a means of aggregating individual preferences that satisfied all of the conditions of the impossibility theorem except for the information content. If the envisaged degree of measurability could be achieved, then the restrictions of the impossibility theorem are overcome.

This concept of measurable and comparable utility began to be dispelled in the early twentieth century. There were two grounds for this rejection. First, no means of measuring utility had been discovered, and it was becoming clear that the earlier hopes would not be realized. Second, advances in economic theory showed that there was no need to have measurable utility in order to construct a coherent theory of consumer choice. In fact the entire theory of the consumer could be derived by specifying only the consumer’s preference ordering. The role of utility then became strictly secondary—it could be invoked to give a convenient function to represent preferences if necessary but was otherwise redundant. Since utility had no deeper meaning attached to it, any increasing monotonic transformation of a utility function representing a set of preferences would also be an equally valid utility function. Utility was simply an ordinal concept, with no natural zero or units of measurement. By the very construction of utility, comparability between different consumers’ utilities was a meaningless concept. This situation therefore left no scientific basis on which to justify the comparability of different consumer’s utility levels.
This perspective on utility, and the consequent elimination of utility comparisons among consumers, created the need to develop concepts for social comparisons, such as Pareto-efficiency, that were free of interpersonal comparisons. However, the weaknesses of these criteria soon became obvious. The analytical trend since the 1960s has been to explore the consequences of re-admitting interpersonal comparability into the analysis. The procedure adopted is basically to assume that comparisons are possible. This permits the derivation of results from which interpretations can be obtained. These are hoped to provide some general insights into policy that can be applied, even though utility is not actually comparable in the way assumed.

There are even some economists who would argue that comparisons are possible. One basis for this is the claim that all consumers have very similar underlying preference orderings. All prefer to have more income to less, and consumers with equal incomes make very similar divisions of expenditures between alternative groups of commodities. For example, expenditure on food is similar, even though the actual foodstuffs purchased may be very different. In modeling such consumers, it is possible to assert that they all have the same utility function guiding their choices. This makes their utilities directly comparable.

So far comparability has been used as a catch-all phrase for being able to draw some contrast between the utility levels of consumers. In fact many different degrees of comparability can be envisaged. For instance, the claim that one household has a higher level of utility than another requires rather less comparability than claiming it has 15 percent more utility. Different degrees of comparability have implications for the way in which individual utilities can be aggregated into a social preference ordering.

The starting point for discussing comparability is to define the two major forms of utility. The first is *ordinal utility*, which is the familiar concept from consumer theory. Essentially an ordinal utility function is no more than just a numbering of a consumer’s indifference curves, with the numbering chosen so that higher indifference curves have higher utility numbers. These numbers can be subjected to any form of transformation without altering their meaning, provided that the transformation leaves the ranking of the numbers unchanged—higher indifference curves must still have larger utility numbers attached. Because they can be so freely transformed, there is no meaning to differences in utility levels between two situations for a single consumer except which of the two provides the higher utility.

The second form of utility is *cardinal utility*. Cardinal utility imposes restrictions beyond those of ordinal utility. With cardinal utility one can only transform utility numbers by multiplying by a constant and then adding a constant, so an initial utility function \( U \) becomes the transformed utility \( \hat{U} = a + bU \), where \( a \) and \( b \) are the constants. Any
other form of transformation will affect the meaning of a cardinal utility function. The
typical place where cardinal utility is found is in the economics of uncertainty, since
an expected utility function is cardinal. This cardinality is a consequence of the fact
that an expected utility function must provide a consistent ranking for different prob-
ability distributions of the outcomes. (A noneconomic example of a cardinal scale is
temperature. It is possible to convert Celsius to Fahrenheit by multiplying by \( \frac{9}{5} \) and
adding 32. The converse transformation from Fahrenheit to Celsius is to multiply by \( \frac{5}{9} \)
and subtract 32.) With these definitions it now becomes possible to talk in detail about
comparability and noncomparability.

Noncomparability can arise with both ordinal and cardinal utility. What noncompar-
rability means is that we can apply different transformations to different consumers’
utilities. To express this in formal terms, let \( U_1 \) be the utility function of consumer 1
and \( U_2 \) the utility function of consumer 2. Then noncomparability arises if the trans-
formation \( f_1 \) can be applied to \( U_1 \) and a different transformation \( f_2 \) to \( U_2 \), with no
relationship between \( f_1 \) and \( f_2 \). Why is this noncomparable? The reasoning is that
by suitably choosing \( f_1 \) and \( f_2 \), it is always possible to start with one ranking of the
initial utilities and to arrive at a different ranking of the transformed utilities. The utility
information therefore does not provide sufficient information to make a comparison of
the two utility levels.

Comparability exists when the transformations that can be applied to the utility
functions are restricted. With ordinal utility there is only one possible degree of compa-
rability. This occurs when the ordinal utilities for different consumers can be subjected
only to the same transformation. The implication of this is that the transformation pre-
serves the ranking of utilities among different consumers. So, if one consumer has a
higher utility than another before the transformation, the same consumer will have a
higher utility after the transformation. Letting this transformation be denoted by \( f \),
then if \( U_1 \geq U_2 \), it must be the case that \( f(U_1) \geq f(U_2) \). This form of comparability
is called ordinal level comparability.

If the underlying utility functions are cardinal, there are two forms of comparability
that are worth discussing. The first form of comparability is to assume that the constant
multiplying of utility in the transformation must be the same for all consumers, but the
constant that is added can differ. Hence for two consumers the transformed utilities are
\( \tilde{U}_1 = a^1 + bU^1 \) and \( \tilde{U}_2 = a^2 + bU^2 \), so the constant \( b \) is the same for both. This is
called cardinal unit comparability. The implication of this transformation is that it now
becomes meaningful to talk about the effect of changes in utility, meaning that gains to
one consumer can be measured against losses to another—and whether the gain exceeds
the loss is not affected by the transformation. The second degree of comparability for
cardinal utility is to further restrict the constant $a$ in the transformation to be the same for both consumers. For all consumers the transformed utility becomes $	ilde{U}^h = a + bU^h$. It is now possible for both changes in utility and levels of utility to be compared. This form of comparability is called \textit{cardinal full comparability}.

The next step is to explore the implications of these comparabilities for the construction of social welfare functions. It will be shown that each form of comparability implies different permissible social welfare functions.

13.10 Comparability and Social Welfare

The discussion of Arrow’s Impossibility Theorem showed that the failure to successfully generate a social preference ordering from a set of individual preference orderings was the result of limited information. The information content of an individual’s preference order involves nothing more than knowing how they rank the alternatives. A preference order does not convey any information on the strength of preferences or allow comparison of utility levels across consumers. When more information is available, it becomes possible to find social preference orderings that satisfy the conditions I, N, P, U, T. Such information can be introduced by building social preferences on individual utility functions that allow for comparability.

What this section shows is that for each form of comparability there is a specification of social welfare function that is consistent with the information content of the comparable utilities. To explain what is meant by consistent, recall that comparability is described by a set of permissible transformations of utility. A social welfare function is \textit{consistent} if it ranks the set of alternative social states in the same way for all permissible transformations of the utility functions. Since increasing the degree of comparability reduces the number of permissible transformations, it has the effect of increasing the set of consistent social welfare functions.

Let the utility obtained by consumer $h$ from allocation $s$ be $U^h(s)$. A transformation of this basic utility function is denoted by $	ilde{U}^h(s) = f^h(U^h(s))$. The value of social welfare at allocation $s$ using the basic utilities is $W(s) = W(U^1(s), \ldots, U^H(s))$, and that from using the transformed utilities is $	ilde{W}(s) = W\left(\tilde{U}^1(s), \ldots, \tilde{U}^H(s)\right)$. Given alternative allocations $A$ and $B$, the social welfare function is consistent with the transformation (and hence the form of comparability) if $W(A) \geq W(B)$ implies $	ilde{W}(A) \geq \tilde{W}(B)$. In words, if $A$ generates higher social welfare than $B$ for the basic utilities, it will also do so for the transformed utilities.
To demonstrate these points, assume there are two consumers with the basic utility functions $U_1 = \frac{x_1}{2} \frac{y_1}{2}$ and $U_2 = x + y$, where $x$ and $y$ are the consumption levels of the two goods. Further assume that there are two allocations $A$ and $B$ with the consumption levels, and the resulting utilities, as shown in table 13.1.

The first point to establish is that it is possible to find a social welfare function that is consistent with ordinal level comparability but none that is consistent with ordinal noncomparability. What level comparability allows is the ranking of consumers by utility level (think of placing the consumers in a line with the lowest utility level first). A position in this line (e.g., the first, or the tenth, or the $n$th) can be chosen, and the level of utility of the consumer in that position used as the measure of social welfare. This process generates a positional social welfare function. The best known example is the Rawlsian social welfare function, $W = \min\{U_h\}$, which judges social welfare by the minimum level of utility in the population. An alternative that shows other positions can be employed (though not one that is often used) is to measure social welfare by the maximum level of utility, $W = \max\{U_h\}$.

That such positional welfare functions are consistent with ordinal level comparability but not with ordinal noncomparability is shown in table 13.2 using the allocations $A$ and $B$ introduced above. For the social welfare function $W = \min\{U^h\}$, the welfare level in allocation $A$ is 5 and that in allocation $B$ is 4. Therefore allocation $A$ is judged superior using the basic utilities. An example of a pair of transformations that satisfy ordinal noncomparability are $\tilde{U}^1 = f^1(U^1) = 3U^1$ and $\tilde{U}^2 = f^2(U^2) = 2U^2$. The levels of utility and resulting social welfare are displayed in the upper part of table 13.2. The table shows that the preferred allocation is now $B$, so the transformation has changed the preferred social outcome. With ordinal level comparability, the transformations $f^1(U^1)$ and $f^2(U^2)$ must be the same. For example, let the transformation be given by $\tilde{U}^h = f(U^h) = (U^h)^2$. The values of the transformed utilities in the lower part of the table confirm that allocation $A$ is preferred—as it was with the basic utilities. The positional social welfare function is therefore consistent with ordinal level comparability.
Although cardinal utility is often viewed as stronger concept than ordinal utility, cardinality alone does not permit the construction of a consistent social welfare function. Recalling that transformations of the form $f^h = a^h + b^h U^h$ can be applied with noncomparability, it can be seen that even positional welfare functions will not be consistent, since $a^h$ can always be chosen to change the social ranking generated by the transformed utilities compared to that generated by the basic utilities. In contrast, if utility satisfies cardinal unit comparability, it is possible to use social welfare functions of the form

$$W = \sum_{h=1}^{H} \alpha^h U^h,$$

(13.9)

where the $\alpha^h$ are constants. To demonstrate this, and to show that social welfare function is not consistent with cardinal noncomparability, assume that $\alpha^1 = 2$ and $\alpha^2 = 1$. Then, under the basic utility functions, the social welfare levels in the two allocations are $W(A) = 2 \times 6 + 5 = 17$ and $W(B) = 2 \times 4 + 7 = 15$, so allocation $A$ is preferred. The upper part of table 13.3 displays two transformations satisfying noncomparability and the implied value of social welfare. This shows that allocation $B$ will be preferred with the transformed utility. Therefore the social welfare function is not consistent with the transformations. With cardinal unit comparability, the transformations are restricted to have a common value for $b^h$, so $\tilde{U}^h = a^h + b U^h$. Two such transformations are selected, and the resulting utility levels are given in the lower part of the table. Calculation of the social welfare shows the preferred allocation to be $A$ as it was with the basic utilities. Therefore with cardinal level comparability, social welfare functions
Table 13.3
Cardinal utility

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noncomparability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \hat{U}^1 = f^1(U^1) = 2 + 2U^1 )</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>( \hat{U}^2 = f^2(U^2) = 5 + 6U^2 )</td>
<td>35</td>
<td>47</td>
</tr>
<tr>
<td>( W = 2\hat{U}^1 + \hat{U}^2 )</td>
<td>63</td>
<td>67</td>
</tr>
<tr>
<td>Level comparability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \hat{U}^1 = f^1(U^1) = 2 + 3U^1 )</td>
<td>20</td>
<td>14</td>
</tr>
<tr>
<td>( \hat{U}^2 = f^2(U^2) = 5 + 3U^2 )</td>
<td>20</td>
<td>26</td>
</tr>
<tr>
<td>( W = 2\hat{U}^1 + \hat{U}^2 )</td>
<td>60</td>
<td>54</td>
</tr>
</tbody>
</table>

of the form (13.9) are consistent and provide a social ranking that is invariant for the permissible transformations.

With cardinal full comparability the transformations must satisfy \( \hat{U}^h = a + bU^h \). One interesting example of the forms of social welfare function that are consistent with such transformations is

\[
W = \bar{U} + \gamma \min\{U^h - \bar{U} \}, \quad \bar{U} = \frac{\sum_{h=1}^H U^h}{H},
\]

(13.10)

where \( \gamma \) is a parameter that can be chosen. This form of social welfare function is especially interesting because it is the utilitarian social welfare function when \( \gamma = 0 \) and Rawlsian when \( \gamma = 1 \). To show that this function is not consistent for cardinal unit comparability, assume \( \gamma = \frac{1}{2} \). For the basic utilities it follows for allocation \( A \) that \( \bar{U} = \frac{6 + 5}{2} = 5.5 \) and for allocation \( B \), \( \bar{U} = \frac{4 + 7}{2} = 5.5 \). The social welfare levels are then \( W = 5.5 + \frac{1}{2} \min\{6 - 5.5, 5 - 5.5\} = 5.25 \) for allocation \( A \) and \( W = 5.5 + \frac{1}{2} \min\{4 - 5.5, 7 - 5.5\} = 4.75 \) for allocation \( B \). The social welfare function would select allocation \( A \). The upper part of table 13.4 displays the welfare levels for two transformations that satisfy cardinal level comparability. With these transformed utilities the welfare function would select allocation \( B \), so the social welfare function is not valid for these transformations. The lower part of the table displays a transformation that satisfies cardinal full comparability. For this transformation the social welfare function selects allocation \( A \) for both the basic and the transformed utilities. This demonstrates the consistency.

These calculations have demonstrated that if we can compare utility levels among consumers, then a consistent social welfare function can be constructed. The resulting
social welfare function must agree with the information content in the utilities, so each form of comparability leads to a different consistent social welfare function. As the information increases, so does the range of consistent social welfare functions. Expressed differently, for each of the cases of comparability the problem of aggregating individual preferences leads to a well-defined form of social welfare function. All these social welfare functions will generate a social preference ordering that completely ranks the alternative states. They are obviously stronger in content than majority voting or Pareto-efficiency. The drawback is that they are reliant on stronger utility information that may simply not exist.

13.11 Conclusions

This chapter has cast a critical eye over the efficiency theorems of chapter 2. Although these theorems are important for providing a basic framework in which to think about policy, they are not an end in their own right. This perspective is based on the limited practical applicability of the lump-sum transfers needed to support the decentralization in the Second Theorem and the weakness of Pareto-efficiency as a method of judging among economic states.

Although at first sight the theorems apparently have very strong policy implications, they become weakened when placed under critical scrutiny. But they are not without value. Much of the subject matter of public economics takes as its starting point the practical shortcomings of these theorems and attempts to find a way forward to something that is applicable. A knowledge of what could be achieved if the
optimal lump-sum transfers were available provides a means of assessing the success of what can be achieved and shows ways in which improvements in policy can be made.

The other aspect involved in the Second Theorem is the selection of the optimal allocation to be decentralized. This choice requires a social welfare function that can be used to judge different allocations of utility among consumers. Such a social welfare function can only be constructed if the consumers’ utilities are comparable. The chapter described several different forms of comparability and of the social welfare functions that are consistent with them.

Further Reading

Arrow’s Impossibility Theorem was first demonstrated in:

The theorem is further elaborated in:

A comprehensive textbook treatment is given by:

The concept of a social welfare function was first introduced by:

An analysis of limitations on the use of lump-sum taxation is contained in:

An economic assessment of the UK poll tax is conducted in:

For a more complete theoretical treatment of the information constraint on redistribution see:


Two excellent reviews of the central issues that arise with redistribution:
Part V: Equity and Distribution


The self-selection argument for in-kind redistribution is in:


Estimates of the incentive effects of welfare programs are in:


The government time-consistency problem is in:


Comparability of utility is discussed in:


A discussion of the relation between social welfare functions and Arrow’s theorem can be found in:


Several of Sen’s papers that discuss these issues are collected in:


The application of Arrow’s theorem to economic allocation problems is discussed in:

Exercises

13.1 Should a social planner be concerned with the distribution of income or the distribution of utility? How does the answer relate to needs and abilities?

13.2 Sketch the indifference curves of the Bergson–Samuelson social welfare function \( W = U^1 + U^2 \). What do these indifference curves imply about the degree of concern for equity of the social planner? Repeat for the welfare function \( W = \min\{U^1, U^2\} \).

13.3 Show that an anonymous social welfare function must have indifference curves that are symmetric about the 45 degree line. Will an optimal allocation with an anonymous social welfare function and a symmetric utility possibility frontier always be equitable?

13.4 Assume that the preferences of the social planner are given by the function \( W = \left[ \frac{U^1}{\varepsilon} \right] + \left[ \frac{U^2}{\varepsilon} \right] \). What effect does an increase in \( \varepsilon \) have on the curvature of a social indifference curve? Use this result to relate the value of \( \varepsilon \) to the planner’s concern for equity.

13.5 There are \( H \) consumers who each have utility function \( U^h = \log(M^h) \). If the social welfare function is given by \( W = \sum U^h \), show that a fixed stock of income will be allocated equitably. Explain why this is so.

13.6 For a social welfare function \( W = W(U^1(M^1), \ldots, U^H(M^H)) \), where \( M^h \) is income, the “social marginal utility of income” is defined by \( \frac{\partial W}{\partial M^h} \). If \( U^h = [M^h]^{1/2} \) for all \( h \), show that the social marginal utility of income is decreasing in \( M^h \) for a utilitarian social welfare function. Use this to argue that a fixed stock of income will be distributed equally. Show that the argument extends to any anonymous and concave social welfare function when all consumers have the same utility function.

13.7 The two consumers that constitute an economy have utility functions \( U^1 = x^1_1 x^2_1 \) and \( U^2 = x^2_1 x^2_2 \).

a. Graph the indifference curves of the consumers, and show that at every Pareto-efficient allocation \( \frac{x^1_1}{x^2_1} = \frac{x^1_2}{x^2_2} \).

b. Employ the feasibility conditions and the result in part a to show that Pareto-efficiency requires \( \frac{x^1_1}{x^2_1} = \frac{\omega_1}{\omega_2} \), where \( \omega_1 \) and \( \omega_2 \) denote the endowments of the two goods.

c. Using the utility function of consumer 2, solve for \( x^2_1 \) and \( x^2_2 \) as functions of \( \omega_1 \), \( \omega_2 \), and \( U^2 \).

d. Using the utility function of consumer 1, express \( U^1 \) as a function of \( \omega_1 \), \( \omega_2 \), and \( U^2 \).

e. Assuming that \( \omega_1 = 1 \) and \( \omega_2 = 1 \), plot the utility possibility frontier.

f. Which allocation maximizes the social welfare function \( W = U^1 + U^2 \)?

13.8 “Government intervention in markets is essential if we wish to achieve a fair allocation of resources.” Is this correct?
13.9 Consider three individuals with utility indicators $U^A = M^A$, $U^B = \nu M^B$ and $U^C = \gamma M^C$.

a. Show that there are values of $\nu$ and $\gamma$ that can generate any social ordering of the income allocations $a = (5, 2, 5)$, $b = (4, 6, 1)$, and $c = (3, 4, 8)$ when evaluated by the social welfare function $W = U^A + U^B + U^C$.

b. Assume instead that $U^A = \nu + \gamma M^A$, $U^B = \nu + \gamma M^B$ and $U^C = \nu + \gamma M^C$. Show that the evaluation via the utilitarian social welfare function is unaffected by the choices of $\nu$ and $\gamma$.

c. Now assume $U^h = [M^h]^{\gamma}$, where $h = A, B, C$. Show that the preferred outcome under the social welfare function $W = \min_h\{U^A, U^B, U^C\}$ is unaffected by choice of $\gamma$ but that for the welfare function $W = U^A + U^B + U^C$ is affected.

d. Explain the answers to parts a through c in terms of the comparability of utility.

13.10 Provide an argument to establish that the optimal allocation must be Pareto-efficient. What assumptions have you placed on the social welfare function?

13.11 Consider an economy with two individuals (1 and 2). $A$, $B$, and $C$ are three points that belong to the utility possibility frontier of the economy. The individual utilities $(U^1, U^2)$ at the three points are as follows:

<table>
<thead>
<tr>
<th>Points</th>
<th>$U^1$</th>
<th>$U^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4</td>
<td>21</td>
</tr>
<tr>
<td>B</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td>C</td>
<td>14</td>
<td>6</td>
</tr>
</tbody>
</table>

Now consider point $D$.

<table>
<thead>
<tr>
<th>Point</th>
<th>$U^1$</th>
<th>$U^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>9</td>
<td>15</td>
</tr>
</tbody>
</table>

a. Does point $D$ lie on the utility possibility frontier? How does the answer change if you know the utility possibility frontier is concave?

b. Are there any points on the utility possibility frontier that are Pareto-preferred to $C$? Justify your answer.

c. Which of the points $A$, $B$, $C$, or $D$ lies on the highest indifference curve of a utilitarian social welfare function?

d. Which point is on the highest indifference curve of a Rawlsian social welfare function? Explain the answer using the solution to part a.

13.12 The most general form of a social welfare function $SWF$ can be written as $W = W(U^1, \ldots, U^H)$.

a. Explain the following properties that a $SWF$ may satisfy: nonpaternalism, Pareto principle, anonymity (the names of the agents do not matter), and concavity (aversion to inequality).

b. Consider two agents $h = 1, 2$ with utilities $U^1$ and $U^2$. Depict the social indifference curve of the utilitarian $SWF$ in $(U^1, U^2)$-space. Which of the properties in part a does it satisfy?
c. Depict the social indifference curves of the maxi-min or Rawlsian SWF. Contrast to the utilitarian SWF with respect to the aversion to inequality. Which properties does the Rawlsian SWF satisfy?

d. The Bernoulli–Nash social welfare function is given by the product of individual utilities. Discuss the distributional properties of the Bernoulli–Nash SWF.

13.13 Consider the SWF of the form \( W = \left( \frac{1}{\eta} \right)^{1/\eta} \) for \(-\infty < \eta \leq 1\). Show that this SWF reduces to the utilitarian SWF when \( \eta = 1 \), to the Bernoulli–Nash SWF when \( \eta = 0 \), and to the maxi-min Rawlsian SWF when \( \eta \to -\infty \).

13.14 Are the following statements true or false? Provide examples to demonstrate your answer.
   a. A Pareto improvement is always obtained when the economy moves from a point inside the utility possibility frontier to a point on the frontier.
   b. A policy intervention will increase social welfare if and only if it is a Pareto improvement.
   c. A policy intervention will increase social welfare for every Bergson–Samuelson social welfare function if and only if it is a Pareto improvement.

13.15 A fixed amount \( x \) of a good has to be allocated between two individuals, \( h = 1, 2 \) with utility functions \( U_h = \alpha_h x_h^\beta \) (with \( \alpha_h > 0 \)), where \( x_h \) is the amount of the good allocated to consumer \( h \).
   a. How should \( x \) be allocated to maximize a utilitarian SWF? Illustrate the answer graphically.
   b. What is the allocation maximizing the Bernoulli–Nash SWF? Illustrate graphically.
   c. What is the allocation maximizing the maxi-min Rawlsian SWF? Illustrate graphically.

13.16 Show how the results of the previous exercise change if we assume a utility function of the form \( U_h = a_h \sqrt{x_h} \).

13.17 The 31 professors in an economics department have to vote on the location of a new coffee machine. The offices of the professors are located along one side of a corridor. Every professor would receive a utility of 20 if the coffee machine was placed outside his office. Utility is reduced by one unit for each office the professor has to pass to reach the machine.
   a. Which location for the coffee machine wins a simple majority vote?
   b. Which location would be chosen by a benevolent Rawlsian planner?
   c. Which location would be chosen by a benevolent utilitarian planner?
   d. How would the answer change if the coffee machine was an irritation, so each professor gained a unit of utility for each office he had to pass to reach the machine?

13.18 Consider a two-good exchange economy with two types of consumers. Type \( A \) have the utility function \( U^A = 2 \log(x^A_1) + \log(x^A_2) \) and an endowment of 3 units of good 1 and \( k \) units of good 2. Type \( B \) have the utility function \( U^B = \log(x^B_1) + 2 \log(x^B_2) \) and an endowment of 6 units of good 1 and \( 21 - k \) units of good 2.
a. Find the competitive equilibrium outcome and show that the equilibrium price \( p^* = \frac{p_1}{p_2} \) of good 1 in terms of good 2 is \( p^* = \frac{21+k}{15} \).

b. Find the income levels \((M^A, M^B)\) of both types in equilibrium as a function of \( k \).

c. Suppose that the government can make a lump-sum transfer of good 2, but it is impossible to transfer good 1. Use your answer to part b to describe the set of income distributions attainable through such transfers. Draw this in a diagram.

d. Suppose that the government can affect the initial distribution of resources by varying \( k \). Find the optimal distribution of income if (i) the SWF is \( W = \log(M^A) + \log(M^B) \) and (ii) \( W = M^A + M^B \).

13.19 Are the following true or false? Explain your answer.

a. Cardinal utilities are always interpersonally comparable.

b. A Rawlsian social welfare function can be consistent with ordinal utility.

c. The optimal allocation with a utilitarian social welfare function is always inequitable.

13.20 The purpose of this exercise is to illustrate the potential conflict between personal liberty and the Pareto principle (first studied by Sen). Assume there is a copy of Lady Chatterley’s Lover available to be read by two persons, A and B. There are three possible options: (a) A reads the book and B does not; (b) B reads the book and A does not; (c) neither reads the book. The preference ordering of A (the prude) is \( c \succ_A a \succ_A b \) and the preference ordering of B (the lascivious) is \( a \succ_B b \succ_B c \). Hence c is the worst option for one and the best option for the other; while both prefer a to b. Define the personal liberty rule as allowing everyone to choose freely on personal matters (like the color of one’s own hair) with society as a whole accepting the choice, no matter what others think.

a. Apply the personal liberty rule to the example to derive social preferences \( b \succ c \) and \( c \succ a \).

b. Show that by the Pareto principle we must have a social preference cycle \( a \succ b \succ c \succ a \).

c. Suppose that liberalism is constrained by the requirement that the prude A decides to respect B’s preferences such that A’s preference for c over b is ignored. Similarly for B, only his preference for b over c is relevant but not his preference for a over c. What are the modified preference orderings of each person? Show that it leads to acyclic (transitive) social preference.

d. The second possibility to solve the paradox is to suppose that each is willing to respect the other’s choice. Thus A respects B’s preference for b over c and B respects A’s preference for c over a. What are the modified preference orderings of each person? Show that it leads to acyclic social preference. What is then the best social outcome?
A social welfare function permits the evaluation of economic policies that cause re-
distribution among consumers—a task that Pareto-efficiency can never accomplish.
Although the concept of a social welfare function is a simple one, previous chapters
have identified numerous difficulties on the path between individual utility and ag-
gregate social welfare. The essence of these difficulties is that if the individual utility
function corresponds with what is theoretically acceptable, then its information content
is too limited for social decision-making.

The motivation for employing a social welfare function was to be able to address
issues of equity as well as issues of efficiency. Fortunately a social welfare function is
not the only way to do this, and as this chapter will show, we can construct measures
of the economic situation that relate to equity and that are based on observable and
measurable information. This provides a set of tools that can be, and frequently are,
applied in economic policy analysis. They may not meet some of the requirements
of the ideal social welfare function, but they have the distinct advantage of being
practically implementable.

Inequality and poverty provide two alternative perspectives on the equity of the
income distribution. Inequality of income means that some households have higher
incomes than others—which is a basic source for an inequity in welfare. Poverty exists
when some households are too poor to achieve an acceptable standard of living. An
inequality measure is a means of assigning a single number to the observed income
distribution that reflects its degree of inequality. A poverty measure achieves the same
for poverty. Although measures of inequality and poverty are not directly social welfare
functions, the chapter will reveal the closeness of the link between the measures and
welfare.

The starting point of the chapter is a discussion of income. There are two aspects
to this: the definition of income and the comparison of income across families with
different compositions. In a setting of certainty, income is a clearly defined concept.
When there is uncertainty, differences can arise between ex ante and ex post definitions.
Given this, we look at alternative definitions and relate these to the treatment of income
for tax purposes. If two households differ in their composition (e.g., one household is
a single person and the other is a family of four), a direct comparison of their income
levels will reveal little about the standard of living they achieve. Instead, the incomes must be adjusted to take account of composition and then compared. The tool used to make the adjustment is an equivalence scale. We review the use of equivalence scales and some of the issues that they raise.

Having arrived at a set of correctly defined income levels that have been adjusted for family composition using an equivalence scale, it becomes possible to evaluate inequality and poverty. A number of the commonly used measures of each of these concepts are discussed and their properties investigated. Importantly, the link is drawn between measures of inequality and the welfare assumptions that are implicit within them. This leads into the idea of making the welfare assumptions explicit and building the measure up from these assumptions. To measure poverty, it is necessary to determine who is “poor,” which is achieved by choosing a level of income as the poverty line and labeling as poor all those who fall below it. As well as discussing measures of poverty, we also review issues concerning the definition of the poverty line and the very concept of poverty.

Although the aim of this chapter is to move away from utility concepts toward practical tools, it is significant that we keep returning to utility in the assessment and improvement of the tools. In attempting to refine, for example, an equivalence scale or a measure of inequality, it is found that it is necessary to comprehend the utility basis of the measure. Despite intentionally starting in a direction away from utility, the theory returns us back to utility on every occasion.

14.2 Measuring Income

What is income? The obvious answer is that it is the additional resources a consumer receives over a given period of time. The reference to a time period is important here, since income is a flow, so the period over which measurement takes place must be specified. Certainly evaluating the receipt of resources is the basis of the definition used in the assessment of income for tax purposes. This definition works in a practical setting but only in a backward-looking sense. What an economist needs in order to understand behavior, especially when choices are made in advance of income being received, is a forward-looking measure of income. If the flow of income is certain, then there is no distinction between backward- and forward-looking measures. It is when income is uncertain that differences emerge.

The relevance of this issue is that both inequality and poverty measures use income data as their basic input. The resulting measures will only be as accurate as the data that are employed to evaluate them. The data will be accurate when information is carefully
collected and a consistent definition is used of what is to be measured. To evaluate the level of inequality or poverty, a necessary first step is to resolve the issues surrounding the definition of income.

The classic backward-looking definition of income was provided by Henry C. Simons in 1938. This definition is “Personal income may be defined as the algebraic sum of (1) the market value of rights exercised in consumption and (2) the change in the value of the store of property rights between the beginning and end of the period in question.” The essential feature of this definition is that it makes an attempt to be inclusive so as to incorporate all income regardless of the source.

Although income definitions for tax purposes also adopt the backward-looking viewpoint, they do not precisely satisfy the Simons’s definition. The divergence arises through the practical difficulties of assessing some sources of income especially those arising from capital gains. According to the Simons’s definition, the increase in the value of capital assets should be classed as income. However, if the assets are not liquidated, the capital gain will not be realized during the period in question and will not be received as an income flow. For this reason capital gains are taxed only on realization. In the converse situation when capital losses are made, most tax codes place limits on the extent to which they can be offset against income.

We have so far worked with the natural definition of income as the flow of additional resources. To proceed further, it becomes more helpful to adopt a different perspective and to view the level of income by the benefits it can deliver. Since income is the means to achieve consumption, the flow of income during a fixed time period can be measured as the value of consumption that can be undertaken, while leaving the household with the same stock of wealth at the end of the period as it had at the start of the period. The benefit of this perspective is that it extends naturally to situations where the income flow is uncertain. Building on it, in 1939 John R. Hicks provided what is generally taken as the standard definition of income with uncertainty. This definition states that “income is the maximum value which a man can consume during a week and still expect to be as well-off at the end of the week as he was at the beginning.”

This definition can clearly cope with uncertainty, since it operates in expectational terms. But this advantage is also its major shortcoming when a move is made toward applications. Expectations may be ill-defined or even irrational, so evaluation of the expected income flow may be unreasonably high or low. A literal application of the definition would not count windfall gains, such as unexpected gifts or lottery wins, as income because they are not expected despite such gains clearly raising the potential level of consumption. For these reasons the Hicks definition of income is informative but not perfect.
These alternative definitions of income have highlighted the distinctions between ex ante and ex post measures. Assessments of income for tax purposes use the backward-looking viewpoint and measure income as all relevant payments received over the measurement period. Practical issues limit the extent to which some sources of income can be included, so the definition of income in tax codes does not precisely satisfy any of the formal definitions. This observation just reflects the fact that there is no unambiguously perfect definition of income.

14.3 Equivalence Scales

The fact that households differ in size and age distribution means that welfare levels cannot be judged just by looking at their income levels. A household of one adult with no children needs less income to achieve a given level of welfare than a household with two adults and one child. In the words of the economist William M. “Terence” Gorman, “When you have a wife and a baby, a penny bun costs threepence.” A larger household obviously needs more income to achieve a given level of utility, but the question is how much more income? Equivalence scales are the economist’s way of answering this question and provide the means of adjusting measured incomes into comparable quantities.

Differences among households arise in the number of adults and the number and ages of dependants. These are called demographic variables. The general problem in designing equivalence scales is to achieve the adjustment of observed income to take account of demographic differences in household composition. Several ways exist to do this, and these are now discussed.

The first approach to equivalence scales is based on the concept of minimum needs. A bundle of goods and services that is seen as representing the minimum needs for the household is identified. The exact bundle will differ among households of varying size, but it typically involves only very basic commodities. The cost of this bundle for families with different compositions is then calculated and the ratio of these costs for different families provides the equivalence scale. The first application of this approach was by Seebohm Rowntree in 1901 in his pioneering study of poverty. The bundle of goods employed was just a minimum acceptable quantity of food, rent, and a small allowance for “household sundries.” The equivalence scale was constructed by assigning the expenditure for a two-adult household with no children the index of 100 and measuring costs for all other household compositions relative to this. The scale obtained from expenditures calculated by Rowntree is given in the first column of table 14.1.
Table 14.1
Minimum needs equivalence scales

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Single person</td>
<td>60</td>
<td>59</td>
<td>78</td>
</tr>
<tr>
<td>Couple</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>+1 Child</td>
<td>124</td>
<td>122</td>
<td>120</td>
</tr>
<tr>
<td>+2 Children</td>
<td>161</td>
<td>144</td>
<td>151</td>
</tr>
<tr>
<td>+3 Children</td>
<td>186</td>
<td>166</td>
<td>178</td>
</tr>
<tr>
<td>+4 Children</td>
<td>223</td>
<td>188</td>
<td>199</td>
</tr>
</tbody>
</table>


interpretation of these figures is that the minimum needs of a couple with one child cost 24 percent more than for a couple with no children.

A similar approach was taken by William Beveridge in his 1942 construction of the expenditure requirements that provided the foundation for the introduction of social assistance in the United Kingdom. In addition to the goods in the bundle of Rowntree, Beveridge added fuel, light, and a margin for “inefficiency” in purchasing. Also the cost assigned to children increased with their age. The values of the Beveridge scale in the second column of table 14.1 are for children in the 5 to 10 age group.

The final column of the table is generated from the income levels that are judged to represent poverty in the United States for families with different compositions. The original construction of these poverty levels was undertaken by Mollie Orshansky in 1963. The method she used was to evaluate the cost of food for each family composition using the 1961 Economy Food Plan. Next it was observed that if expenditure on food, \( F \), constituted a proportion \( \theta \) of the family’s budget, then total needs would be \( \frac{1}{\theta} F \).

For a family of two, \( \frac{1}{\theta} \) was taken as 3.7, and for a family of three or more, \( \frac{1}{\theta} \) was 3. The exception to this process was to evaluate the cost for a single person as 80 percent of that of a couple. The minimum expenditures obtained have been continually updated, and the third column of the table gives the equivalence scale implied by the poverty line used in 2003.

Table 14.1 shows that these equivalence scales all assume that there are returns to scale in household size so that, for example, a family of two adults does not require twice the income of a single person. Observe also that the US poverty scale is relatively generous for a single person compared to the other two scales. The fact that the
single-person value was constructed in a different way from the other values for the poverty scale (as a fixed percentage of that for a couple rather than as a multiple of food costs) has long been regarded as a contentious issue. Furthermore only for the Beveridge scale is the cost of additional children constant. The fact that the cost of children is nonmonotonic for the poverty scale is a further point of contention.

There are three major shortcomings of this method of computing equivalence scales. First, by focusing on the cost of meeting a minimum set of needs, they are inappropriate for applying to incomes above the minimum level. Second, they are dependent on an assessment of what constitutes minimum needs—and this can be contentious. Most important, the scales do not take into account the process of optimization by the households. The consequence of optimization is that as income rises, substitution between goods can take place, and the same relativities need no longer apply. Alternative methods of constructing equivalence scales that aim to overcome these difficulties are now considered.

In a similar way to the Orshansky construction of the US poverty scale, the Ernst Engel approach to equivalence scales is based on the hypothesis that the welfare of a household can be measured by the proportion of its income that is spent on food. This is a consequence of Engel’s law, which asserts that the share of food in expenditure falls as income rises. If this is accepted, equivalence scales can be constructed for households of different compositions by calculating the income levels at which their expenditure share on food is equal. This is illustrated in figure 14.1 in which the expenditure share on food, as a function of income, is shown for two households with family compositions \( d^1 \) and \( d^2 \). For example, \( d^1 \) may refer to a couple and \( d^2 \) to a couple with one child. Incomes \( M^1 \) and \( M^2 \) lead to the same expenditure share, \( s \), and so are equivalent for the Engel method. The equivalence scale is then formed from the ratio \( \frac{M^2}{M^1} \).

Although Engel’s law may be empirically true, it does not necessarily provide a basis for making welfare comparisons, since it leaves unexplored the link between household composition and food expenditure. In fact there is ground for believing that the Engel method overestimates the cost of additional children because a child is largely a food-consuming addition to a household. If this is correct, a household compensated sufficiently to restore the share of food in its expenditure to its original level after the addition of a child would have been overcompensated with respect to other commodities. The approach of Engel has been extended to the more general iso-prop method in which the expenditure shares of a basket of goods, rather than simply food, becomes the basis for the construction of scales. However, considering a basket of goods does not overcome the basic shortcomings of the Engel method.
A further alternative is to select for attention a set of goods that are consumed only by adults, termed “adult goods,” and such that the expenditure on them can be treated as a measure of welfare. Typical examples of such goods that have been used in practice are tobacco and alcohol. If these goods have the property that changes in household composition only affect their demand via an income effect (so changes in household composition do not cause substitution between commodities), then the extra income required to keep their consumption constant when household composition changes can be used to construct an equivalence scale. The use of adult goods to construct an equivalence scale is illustrated in figure 14.2. On the basis that they generate the same level of demand, $\bar{x}$, as family composition changes, the income levels $M^1$ and $M^2$ can be classed as equivalent, and the equivalence scale can be constructed from their ratio.

There are also a number of difficulties with this approach. It rests on the hypotheses that consumption of adult goods accurately reflects welfare and that household composition affects the demand for these goods only via an income effect. Furthermore the ratio of $M^1$ to $M^2$ will depend on the level of demand chosen for the comparison except in the special case where the demand curves are straight lines through the origin. The ratios may also vary for different goods. This leads into a further problem of forming some average ratio out of the ratios for the individual goods.
All of the methods described so far have attempted to derive the equivalence scale from an observable proxy for welfare. A general approach that can, in principal, overcome the problems identified in the previous methods is illustrated in figure 14.3. To understand this figure, assume that there are just two goods available. The outer indifference curve represents the consumption levels of these two goods necessary for a family of composition \( d^2 \) to obtain welfare level \( U^* \), and the inner indifference curve the consumption requirements for a family with composition \( d^1 \) to obtain the same utility. The extent to which the budget line has to be shifted outward to reach the higher curve determines the extra income required to compensate for the change in family structure. This construction incorporates both the potential change in preferences as family composition changes and the process of optimization subject to budget constraint by the households.

To formalize this process, let the household have preferences described by the utility function \( U(x_1, x_2; d) \), where \( x_i \) is the level of consumption of good \( i \) and \( d \) denotes information on family composition. For example, \( d \) will describe the number of adults, the number and ages of children, and any other relevant information. The consumption plan needed to attain a given utility level, \( U \), at least cost is the solution to

\[
\min_{\{x_1, x_2\}} p_1 x_1 + p_2 x_2 \quad \text{subject to} \quad U(x_1, x_2; d) \geq U^*.
\]  

(14.1)
Denoting the (compensated) demand for good $i$ by $x_i(U^*, d)$, the minimum cost of attaining utility $U$ with characteristics $d$ is then given by

$$M(U^*, d) = p_1x_1(U^*, d) + p_2x_2(U^*, d).$$  \hspace{1cm} (14.2)

The equivalent incomes at utility $U^*$ for two households with compositions $d^1$ and $d^2$ are then given by $M(U^*, d^1)$ and $M(U^*, d^2)$. The equivalence scale is derived by computing their ratio. The important point obtained by presenting the construction in this way is the observation that the equivalence scale will generally depend on the level of utility at which the comparison is made. If it does, there can be no single equivalence scale that works at all levels of utility.

The construction of an equivalence scale from preferences makes two further issues apparent. First, the minimum needs and budget share approaches do not take account of how changes in family structure may shift the indifference map. For instance, the pleasure of having children may raise the utility obtained from any given consumption plan. With the utility approach it then becomes cheaper to attain each indifference curve, so the value of the equivalence scale falls as family size increases. This conclusion, of course, conflicts with the basic sense that it is more expensive to support a larger family.

Figure 14.3
General equivalence scale
The second problem centers around the use of a household utility function. Many economists would argue that a household utility function cannot exist; instead, they would observe that households are composed of individuals with individual preferences. Under the latter interpretation, the construction of a household utility function suffers from the difficulties of preference aggregation identified by Arrow’s Impossibility Theorem. Among the solutions to this problem now being investigated is to look within the functioning of the household and to model its decisions as the outcome of an efficient resource allocation process.

14.4 Inequality Measurement

Inequality is a concept that has immediate intuitive implications. The existence of inequality is easily perceived: differences in living standards between the rich and poor are only too obvious both across countries and, sometimes to a surprising extent, within countries. The obsession of the media with wealth and celebrity provides a constant reminder of just how rich the rich can be. An increase in inequality can also be understood at a basic level. If the rich become richer, and the poor become poorer, then inequality must have increased.

The substantive economic questions about inequality arise when we try to move beyond these generalizations to construct a quantitative measure of inequality. Without a quantitative measure it is not possible to provide a precise answer to questions about inequality. For example, a measure is required to determine which of a range of countries has the greatest level of inequality and to determine whether inequality has risen or fallen over time.

What an inequality measure must do is to take data on the distribution of income and generate a single number that captures the inequality in that distribution. A first approach to constructing such a measure is to adopt a standard statistical index. We describe the most significant of these indexes. Looking at the statistical measures reveals that there are properties, particularly how the measure is affected by transfers of income between households, that we may wish an inequality measure to possess. These properties can also be used to assess the acceptability of alternative measures. It is also shown that implicit within a statistical measure are a set of welfare implications. Rather than just accept these implications, the alternative approach is explored of making the welfare assumptions explicit and building the inequality measure on them.
14.4.1 The Setting

The intention of an inequality measure is to assign a single number to an income distribution that represents the degree of inequality. This section sets out the notation employed for the basic information that is input into the measure and defines precisely what is meant by a measure.

We assume that there are \( H \) households and label these \( h = 1, \ldots, H \). The labeling of the households is chosen so that the lower is the label, the lower is the household’s income. The incomes, \( M^h \), then form an increasing sequence with

\[
M^1 \leq M^2 \leq M^3 \leq \ldots \leq M^H.
\] (14.3)

The list \( \{M^1, \ldots, M^H\} \) is the income distribution whose inequality we wish to measure. Given the income distribution, the mean level of income, \( \mu \), is defined by

\[
\mu = \frac{1}{H} \sum_{h=1}^{H} M^h.
\] (14.4)

The purpose of an inequality measure is to assign a single number to the distribution \( \{M^1, \ldots, M^H\} \). Let \( I(M^1, \ldots, M^H) \) be an inequality measure. Then income distributions \( \{\hat{M}^1, \ldots, \hat{M}^H\} \) has greater inequality than distribution \( \{\tilde{M}^1, \ldots, \tilde{M}^H\} \) if \( I(\hat{M}^1, \ldots, \hat{M}^H) > I(\tilde{M}^1, \ldots, \tilde{M}^H) \). Typically the inequality measure is constructed so that a value of 0 represents complete equality (the position where all incomes are equal) and a value of 1 represents maximum inequality (all income is received by just one household).

The issues that arise in inequality measurement are encapsulated in determining the form that the function \( I(M^1, \ldots, M^H) \) should take. We now investigate some alternative forms and explore their implications.

14.4.2 Statistical Measures

Under the heading of “statistical” fall inequality measures that are derived from the general statistical literature. That is, the measures have been constructed to characterize the distribution of a set of numbers without thought of any explicit economic application or motivation. Even so, the discussion will later show that these statistical measures make implicit economic value judgments. Accepting any one of these measures as the “correct” way to measure inequality means the acceptance of these implicit
assumptions. The measures that follow are presented in approximate order of sophistication. Each is constructed to take a value between 0 and 1, with a value of 0 occurring when all households have identical income levels.

Probably the simplest conceivable measure, the range calculates inequality as being the difference between the highest and lowest incomes expressed as a proportion of total income. As such, it is a very simple measure to compute. The definition of the range, \( R \), is

\[
R = \frac{M^H - M^1}{H\mu}.
\]  

The division by \( H\mu \) in (14.5) is a normalization that ensures the index is independent of the scale of incomes (or the units of measurement of income). Any index that has this property of independence is called a relative index.

As an example of the use of the range, consider the income distribution \( \{1, 3, 6, 9, 11\} \). For this distribution \( \mu = 6 \) and

\[
R = \frac{11 - 1}{5 \times 6} = 0.3333.
\]  

The failure of the range to take account of the intermediate part of the distribution can be illustrated by taking income from the second household in the example and giving it to the fourth to generate new income distribution \( \{1, 1, 6, 11, 11\} \). This new distribution appears to be more unequal than the first, yet the value of the range remains at \( R = 0.3333 \).

Given the simplicity of its definition, it is not surprising that the range has deficiencies. Most important, the range takes no account of the dispersion of the income distribution between the highest and the lowest incomes. Consequently it is not sensitive to any features of the income distribution between these extremes. For instance, an income distribution with most of the households receiving close to the maximum income would be judged just as unequal as one in which most received the lowest income. An ideal measure should possess more sensitivity to the value of intermediate incomes than the range.

The relative mean deviation, \( D \), takes account of the deviation of each income level from the mean so that it is dependent on intermediate incomes. It does this by calculating the absolute value of the deviation of each income level from the mean and then summing. This summation process gives equal weight to deviations both above and below the mean and implies that \( D \) is linear in the size of deviations. Formally, \( D \) is defined by
Chapter 14: Inequality and Poverty

\[ D = \frac{\sum_{h=1}^{H} |\mu - M_h|}{2(H - 1)\mu}. \]  
(14.7)

The division by \(2(H - 1)\mu\) again ensures that \(D\) takes values between 0 and 1.

The advantage of the relative mean deviation over the range is that it takes account of the entire income distribution and not just the end points. Taking the example used for the range, the inequality in the distribution \(\{1, 3, 6, 9, 11\}\) as measured by \(D\) is

\[ D = \frac{|-5| + |-3| + |0| + |3| + |5|}{2 \times 4 \times 6} = 0.3333, \]  
(14.8)

and the inequality of \(\{1, 1, 6, 11, 11\}\) is

\[ D = \frac{|-5| + |-5| + |0| + |5| + |5|}{2 \times 4 \times 6} = 0.4167. \]  
(14.9)

Unlike the range, the relative mean deviation measures the second distribution as having more inequality. Due to the division by \(2(H - 1)\mu\) it is easily seen that \(D = 1\) with the maximum inequality distribution \(\{0, 0, 0, 0, 30\}\) where all income is received by just one household.

Although it does take account of the entire distribution of income, the linearity of \(D\) has the implication that it is insensitive to transfers from richer to poorer households when the households involved in the transfer remain on the same side of the mean income level. To see an example of this, assume that the mean income level is \(\mu = \$20,000\). Now take two households with incomes \(\$25,000\) and \(\$100,000\). Transferring \(\$4,000\) from the poorer of these two households to the richer, so that the income levels become \(\$21,000\) and \(\$104,000\), does not change the value of \(D\)—one term in the summation rises by \(\$4,000\) and the other falls by \(\$4,000\). (Notice that if the two households were on different sides of the mean, then a similar transfer would raise two terms in the summation by \(\$4,000\) and increase inequality.) The fact that \(D\) can be insensitive to transfers seems unsatisfactory, since it is natural to expect that a transfer from a poorer household to a richer one should raise inequality.

This line of reasoning is enshrined in the Pigou–Dalton Principle of Transfers, which is a central concept in the theory of inequality measurement. The basis of this principle is precisely the requirement that any transfer from a poor household to a rich one must increase inequality regardless of where the two households are located in the income distribution.
**Definition 14.1** (Pigou–Dalton Principle of Transfers) The inequality index must decrease if there is a transfer of income from a richer household to a poorer household that preserves the ranking of the two households in the income distribution and leaves total income unchanged.

Any inequality measure that satisfies this principle is said to be *sensitive to transfers*. The Pigou–Dalton Principle is generally viewed as a feature that any acceptable measure of inequality should possess and is therefore expected in an inequality measure. Neither the range nor the relative mean deviation satisfy this principle.

The reason why $D$ is not sensitive to transfers is its linearity in deviations from the mean. The removal of the linearity provides the motivation for considering the *coefficient of variation*, which is defined using the sum of squared deviations. The procedure of forming the square places more weight on incomes that are further away from the mean and so introduces a sensitivity to transfers. The coefficient of variation, $C$, is defined by

$$C = \frac{\sigma}{\mu [H - 1]^{1/2}}, \quad (14.10)$$

where $\sigma^2 = \frac{\sum_{h=1}^{H} (M_h - \mu)^2}{H}$ is the variance of the income distribution, so $\sigma$ is its standard deviation. The division by $\mu [H - 1]^{1/2}$ ensures the $C$ lies between 0 and 1. For the income distribution $\{1, 3, 6, 9, 11\}$, $\sigma^2 = \frac{(-5)^2 + (-3)^2 + 0^2 + 3^2 + 5^2}{5} = 13.6$, so

$$C = \frac{[13.6]^{1/2}}{6[4]^{1/2}} = 0.3073, \quad (14.11)$$

and for $\{1, 1, 6, 11, 11\}$, $\sigma^2 = 20$, giving

$$C = \frac{[20]^{1/2}}{6[4]^{1/2}} = 0.3727. \quad (14.12)$$

To see that the coefficient of variation satisfies the Pigou–Dalton Principle, consider a transfer of an amount of income $d \varepsilon$ from household $i$ to household $j$, with the households chosen so that $M_i < M_j$. Then

$$\frac{dC}{d\varepsilon} = \frac{1}{\mu [H - 1]^{1/2}} \frac{d\sigma}{d\varepsilon} = \frac{M_j - M_i}{\sigma H \mu [H - 1]^{1/2}} > 0, \quad (14.13)$$

so the transfer from the poorer household to the richer household decreases measured inequality as required by the Pigou–Dalton Principle. It should be noted that the value of
the change in $C$ depends on the difference between the incomes of the two households. This has the consequence that a transfer of $100$ of income from a household with an income of $1,000,100$ to one with an income of $999,900$ produces the same change in $C$ as a transfer of $100$ between households with incomes $1,100$ and $900$. Most interpretations of equity would suggest that the latter transfer should be of greater consequence for the index because it involves two households of relatively low incomes. This reasoning suggests that satisfaction of the Pigou–Dalton Principle may not be a sufficient requirement for an inequality measure; the manner in which the measure satisfies it may also matter.

Before moving on to further inequality measures, it is worth describing the Lorenz curve. The Lorenz curve is a helpful graphical device for presenting a summary representation of an income distribution, and it has played an important role in the measurement of inequality. Although not strictly an inequality measure as defined above, Lorenz curves are considered because of their use in illustrating inequality and the central role they play in the motivation of other inequality indexes.

The Lorenz curve is constructed by arranging the population in order of increasing income and then graphing the proportion of income going to each proportion of the population. The graph of the Lorenz curve therefore has the proportion of population on the horizontal axis and the proportion of income on the vertical axis. If all households in the population had identical incomes the Lorenz curve would then be the diagonal line connecting the points $(0, 0)$ and $(1, 1)$. If there is any degree of inequality, the ordering in which the households are taken ensures that the Lorenz curve lies below the diagonal since, for example, the poorest half of the population must have less than half the total income.

To see how the Lorenz curve is plotted, consider a population of $10$ with income distribution $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. The total quantity of income is $55$, so the first household (which represents $10$ percent of the population) receives $\frac{1}{55} \times 100$ percent of the total income. This is the first point plotted in the lower left corner of figure 13.4. Taking the two lowest income households (which are $20$ percent of the population), we have their combined income as $\frac{3}{55} \times 100$ percent of the total. Adding the third household awards $30$ percent of the population $\frac{6}{55} \times 100$ percent of total income. Proceeding in this way, we plot the ten points in the figure. Joining them gives the Lorenz curve. In summary, the larger the population, the smoother is this curve.

The Lorenz curve can be employed to unambiguously rank some income distributions with respect to income inequality. This claim is based on the fact that a transfer of income from a poor household to a richer household moves the Lorenz curve farther away from the diagonal. (This can be verified by re-plotting the Lorenz curve in figure 14.4 for
the income distribution \{1, 1, 3, 4, 5, 6, 7, 8, 10, 10\}, which is the same as the original except for the transfer of one unit from household 2 to household 9.) Because of this property the Lorenz curve satisfies the Pigou–Dalton Principle, with the curve farther from the diagonal indicating greater inequality.

Income distributions that can, and cannot, be ranked are displayed in figure 14.5. In the left-hand panel, the Lorenz curve for income distribution \(B\) lies entirely outside that for income distribution \(A\). In such a case distribution \(B\) unambiguously has more inequality than \(A\). One way to see this is to observe that distribution \(B\) can be obtained from distribution \(A\) by transferring income from poor households to rich households. Applying the Pigou–Dalton Principle, we see that this raises inequality. If the Lorenz curves representing the distributions \(A\) and \(B\) cross, it is not possible to obtain an unambiguous conclusion by the Lorenz curve alone. The Lorenz curve therefore provides only a partial ranking of income distributions. Despite this limitation the Lorenz curve is still a popular tool in applied economics, since it presents very convenient and easily interpreted visual summary of an income distribution.

The next measure, the Gini, has been the subject of extensive attention in discussions of inequality measurement and has been much used in applied economics. The Gini, \(G\), can be expressed by considering all possible pairs of incomes and out of each pair selecting the minimum income level. Summing the minimum income levels and
dividing by $H^2\mu$ to ensure a value between 0 and 1 provides the formula for the Gini:

$$G = 1 - \frac{1}{H^2\mu} \sum_{i=1}^{H} \sum_{j=1}^{H} \min\left\{M^i, M^j\right\}. \quad (14.14)$$

It should be noted that in the construction of this measure, each level of income is compared to itself as well as all other income levels. For example, if there are three income levels \{3, 5, 10\}, the value of the Gini is

$$G = 1 - \frac{1}{3^2 \times 6} \left[ \min\{3, 3\} + \min\{3, 5\} + \min\{3, 10\} + \min\{5, 3\} + \min\{5, 5\} + \min\{5, 10\} + \min\{10, 3\} + \min\{10, 5\} + \min\{10, 10\} \right]$$

$$= 1 - \frac{1}{54} \left[ 3 + 3 + 3 + 5 + 3 + 5 + 3 + 5 + 10 \right]$$

$$= 0.259. \quad (14.15)$$

By counting the number of times each income level appears, we can also write the Gini as

$$G = 1 - \frac{1}{H^2\mu} \left[ (2H - 1) M^1 + (2H - 3) M^2 + (2H - 5) M^3 + \ldots + M^H \right]. \quad (14.16)$$
This second form of the Gini makes its computation simpler but hides the construction behind the measure.

The Gini also satisfies the Pigou–Dalton Principle. This can be seen by considering a transfer of income of size \( \Delta > 0 \) from household \( i \) to household \( j \), with the households chosen so that \( M^j > M^i \). From the ranking of incomes this implies \( j > i \). Then

\[
\Delta G = \frac{2}{H^2 \mu} (j - i) \Delta > 0,
\]

as required. In the case of the Gini, the effect of the transfer of income on the measure depends only on the locations of \( i \) and \( j \) in the income distribution. For example, a transfer from the household at position \( i = 1 \) to the household at position \( j = 11 \) counts as much as one from position \( i = 151 \) to position \( j = 161 \). It might be expected that an inequality should be more sensitive to transfers between households low in the income distribution.

There is an important relationship between the Gini and the Lorenz curve. As shown in figure 14.6, the Gini is equal to the area between the Lorenz curve and the line of equality as a proportion of the area of the triangle beneath the line of equality. As the area of the triangle is \( \frac{1}{2} \), the Gini is twice the area between the Lorenz curve and the equality line. This definition makes it clear that the Gini, in common with \( R \), \( C \), and \( D \), can be used to rank distributions when the Lorenz curves cross, since the relevant area is always well defined. All these measures provide a stronger ranking of income distributions than the Lorenz curve; hence they must each impose additional restrictions that allow a comparison to be made between distributions even when their Lorenz curves cross.
A final statistical measure that displays a different form of sensitivity to transfers is the Theil entropy measure. This measure is drawn from information theory and is used in that context to measure the average information content of a system of information. The definition of the Theil entropy measure, $T$, is given by

$$T = \frac{1}{\log H} \sum_{h=1}^{H} \frac{M^h}{H \mu} \left[ \log \frac{M^h}{H \mu} - \log \frac{1}{H} \right]$$

(14.18)

The effect of an income transfer, $d\epsilon$, between households $i$ and $j$ on the entropy index is given by

$$\frac{dT}{d\epsilon} = \frac{1}{H \log H} \log \frac{M^j}{M^i} < 0,$$

(14.19)

so the entropy measure also satisfies the Pigou–Dalton Principle. For the Theil entropy measure, the change is dependent on the relative incomes of the two households involved in the transfer. This provides an alternative form of sensitivity to transfers.

### 14.4.3 Inequality and Welfare

The analysis of the statistical measures of inequality has made reference to “acceptable” criteria for a measure to possess. One of these was made explicit in the Pigou–Dalton Principle, while other criteria relating to additional desirable sensitivity properties have been implicit in the discussion. To be able to say that something is acceptable or not implies that there is some notion of distributive justice or social welfare underlying the judgment. It is then interesting to consider the relationship between inequality measures and welfare.

The first issue to address is the extent to which income distributions can be ranked in terms of welfare with minimal restrictions imposed on the social welfare function. To investigate this, let the level of social welfare be determined by the function $W = W (M^1, \ldots, M^H)$. It is assumed that this social welfare function is symmetric and concave. Symmetry means that the level of welfare is unaffected by changing the ordering of the households. This is just a requirement that all households are treated equally. Concavity ensures that the indifference curves of the welfare function have...
the standard shape with mixtures preferred to extremes. This assumption imposes a concern for equity on the welfare function.

The critical theorem relating the ranking of income distributions to social welfare is now given.

**Theorem 14.2 (Atkinson)** Consider two distributions of income with the same mean. If the Lorenz curves for these distributions do not cross, every symmetric and concave social welfare function will assign a higher level of welfare to the distribution whose Lorenz curve is closest to the main diagonal.

The proof of this theorem is very straightforward. Since the welfare function is symmetric and concave, it follows that \( \frac{\partial W}{\partial M_i} \geq \frac{\partial W}{\partial M_j} \) if \( M_i < M_j \). Hence the marginal social welfare of income is greater for a household lower in the income distribution. If the two Lorenz curves do not cross, the income distribution represented by the inner one (that closest to the main diagonal) can be obtained from that of the outer one by transferring income from richer to poorer households. Since the marginal social welfare of income to the poorer households is never less than that from richer, this transfer must raise welfare as measured by any symmetric and concave social welfare function.

The converse of this theorem is that if the Lorenz curves for two distributions cross, then two symmetric and concave social welfare functions can be found that will rank the two distributions differently. This is because the income distributions of two Lorenz curves that cross are not related by simple transfers from rich to poor. So, if the Lorenz curves do cross, the income distributions cannot be unambiguously ranked without specifying the social welfare function.

Taken together, the theorem and its converse show that the Lorenz curve provides the most complete ranking of income distributions that is possible without our making assumptions on the form of the social welfare function other than symmetry and concavity. To achieve a complete ranking when the Lorenz curves cross requires restrictions to be placed on the structure of the social welfare function. In addition any measure of inequality is necessarily stronger than the Lorenz curve because it generates a complete ranking of distributions. This is true of all the statistical measures, which is why it can be argued that they all carry implicit welfare judgments.

This argument can be taken a stage further. It is in fact possible to construct the social welfare function that is implied by an inequality measure. To see how this can be done, consider the Gini. Assume that the total amount of income available is constant. Any redistribution of this that leaves the Gini unchanged must leave the implied level of welfare unchanged. A redistribution of income will not affect the Gini if the term
The welfare function must thus be a function of this expression. Furthermore, the Gini is defined to be independent of the total level of income, but a welfare function will increase if total income rises and distribution is unaffected. This can be incorporated by not dividing through by the mean level of income. Putting these arguments together, the welfare function implied by the Gini is given by

$$W_G(M) = \frac{1}{H^2} \left[ [2H - 1]M^1 + [2H - 3]M^2 + \ldots + M^H \right].$$

(14.20)

The form of $W_G(M)$ is interesting, since it shows that the Gini implies a social welfare function that is linear in incomes. It also shows a clear structure of increasing welfare weights for lower income consumers. The welfare function further has indifference curves that are straight lines above and below the line of equal incomes but kinked on this line. This is illustrated in figure 14.7.

In the same way a welfare function can be constructed for all the statistical measures. Therefore acceptance of the measure is acceptance of the implied welfare function. As shown by the linear social indifference curves and increasing welfare weights for the Gini social welfare function, the implied welfare functions can have a very restrictive form. We do not need to merely accept such welfare restrictions. The fact that each inequality measure implies a social welfare function suggests that the relationship can be inverted to move from a social welfare function to an inequality measure. By assuming
a social welfare function at the outset, it is possible to make welfare judgments explicit and, by deriving the inequality measure from the social welfare function, to ensure that these judgments are incorporated in the inequality measure.

To implement this approach, assume that the social welfare function is utilitarian with

\[ W = \sum_{h=1}^{H} U(M^h). \]  

(14.21)

The household utility of income function, \( U(M) \), is taken to satisfy the conditions that \( U'(M) > 0 \) and \( U''(M) < 0 \). The utility function \( U(M) \) can either be the households’ true cardinal utility function or be chosen by the policy analyst as in the evaluation of the utility of income to each household. In this second interpretation, since social welfare is obtained by summing the individual utilities, the importance given to equity can be captured in the choice of \( U(M) \). This is because increasing the concavity of the utility function places a relatively higher weight on low incomes in the social welfare function.

A measure of inequality can be constructed from the social welfare function by defining \( M_{EDE} \) as the solution to

\[ \sum_{h=1}^{H} U(M^h) = HU(M_{EDE}). \]  

(14.22)

\( M_{EDE} \) is called the equally distributed equivalent income and is the level of income that, if given to all households, would generate the same level of social welfare as the initial income distribution. Using \( M_{EDE} \), the Atkinson measure of inequality is defined by

\[ A = 1 - \frac{M_{EDE}}{\mu}. \]  

(14.23)

For the case of two households the construction of \( M_{EDE} \) is illustrated in figure 14.8. The initial income distribution is given by \( \{M^1, M^2\} \), and this determines the relevant indifference curve of the social welfare function. \( M_{EDE} \) is found by moving around this indifference curve to the 45 degree line where the two households’ incomes are equal. The figure makes clear that because of the concavity of the social indifference curve, \( M_{EDE} \) is less than the mean income, \( \mu \). This fact guarantees that \( 0 \leq A \leq 1 \). Furthermore, for a given level of mean income, a more diverse income distribution will achieve a lower social indifference curve and be equivalent to a lower \( M_{EDE} \).
The flexibility in this measure lies in the freedom of choice of the household utility of income function. Given the assumption of a utilitarian social welfare function, it is the household utility that determines the importance attached to inequality by the measure. One commonly used form of utility function is

\[ U(M) = \frac{M^{1-\varepsilon}}{1-\varepsilon}, \quad \varepsilon \neq 1, \quad (14.24) \]

which allows the welfare judgments of the policy analyst to be contained in the chosen value of the parameter \( \varepsilon \). The value of \( \varepsilon \) determines the degree of concavity of the utility function: it becomes more concave as \( \varepsilon \) increases. An increase in concavity raises the relative importance of low incomes because it causes the marginal utility of income to decline at a faster rate. The utility function is isoelastic, and concave if \( \varepsilon \geq 0 \). When \( \varepsilon = 1 \), \( U(M) = \log(M) \), and when \( \varepsilon = 0 \), \( U(M) = M \).

The Atkinson measure can be illustrated using the example of the income distribution \( \{1, 3, 6, 9, 11\} \). If \( \varepsilon = \frac{1}{2} \) the household utility function is \( U = 2M^{1/2} \) so the level of social welfare is

\[ W = 2 \times 1^{1/2} + 2 \times 3^{1/2} + 2 \times 6^{1/2} + 2 \times 9^{1/2} + 2 \times 11^{1/2} = 22.996. \quad (14.25) \]

The equally distributed equivalent income then solves
\[ 5 \times \left[ 2 \times (M)^{1/2} \right] = 22.996, \]  
so \( M_{EDE} = 5.2882 \). This gives the value of the Atkinson measure as 
\[ A = 1 - \frac{5.2882}{6} = 0.1186. \]  

14.4.4 An Application

As has been noted in the discussion, inequality measures are frequently used in practical policy analysis. Table 14.2 summarizes the results of an OECD study into the change in inequality over time in a wide range of countries. This is undertaken by calculating inequality at two points in time and determining the percentage change in the measure. Inequality is calculated for income prior to taxes and transfers, and for income after taxes and transfers. The difference between the inequality levels in these two situations gives an insight into the extent to which the tax and transfer system succeeds in redistributing income.

Looking at the results, in all cases inequality is smaller after taxes and transfers than before, so the tax systems in the countries studied are redistributive. For instance, in Denmark inequality is 0.0420 when measured by Gini before taxes and transfers but only 0.0217 after. The second general message of the results is that inequality has tended to rise in these countries—only in three cases has it been reduced, and in every case this is after taxes and transfers.

It is also interesting to look at the rankings of inequality and changes in inequality under the different measures. If there is general agreement for different measures, then we can be reassured that the choice of measure is not too critical for what we observe. For the level of inequality all four measures are in agreement for both the before-tax and after-tax cases except for the SCV (squared coefficient of variation), which reverses the after-taxes and transfers ranking of Denmark and Sweden, and the Atkinson, which reverses the before-taxes and transfers ranking of Denmark and the United States. For these four measures there is a considerable degree of consistency in the rankings. Taking the majority opinion, observe that before taxes and transfers the ranking (with the highest level of inequality first) is Italy, Sweden, United States, Denmark, and Japan. After the operation of taxes and transfers this ranking becomes Italy, United States, Japan, Sweden, and Denmark. This change in rankings is evidence of the highly redistributive tax and transfer systems operated in the Nordic countries.

The rankings for the change in inequality are not quite as consistent across the four measures, but there is still considerable agreement. The majority order for the before
Chapter 14: Inequality and Poverty

Table 14.2
Inequalities before and after taxes and transfers

<table>
<thead>
<tr>
<th>Measure</th>
<th>SCV Before</th>
<th>SCV After</th>
<th>Gini Before</th>
<th>Gini After</th>
<th>Atkinson Before</th>
<th>Atkinson After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denmark 1994</td>
<td>0.671</td>
<td>0.229</td>
<td>0.420</td>
<td>0.217</td>
<td>0.209</td>
<td>0.041</td>
</tr>
<tr>
<td>% Change 1983–1994</td>
<td>4.9</td>
<td>2.0</td>
<td>11.2</td>
<td>−4.9</td>
<td>25.3</td>
<td>−11.1</td>
</tr>
<tr>
<td>Italy 1993</td>
<td>1.19</td>
<td>0.584</td>
<td>0.570</td>
<td>0.345</td>
<td>0.299</td>
<td>0.105</td>
</tr>
<tr>
<td>% Change 1984–1993</td>
<td>59.6</td>
<td>44.7</td>
<td>20.8</td>
<td>12.8</td>
<td>43.8</td>
<td>33.1</td>
</tr>
<tr>
<td>Japan 1994</td>
<td>0.536</td>
<td>0.296</td>
<td>0.340</td>
<td>0.265</td>
<td>0.124</td>
<td>0.059</td>
</tr>
<tr>
<td>% Change 1984–1994</td>
<td>33.7</td>
<td>21.7</td>
<td>14.0</td>
<td>4.9</td>
<td>47.3</td>
<td>10.9</td>
</tr>
<tr>
<td>Sweden 1995</td>
<td>0.894</td>
<td>0.217</td>
<td>0.487</td>
<td>0.230</td>
<td>0.262</td>
<td>0.049</td>
</tr>
<tr>
<td>% Change 1975–1995</td>
<td>49.1</td>
<td>36.9</td>
<td>17.2</td>
<td>−1.0</td>
<td>28.7</td>
<td>3.2</td>
</tr>
<tr>
<td>United States 1995</td>
<td>0.811</td>
<td>0.441</td>
<td>0.455</td>
<td>0.344</td>
<td>0.205</td>
<td>0.100</td>
</tr>
<tr>
<td>% Change 1974–1995</td>
<td>32.0</td>
<td>25.4</td>
<td>13.1</td>
<td>10.0</td>
<td>19.6</td>
<td>18.6</td>
</tr>
</tbody>
</table>

Source: OECD ECO/WKP(98)2.
Notes: The squared coefficient of variation (SCV) is defined by \[ SCV = \left[ H - 1 \right] C. \] For the Atkinson measure, \( \varepsilon = 0.5 \).

taxes and transfers case (with the greatest increase in inequality first) is Italy, Sweden, Japan, United States, and Denmark. The Atkinson measure places Japan at the top and reverses Denmark and the United States. For the after-taxes and transfers ranking, the Gini and the Atkinson measures produce the same ranking, but the SCV places Sweden above the United States and Japan. But what is clear is the general agreement on an increase in inequality.

The review of this application has shown that the different measure can produce a fairly consistent picture about ranking by inequality, about the changes in inequality, and on the effect of taxes and transfers. Despite the differences emphasized in the analysis of the measures, when put into practice in this way, the differences need not lead to widespread disagreement among the measures. In fact a fairly harmonious picture can emerge.

14.5 Poverty

The essential feature of poverty is the possession of fewer resources than are required to achieve an acceptable standard of living. What constitutes poverty can be understood in the same intuitive way as what constitutes inequality, but similar issues about the correct measure arise again once we attempt to provide a quantification. This section
Part V: Equity and Distribution

first discusses concepts of poverty and the poverty line, and then proceeds to review a number of common poverty measures.

14.5.1 Poverty and the Poverty Line

Before measuring poverty, it is first necessary to define it. It is obvious that poverty refers to a situation involving a lack of income and a consequent low level of consumption and welfare. What is not so clear is the standard against which the level of income should be judged. Two possibilities arise in this context: an absolute conception of poverty and a relative one. The distinction between these has implications for changes in the level of poverty over time and the success of policy in alleviating poverty.

The concept of **absolute poverty** assumes that there is some fixed minimum level of consumption (and hence of income) that constitutes poverty and is independent of time or place. Such a minimum level of consumption can be a diet that is just sufficient to maintain health and limited housing and clothing. Under the concept of absolute poverty, if the incomes of all households rise, there will eventually be no poverty. Although a concept of absolute poverty was probably implicit in early studies of poverty, such as that of Rowntree in 1901, the appropriateness of absolute poverty has since generally been rejected. In its place has been adopted the notion of relative poverty.

**Relative poverty** is not a recent concept. Even in 1776 Adam Smith was defining poverty as the lack of necessities, where necessities are defined as “what ever the custom of the country renders it indecent for creditable people, even of the lowest order, to be without.” This definition makes it clear that relative poverty is defined in terms of the standards of a given society at a given time and that the level that represents poverty rises as does the income of that society. Operating under a relative standard, it becomes much more difficult to eliminate poverty. Relative poverty has also been defined in terms of the ability to “participate” in society. Poverty then arises whenever a household possesses insufficient resources to allow it to participate in the customary activities of its society.

The starting point for the measurement of poverty is to set a poverty line that separates those viewed as living in poverty from those who are not. Of course, this poverty line applies to the incomes levels after application of an equivalence scale. Whether poverty is viewed as absolute or relative matters little for setting a poverty line at any particular point in time (though advocates of an absolute poverty concept may choose to set it lower). Where the distinction matters is whether and how the poverty line is adjusted over time. If an absolute poverty standard were adopted, then there would be no revision.
Conversely, with relative poverty the level of the line would rise or fall in line with average incomes.

In practice, poverty lines have often been determined by following the minimum needs approach that was discussed in connection with equivalence scales. As noted in section 14.3, this is the case with the US poverty line that was fixed in 1963 and has since been updated annually. As the package of minimum needs has not changed, the underlying concept is that of an absolute poverty measure. In the United Kingdom the poverty line has been taken as the level of income that is 120 or 140 percent of the minimum supplementary benefit level. As this level of benefit is determined by minimum needs, a minimum needs poverty line is implied. In addition benefits have risen with increases in average income, so causing the poverty line to rise. The UK poverty line thus represents the use of a relative concept of poverty.

The assumption that there is a precise switch between poverty and nonpoverty as the poverty line is crossed is very strong. It is much more natural for there to be a gradual move out of poverty as income increases. The precision of the poverty line may also lead to difficulty in determining where it should lie if the level of poverty is critically dependent on the precise choice. Both of these difficulties can be overcome by observing that often it is not the precise level of poverty that matters but changes in the level of poverty over time and across countries. In these instances the poverty value is not too important but only the rankings. This suggest the procedure of calculating poverty for a range of poverty lines. If poverty is higher today for all poverty lines than it was yesterday, then it seems unambiguous that poverty has risen. In this sense the poverty line may not actually be of critical importance for the uses to poverty measurement is often put. An application illustrating this argument is given below.

### 14.5.2 Poverty Measures

The poverty line is now taken as given, and we proceed to discuss alternative measures of poverty. The basic issue in this discussion is how best to combine two pieces of information (how many households are poor, and how poor they are) into a single quantitative measure of poverty. By describing a number of measures, the discussion will draw out the properties that are desirable for a poverty measure to possess.

Throughout the discussion the poverty line is denoted by the income level $z$ such that a household with an income level below or equal to $z$ is classed as living in poverty. For a household with income $M^h$ the *income gap* measures how far their income is below the poverty line. Denoting the income gap for household $h$ by $g_h$, it follows that $g_h = z - M^h$. Given the poverty line $z$ and an income distribution $\{M^1, \ldots, M^H\}$,
where $M^1 \leq M^2 \leq \ldots \leq M^H$, the number of households in poverty is denoted by $q$. The value of $q$ is defined by the facts that the income of household $q$ is on or below the poverty line, so $M^q \leq z$, but that of the next household is above $M^{q+1} > z$.

The simplest measure of poverty is the head-count ratio, which determines the extent of poverty by counting the number of households whose incomes are not above the poverty line. Expressing the number as a proportion of the population, the head-count ratio is defined by

$$E = \frac{q}{H}. \quad (14.28)$$

This measure of poverty was first used by Rowntree in 1901 and has been employed in many subsequent studies. The major advantage of the head-count ratio is its simplicity of calculation.

The head-count ratio is clearly limited because it is not affected by how far below the poverty line the households are. For example, with a poverty line of $z = 10$ the income distributions $\{1, 1, 20, 40, 50\}$ and $\{9, 9, 20, 40, 50\}$ would both have a headcount ratio of $E = \frac{2}{5}$. A policy maker may well see these income distributions differently, since the income required to alleviate poverty in the second case (2 units) is much less than that required for the first (18 units). The head-count ratio is also not affected by any transfer of income from a poor household to one that is richer if both households remain on the same side of the poverty line. Even worse, observe that if we change the second distribution to $\{7, 11, 20, 40, 50\}$ the head-count ratio falls to $E = \frac{1}{5}$, so a regressive transfer has actually reduced the head-count ratio. This will happen whenever a transfer takes the income of the recipient of the transfer above the poverty line.

Only one of the two pieces of information on poverty are used in the head count. A measure that uses only information on how far below the poverty line the incomes of the poor households are is the aggregate poverty gap. This is defined as the simple sum of the income gaps of the households that are in poverty. Recalling that it is the first $q$ households that are in poverty, the aggregate poverty gap is

$$V = \sum_{h=1}^{q} g_h. \quad (14.29)$$

The interpretation of this measure is that it is the additional income for the poor that is required to eliminate poverty. It provides some information but is limited by the fact that it is not sensitive to changes in the number in poverty. In addition the aggregate poverty gap gives equal weight to all income shortfalls regardless of how far they are from the
poverty line. It is therefore insensitive to transfers unless the transfer takes one of the households out of poverty. To see this latter point, for the poverty line of $z = 10$ the income distributions \{5, 5, 20, 40, 50\} and \{1, 9, 20, 40, 50\} have an aggregate poverty gap of $V = 10$. The distribution between the poor is somewhat different in the two cases.

One direct extension of the aggregate poverty gap is to adjust the measure by taking into account the number in poverty. The *income gap ratio* does this by calculating the aggregate poverty gap and then dividing by the number in poverty. Finally the value obtained is divided by the value of the poverty line, $z$, to obtain a measure whose value falls between 0 (the absence of poverty) and 1 (all households in poverty have no income):

$$I = \frac{1}{z} \frac{\sum_{q=1}^{Q} g_h}{\sum_{h=1}^{H} q}.$$  (14.30)

For the income distribution \{1, 9, 20, 40, 50\}, the income gap ratio when $z = 10$ is

$$I = \frac{1}{10} \frac{9 + 1}{2} = 0.5.$$  (14.31)

However, when this income distribution changes to \{1, 10, 20, 40, 50\}, so that only one household is now in poverty, the measure becomes

$$I = \frac{1}{10} \frac{9}{1} = 0.9.$$  (14.32)

This example reveals that the income gap ratio has the unfortunate property of being able to report increased poverty when the income of a household crosses the poverty line and the number in poverty is reduced.

These observations suggest that it is necessary to reflect more carefully on the properties that a poverty measure should possess. In 1976 Amartya K. Sen suggested that a poverty measure should have the following properties:

- Transfers of income between households above the poverty line should not affect the amount of poverty.
- If a household below the poverty line becomes worse off, poverty should increase.
- The poverty measure should be anonymous, meaning it should not depend on who is poor.
- A regressive transfer among the poor should raise poverty.
These are properties that have already been highlighted by the discussion. Two further properties were also proposed:

- The weight given to a household should depend on their ranking among the poor, meaning more weight should be given to those furthest below the poverty line.
- The measure should reduce to the headcount if all the poor have the same level of income.

One poverty measure that satisfies all of these conditions is the Sen measure

\[ S = E \left[ I + (1 - I) G_p \left( \frac{q}{q + 1} \right) \right], \]  

(14.33)

where \( G_p \) is the Gini measure of income inequality among the households below the poverty line. This poverty measure combines a measure of the number in poverty (the head-count ratio), a measure of the shortfall in income (the income gap ratio), and a measure of the distribution of income among the poor (the Gini). Applying the Sen measure to the income distribution \{1, 9, 20, 40, 50\}, when \( z = 10 \), we have \( E = \frac{2}{5} \) and \( I = 0.5 \). The Gini is calculated for the distribution of income of the poor \{1, 9\}, so \( G_p = 1 - \frac{1}{2 \times 3} [3 \times 1 + 9] = \frac{4}{10} \). These values give

\[ S = \frac{2}{5} \left[ 0.5 + (1 - 0.5) \frac{4}{10} \left( \frac{2}{2 + 1} \right) \right] = 0.2533. \]  

(14.34)

In contrast, for the distribution \{1, 10, 20, 40, 50\} that was judged worse using the income gap ratio, there is no inequality among the poor (since there is a single poor person), so the Sen measure is

\[ S = \frac{1}{5} \left[ 0.9 + (1 - 0.9) 0 \left( \frac{1}{1 + 1} \right) \right] = 0.18, \]  

(14.35)

which is simply the head-count ratio and records a lower level of poverty.

There is a further desirable property that leads into an alternative and important class of poverty measures. Consider a population that can be broken down into distinct subgroups. For instance, imagine dividing the population into rural and urban dwellers. The property we want is for the measure to be able to assign a poverty level for each of the groups and to aggregate these group poverty levels into a single level for the total society. Further we will also want the aggregate measure to increase if poverty rises in one of the subgroups and does not fall in any of the others. So, if rural poverty rises while urban poverty remains the same, aggregate poverty must rise. Any poverty measures that satisfies this condition are termed subgroup consistent.
Before introducing a form of measure that is subgroup consistent, it is worth providing additional discussion of the effect of transfers. The measures discussed so far have all had the property that the effect of a transfer has been independent of the income levels of the loser and gainer (except when the transfer was between households on different sides of the poverty line or changed the number in poverty). In the same way as in inequality measurement, where we argued for magnifying the effect of deviations far from the mean, we can argue that the effect of a transfer in poverty measurement should be dependent on the incomes of those involved in the transfer. For example, a transfer away from the lowest income household should have more effect on measured poverty than a transfer away from a household close to the poverty line. A poverty measure will satisfy this sensitivity to transfers if the increase in measured poverty caused by a transfer of income from a poor household to a poor household with a higher income is smaller, the larger is the income of the lowest income household.

Let the total population remain at $H$. Assume that this population can be divided into $\Gamma$ separate subgroups. Let $g_h^\gamma$ be the income gap of a poor member of subgroup $\gamma$ and $q_\gamma$ be the number of poor in that subgroup. Using this notation, a poverty measure that satisfies the property of subgroup consistency is the Foster–Greer–Thorbecke (FGT) class, given by

$$P_\alpha = \frac{1}{H} \sum_{\gamma=1}^{\Gamma} \left[ \frac{q_\gamma}{q_\gamma} \sum_{h=1}^{q_\gamma} \frac{g_h^\gamma}{z} \right]^\alpha. \quad (14.36)$$

The form of this measure depends on the value chosen for the parameter $\alpha$. If $\alpha = 0$, then

$$P_0 = \frac{\sum_{\gamma=1}^{\Gamma} q_\gamma}{H} = E, \quad (14.37)$$

the head-count ratio. If instead $\alpha = 1$, then

$$P_1 = \frac{1}{H} \sum_{\gamma=1}^{\Gamma} \left[ \frac{q_\gamma}{q_\gamma} \sum_{h=1}^{q_\gamma} \frac{g_h^\gamma}{z} \right] = EI, \quad (14.38)$$

the product of the head-count ratio and the income gap ratio. Note that $P_0$ is insensitive to transfers, while the effect of a transfer for $P_1$ is independent of the incomes of the households involved. For higher values of $\alpha$ the FGT measure satisfies sensitivity to transfers, and more weight is placed on the income gaps of lower income households.
14.5.3 Two Applications

The use of these poverty measures is now illustrated by reviewing two applications. The first application, taken from Foster, Greer, and Thorbecke (1984), shows how subgroup consistency can give additional insight into the sources of poverty. The second application is extracted from an OECD working paper and illustrates how a range of poverty lines can be used as a check on consistency. It also reveals that there can be a good degree of agreement between different measures of poverty.

Table 14.3 reports an application of the FGT measure. The data are from a household survey in Nairobi and groups the population according to their length of residence in Nairobi. The measure used is the $P_2$ measure, so $\alpha = 2$. As already discussed, the use of the FGT measure allows the contribution of each group to total poverty to be identified. For example, those living in Nairobi between 6 and 10 years have a level of poverty of 0.0343 and contribute 12.1 percent to total poverty—this is also the percentage by which total poverty would fall if this group were all raised above the poverty line. The division into groups also allows identification of where the major contribution to poverty arises. In this case the major contribution is made by those in the 21 to 70 group. Although the actual poverty level in this group is low, the number of households in this group causes them to have a major effect on poverty.

The second application is reported in table 14.4. This OECD analysis studies the change in poverty over (approximately) a ten-year period from the mid-1980s to the

<table>
<thead>
<tr>
<th>Years in Nairobi</th>
<th>Level of poverty</th>
<th>Contribution to total poverty (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.4267</td>
<td>5.6</td>
</tr>
<tr>
<td>0.01–1</td>
<td>0.1237</td>
<td>6.5</td>
</tr>
<tr>
<td>2</td>
<td>0.1264</td>
<td>6.6</td>
</tr>
<tr>
<td>3–5</td>
<td>0.0257</td>
<td>5.1</td>
</tr>
<tr>
<td>6–10</td>
<td>0.0343</td>
<td>12.1</td>
</tr>
<tr>
<td>11–15</td>
<td>0.0291</td>
<td>9.4</td>
</tr>
<tr>
<td>16–20</td>
<td>0.0260</td>
<td>6.6</td>
</tr>
<tr>
<td>21–70</td>
<td>0.0555</td>
<td>23.8</td>
</tr>
<tr>
<td>Permanent resident</td>
<td>0.1659</td>
<td>8.7</td>
</tr>
<tr>
<td>Don’t know</td>
<td>0.2461</td>
<td>15.5</td>
</tr>
<tr>
<td>Total</td>
<td>0.0558</td>
<td>99.9</td>
</tr>
</tbody>
</table>

mid-1990s. The numbers given are therefore the percentage change in the measure and not the value of the measure. What the results show is that the direction of change in poverty as measured by the head-count ratio is not sensitive to the choice of the poverty line—the only inconsistency is the value for Australia with the poverty line as 40 percent of median income. In detail, there has been a decrease in poverty in Australia, Belgium, and the United States but an increase in Germany, Japan, and Sweden. The results in the three central columns report the calculations for three different poverty measures. These show that the Sen measure and the head count are always in agreement about the direction of change. This is not true of the income gap, which disagrees with the other two for Australia and the United States.

### 14.6 Unequal Opportunities

The main focus of the literature on equality of opportunity is on separating sources of inequality of outcomes that are morally acceptable from those that are morally unacceptable. In the seminal book of Roemer, *Equality of Opportunity* (1998), it is shown that some inequality of outcomes is morally acceptable. Unequal outcomes that are a consequence of factors for which individuals are judged to be responsible—referred to as *effort*—are morally acceptable and should not be compensated for. Only inequality that is outside the realm of individual choice—referred to as *circumstances*—should be compensated for. Typical examples of circumstances that may affect individual outcomes are family background and individual attributes such as race, gender, and place of birth. To explore these ideas, it is first necessary to define equality of opportunity.
14.6.1 Defining Equality of Opportunity

When are opportunities equal? The central idea is to observe how the opportunity sets of people vary with their family background, race, gender, and so on. Equality of opportunity is achieved when no particular set of circumstances is preferred to another set of circumstances by all individuals. Given that people facing similar circumstances may produce different outcomes, the problem amounts to the comparison of distributions of outcomes, conditional on circumstances.

Consider a situation where individuals are allowed to choose their circumstances, $s$—which we refer to as their type—before they know their level of effort. Equality of opportunity prevails between circumstances $s$ and $s'$ if $s$ is not preferred to $s'$ by all individuals, and vice versa. In other words, people do not unanimously order the opportunity sets $s$ and $s'$. The opportunity set of an individual can be represented by the conditional distribution function $F(x|s)$ denoting the probability of producing an outcome less than or equal to $x$ for given type $s$. For example, this could describe the distribution of educational attainments for pupils of low-income families (say type $s$) compared to pupils from high-income families (say type $s'$).

Under the (weak) assumption that preferences satisfy the criteria of first-order stochastic dominance (FSD) and second-order stochastic dominance (SSD), stochastic dominance tests can be used to rank conditional distribution functions. Formal definitions of first-order stochastic dominance (FSD) and second-order stochastic dominance (SSD) are given in, respectively, (14.39) and (14.40). Suppose inequality of opportunity where circumstance $s$ is preferred to circumstance $s'$ by all individuals. Inequality of opportunity defined as first-order stochastic dominance between $s$ and $s'$ means that the distribution of outcome $x$ conditional on $s$ is for all $x$ below the distribution of $x$ conditional on circumstance $s'$:

$$s \preceq_{FSD} s' \Leftrightarrow F(x|s) \leq F(x|s'), \quad \forall x \in \mathbb{R}_+.$$

(14.39)

However, it can easily be shown that this is a very weak definition of equality of opportunity. Indeed suppose a situation where the outcome distribution of type $s$ always dominates the outcome distribution of type $s'$, except at the top (possibly, when they exert maximal effort). Under the definition of first-order stochastic dominance, equality of opportunity is not rejected in this case. But it is unfair because type $s'$ must exert maximal effort to get a chance to outperform type $s$.

Second-order stochastic dominance provides extra restrictions. Under second-order stochastic dominance, equality of opportunity prevails when the expected value derived from distribution $F(y|s)$ is not greater than the one derived from $F(y|s')$: 
Chapter 14: Inequality and Poverty

\[ s \leq SSD s' \text{ iff } \int_0^x F(y|s)dy \leq \int_0^x F(y|s')dy, \forall x \in \mathbb{R}_+. \] (14.40)

14.6.2 Measuring Equality of Opportunity

Econometric stochastic dominance techniques can be used to test FSD and SSD of conditional distributions. In relation to classical measure of inequality, we can also estimate inequality of opportunity with a Gini-type index. This index is based on the equivalence between SSD and Lorenz dominance. The Gini opportunity (GO) index with \( k \) types is defined as

\[ GO(x) = \frac{1}{\mu} \sum_{i=1}^{k} \sum_{j>i} p_i p_j (\mu_j (1-G_j) - \mu_i (1-G_i)), \] (14.41)

where \( \mu \) is the mean of the population, \( \mu_k \) the mean of group \( k \), \( p_k \) the population weight of group \( k \), and \( G \) the Gini coefficient. The GO index computes the sum of all pairwise differences of the opportunity sets of all types, where the opportunity sets are defined as twice the area under the generalized Lorenz curve, \( \mu_s (1-G_s) \), for type \( s \). The GO index is in the interval \([0, 1]\). A value of 0 indicates perfect equality of opportunity.

14.6.3 Equal-Opportunity Policy

How does the theory of equal opportunity translate into policy? The distinctive feature of the equal-opportunity policy, compared to classical welfare policy, is that it is a nonwelfarist policy. Welfarism is the view that only the set of vectors of outcome (welfare) possibilities matters for choosing public policy. To be precise, if we represent individual preferences over social alternatives by utility functions, then the choice of a social alternative should depend only on the information that is recoverable from the utility possibilities sets. In this sense, welfarism is a consequentialist view.

The equal-opportunity approach says that one cannot judge the goodness of a social outcome by knowing only the distribution of outcomes; one must also know how hard people tried in order to evaluate that goodness. To put it differently, one must know the role of effort in achievement to pass judgment on the goodness of a public policy. From this perspective income taxation may not be the instrument of choice to equalize opportunities for income (as recommended by the tax principle): one naturally
thinks of using educational finance policy as a method for compensating children from disadvantaged families.

To illustrate, there is a large difference between an equal-resource policy, which invests the same amount in all children, and an equal-opportunity policy, which invests in children so as to compensate for different social circumstances. The United States, with its system of locally financed public education, is in most places less equitable even than the equal resource policy would be: that is, usually more is invested in the public education of advantaged children than of disadvantaged children. Most of the affirmative action policies are grounded in an equal-opportunity approach.

14.7 Intergenerational Inequality

Social mobility occurs when different generations of a family have differing social status. Formally, social mobility refers to a lack of correlation between the educational attainments and earnings of parents and those of their children. To what extent is low educational attainment and consequent poverty transmitted by parents to their children? How can this be measured? And how can it be explained?

14.7.1 Measuring Issues

There are basically two sets of measures of the transmission of income and education across generations. The first set relates the earnings of parents to those of children, or the education of parents to that of children. Both the correlation and elasticities of these variables between parents and children give a basic measure of the intergenerational transmission of income and education. The difference between correlations and elasticities reflects differences in the variance of income/education for parents’ and children’s generations. The correlation is always between 0 and 1, but the elasticity can be greater than 1 if there is an increase in inequality across generations. The complement to the correlation (1 − correlation) is a measure of intergenerational mobility.

The central issues are to measure permanent income by averaging over a sufficient number of years (to accommodate temporary shocks and lifetime income progression) and to compare the income of the parents to that of children at the same point in their life cycle. As a matter of illustration, Nicoletti and Ermisch (2007) obtain income elasticities for the United States of around 0.5 to 0.6, and for the United Kingdom of around 0.3. Elasticities for the Nordic countries are always lower than 0.3. The comparison across countries and over time of these elasticities requires the use of
the same estimation methods, same definitions, and the same sample selection rule. There are also interesting nonlinearities in the elasticities. Bratsberg et al. (2007) find strong nonlinearities in the Nordic countries but not in the United States or the United Kingdom. For example, in Denmark the elasticities are 0.06 at the bottom of the income distribution and 0.31 at the top of the distribution. They suggest that this nonlinearity is related to the strong public education systems that exist in the Nordic countries.

The second set of measures use transition matrices to estimate mobility at each point in the distribution. Transition matrices enable us to compare the mobility across the full distribution rather than just around the mean. Jantti et al. (2006) split the distribution into quintiles and study mobility across quintiles. They find that more than 40 percent of sons in the United States who are born to fathers in the lowest quintile are in the lowest quintile themselves. Mobility from the lowest quintile is found to be much higher in Nordic countries. They find that much of the difference in mobility rates between the United States and the Nordic countries are in fact attributable to difference in the tails of the income distributions. Mobility across the middle three quintiles is very similar across all countries.

Some authors define upward mobility as the probability that a child’s percentile rank in the distribution of children exceeds the father’s percentile rank in the distribution of fathers. In this case more weight is placed on small moves in position compared to the quintile or quartiles approach. It has been shown that this distinction matters as the degree of upward mobility of blacks in the United States is similar to that of whites when the finer mobility metric is used.

14.7.2 Causal Mechanisms

Understanding the determinants of intergenerational correlations is crucial for the development of appropriate public policy. Without uncovering the driving forces behind intergenerational transmission, it is impossible to figure out how to promote changes. This is, of course, a very difficult task, as it is often the case that any particular parental attribute (e.g., as education or earnings) is correlated with a variety of parental nonobservable characteristics.

Nature and Nurture

Strong earnings and educational correlations among siblings reveal the importance of shared genetic and environmental factors, but they are not very helpful for pinning down causal mechanisms. Bjorklund et al. (2005) make use of correlations across
identical twins, fraternal twins, full siblings, half siblings, and adopted siblings, when they are both raised together to distinguish between nature and nurture. They show that genetics are more important than shared environment. Surprisingly, the identical twin correlation is only 0.36.

We must interpret this decomposition between nature and nurture with great prudence because it is possible that twins are treated differently by parents and, more important, because with assortative mating people tend to marry people with similar characteristics and so likely share genetic characteristics. It is just as likely there is positive interaction between environmental and genetic factors that complicates the separation of the effects of the two. Also twins are hardly representative of the population.

Using a sample of adopted children in Wisconsin, Plug (2004) finds a regression coefficient of about 0.28 on adoptive mother’s education. However, the interpretation of this is not clear because adoption is not random and may result in positive selection whereby high-ability adoptees end up in more educated families. Overall, this literature on twins’, siblings’, and adoptees’ correlations suggest that both environmental and genetic factors are important. Even so, we cannot be sure which factors matter most. Taken as a whole, the findings in the literature are very inconsistent. Some studies find the effect of fathers to be more important, other studies find the effects of mothers to be more important.

School System

It has been shown that the school system matters. A standard argument is based on credit constraints. Solon (2004) presents a basic model in which families are credit constrained and must reduce current consumption to invest in human capital. If there are no credit constraints, and thus parents can borrow from their children’s future earnings, each family will optimally invest in the human capital of their children (assuming perfect altruism). The optimality conditions will recommend investing more in high-ability children, and so we should expect intergenerational transmission of inequality only if child ability and parental earnings are correlated. If there are credit constraints, however, poor families cannot invest optimally in their children’s human capital. As a result a higher income level means an increase human capital investment.

The model also predicts that the intergenerational elasticity is increasing in the genetic transmission of ability and the rate of return on human capital investment, but it is decreasing with the public investment in education. The implication is that the low elasticities in the Nordic countries could be explained either by low returns on training investment (the compressed earnings distributions) or the public education system that
Chapter 14: Inequality and Poverty

tends to equalize educational opportunities for children. Suggestive evidence is con-
tained in Ichino et al. (2009) who report a negative correlation of $-0.54$ between public
expenditure on education and intergenerational income elasticities. The correlation is
even stronger with public expenditure on primary education.

**Attitudes and Social Behaviors**

What do parents transmit to their children? Other social scientists have claimed that
attitudes matter more than the socioeconomic status of the family. This is particularly
relevant for low-income groups and minorities. Indeed Cunha and Heckman (2008)
survey a large number of studies that show that nonpecuniary factors (psychic costs,
motivations, etc.) play a major role in explaining why minorities and persons from
low-income families do not attend college even though it is financially profitable to do
so. Because returns to schooling are lower for people less likely to attend college, these
groups may “rationally” choose to underinvest in education. Obviously economists
have much to learn from other social scientists to better grasp the nonpecuniary factors
that shape future economic and educational attainments.

**Segregation**

Segregation is important for understanding the dynamics of group inequality. Members
of different groups with identical distributions of cognitive abilities will invest dif-
ferently in the education of their children when the segregation is sufficiently great.
Furthermore the relationship between group equality and social segregation can exhibit
a discontinuity if there exists a critical level of segregation such that convergence to
group equality occurs if and only if segregation lies below this threshold. Hence a small
increase in social integration, if it takes the economy across the threshold, may have
large effects on long-run group inequality, while a large increase in integration that
does not cross the threshold may have no persistent effect. This suggests that economic
inequality across social groups might arise endogenously under certain conditions,
without preexisting discrimination or group differences in ability or wealth.

Dynamic systems with positive feedback can exhibit multiple types of self-reinforced
behavior at the group level. Positive feedback means that while there is tendency for
members of a group to make similar decisions, such feedback does not say which is
generally made. The implication of positive feedback is that different groups with iden-
tical characteristics can eventually adopt different norms of behaviors. This is certainly
a fruitful direction of investigation but a suitable method of empirical testing is not
obvious. Indeed there is no agreement in the literature on how to measure segregation.
(with nearly 20 different indexes of segregation representing different concepts such as evenness, exposure, concentration, centralization, and clustering), and as a result it is hardly possible to verify whether segregation is a leading cause of racial and social differences on economic outcomes.

14.8 Conclusions

The need to quantify is driven by the aim of making precise comparisons. What economic analysis contributes is an understanding of the bridge between intuitive concepts of inequality and poverty, and specific measures of these phenomena. Analysis can reveal the implications of alternative measures and provide principles that a good measure should satisfy.

The first problem we challenged in this chapter was the comparison of incomes between households of different compositions. It is clearly more expensive to support a large family than a small family, but exactly how much more expensive is more difficult to determine. Equivalence scales were introduced as the analytical tool to solve this problem. These scales were initially based on the cost of achieving a minimum standard of living. Although simple, such an approach does not easily generalize to higher income levels, nor does it take much account of economic optimization. In principle, equivalence scales could be built directly from utility functions, but to do so, issues must be addressed of how the preferences of the individual members of a household are aggregated into a household preference order.

Inequality occurs when some households have a higher income (after the incomes have been equivalized for household composition) than others. The Lorenz curve provides a graphical device for contrasting income distributions. Some income distributions can be ranked directly by the Lorenz curve, in which case there is no ambiguity about which has more inequality, but not all distributions can be. Inequality measures provide a quantitative assessment of inequality by imposing restrictions beyond those incorporated in the Lorenz curve. The chapter investigated the properties of a number of measures of inequality. Of particular importance was the observation that all inequality measures embody implicit welfare judgments. Given this, the Atkinson measure is constructed on the basis that the welfare judgments should be made explicit and the inequality measure constructed on these judgments. In principle, alternative measures can generate different rankings of income distributions, but in practice, as the application showed, they can yield very consistent rankings.
In many ways the measurement of poverty raises similar issues to those of inequality. The additional feature of poverty is the necessity to determine whether households can be classed as living in poverty. The poverty line, which provides the division between the two groups, plays a central role in poverty measurement. Where and how to locate this poverty line is important, but more fundamental is how it should be adjusted over time. At stake here is the key question of whether poverty should be viewed in absolute or relative terms. The practice in developed countries is to use relative poverty. The chapter reviewed a number of poverty measures from the head-count ratio to the Foster–Greer–Thorbecke measure. These measures are also distinguished by a range of sensitivity properties. The applications showed how they could be used and that the different measures could provide a consistent picture of the development of poverty despite their different conceptual bases.

The chapter has revealed how economic analysis is able to provide insights into what we are assuming when we employ a particular inequality or poverty measure. It has also revealed how we can think about the process of improving our measures. Inequality and poverty are significant issues, and better measurement is a necessary starting point for better policy.

Further Reading

The relationship between inequality measures and social welfare was first explored in:

A comprehensive survey of the measurement of inequality is given by:

A textbook treatment is in:

Issues surrounding the definition and implications of the poverty line are treated in:

The derivation of the Sen measure, and a general discussion of constructing measures from a set of axioms is given by:

The FGT measure was first discussed in:
Part V: Equity and Distribution


An in-depth survey of poverty measures is:


The seminal contribution on the equality of opportunity is


Social mobility issues including the references presented in the chapter are treated in:


The issue of nature and nurture is discussed by:


The link between social inequality and the school system:


Exercises

**14.1** In many countries lottery prizes are not taxed. Is this consistent with Hicks’s definition of income?

**14.2** Let the utility function be $U = d^{1/2} \log(M)$, where $d$ is family size. Construct the equivalence scale for the value of $U = 10$. How is the scale changed if $U = 20$?
14.3 What economies of scale are there in family size? Are these greater or smaller at low incomes?

14.4 Take the utility function \( U = \log \left( \frac{x_1}{d} \right) + \log \left( \frac{x_2}{d} \right) \), where \( d \) is family size and good 1 is food.
   a. What proportion of income is spent on food? Can this provide the basis for an equivalence scale? Calculate the exact equivalence scale. Does it depend on \( U \)?
   b. Repeat part a for the utility function \( U = \left[ \frac{x_1}{d} \right]^{1/2} + \left[ \frac{x_2}{d} \right]^{1/2} \).

14.5 If children provide utility for their parents, show on a diagram how an equivalence scale can decrease as family size increases.

14.6 Consider a community with ten persons.
   a. Plot the Lorenz curve for the income distribution \((2, 4, 6, 8, 10, 12, 14, 16, 18, 20)\).
   b. Consider an income redistribution that takes two units of income from each of the four richest consumers and gives two units to each of the four poorest. Plot the Lorenz curve again to demonstrate that inequality has decreased.
   c. Show that the Lorenz curve for the income distribution \((2, 3, 5, 9, 11, 12, 15, 17, 19, 20)\) crosses the Lorenz curve for the distribution in part a.
   d. Show that the two social welfare functions \( W = \sum M^h \) and \( W = \sum \log(M^h) \) rank the income distributions in parts a and c differently.

14.7 What is the Gini index, and how can it be used to determine the impact of taxes and transfers on income inequality?

14.8 Calculate the Gini index for the income distributions used in parts a through c of exercise 14.6. Discuss the values obtained.

14.9 For a utilitarian social welfare function construct \( M_{EDE} \) for the distributions used in exercise 14.6 if the utility of income is logarithmic. Find the Atkinson inequality measure. Repeat the exercise for the Rawlsian social welfare function. Compare and discuss.

14.10 What drawbacks are there to eliminating inequality?

14.11 Should we be concerned with inequality if it is due to differences in ability? What if it is due to differences in effort levels?

14.12 Define inequality aversion. Explain how it is related to the concept of risk aversion.

14.13 Discuss the following quote from Cowell (1995, p. 23): “The main disadvantage of G[ini] is that an income transfer from a rich to a poorer man has a much greater effect on G if the men are near the middle rather than at either end of the parade.” Do you agree? Why or why not? (Hint: Use the formula for the Gini coefficient to determine the effect of a fixed transfer at different points in the income distribution.) Does the Gini have other “disadvantages”?

14.14 Consider a hypothetical island with only ten people. Eight have income of $10,000, one has income of $50,000, and one has income of $100,000.
   a. Draw the Lorenz curve for this income distribution. What is the approximate value of the Gini coefficient?
b. Suppose that a wealthy newcomer arrives on this island with an income of $500,000. How does it change the Lorenz curve? What is the impact on the Gini coefficient?

14.15 Have a look at actual income distribution in the United States available on the website <http://www.census.gov/hhes/income/histinc/histinctb.html>. Select Households and then Table H-2.

a. Plot the Lorenz curve for 1981 and 2001. Clearly label each curve. What can you say about the evolution of inequality over time?

b. Based on your diagram, can we conclude that the Gini coefficient was higher in 1981 or 2001? Explain.

c. Can we conclude from the diagram that the poor were necessarily worse off in either 1981 or 2001? Why or why not? Use Table H-1 on the website to refine your answer.

d. Now suppose that people with similar incomes are more likely to get married than people with dissimilar incomes. How would this change affect the Lorenz curve drawn in part a?

14.16 There are two senior advisors to the government, A and B, both of whom agree that the poverty line is at $4,000 for a single person. However, they have different equivalence scales. Mr. A believes that the scale factor in determining equivalent income should be 0.25 for each additional family member. Mrs. B suggests that the scale factor should be 0.75.

a. Find the poverty line for families of two, three, and four under both values of the scale factors 0.25 and 0.75.

b. Explain how Mr. A and Mrs. B must have very different views about income sharing within a family to end up with such different answers.

c. Suppose that the government is committed to providing welfare eligibility to every family below the poverty line. If this government wishes to keep total spending to a minimum, which of the two views should it support?

14.17 Given the income distributions

(1, 2, 2, 5, 5, 7, 11, 11, 12, 20, 21, 22, 24),
(2, 3, 3, 4, 4, 5, 7, 7, 11, 11, 12, 20, 21, 24),

and a poverty line of $6, calculate the Sen poverty measure. Explain the values obtained for the two distributions.

14.18 Use the two income distributions in exercise 14.17 to evaluate the Foster–Greer–Thorbecke poverty measure for $\alpha = 2$. Pool the distributions to evaluate the poverty measure for the total population. Show that the measure is a weighted sum of the measures for the individual distributions.

14.19 (Decoster) The Pareto distribution is a popular functional form for describing income distributions. It is a two-parameter specification for which the frequency density function reads as follows: $f(x) = \alpha x_0^\alpha x^{-(1+\alpha)}$ for $x \geq x_0$, where $x_0 > 0$ is the lowest income level and $\alpha > 1$ is a parameter.

a. Show that the mean income for the Pareto distribution is $x = \frac{\alpha x_0}{\alpha - 1}$.

b. Show that the distribution function for the Pareto distribution is $F(x) = 1 - \left[\frac{x_0}{x}\right]^\alpha$ for $x \geq x_0$. Discuss the effect of changing the parameter $\alpha$. 
501 Chapter 14: Inequality and Poverty

c. The Pareto distribution parameterized by $\alpha$ can easily be used to construct a very simple inequality measure, which is defined as follows: Take an arbitrary income level, say $x$. Calculate the mean income of the subpopulation of all income earners who have an income larger than $x$. The ratio of this mean income to the income $x$ is given by $I = \frac{\alpha}{x^{\alpha-1}}$. Calculate the values of this inequality index for some different values of $\alpha$ (e.g., $\alpha = 1.5, 2, 3$). Does $\alpha$ represent equality or inequality? What is the limiting value of $I$ for very large $\alpha$? Interpret this result.

d. Show that the Lorenz curve for the Pareto distribution is $L(p) = 1 - [1 - p]^{\alpha - 1}/\alpha$, where $p = F(x)$ and $p \in [0, 1]$. What is the shape of the curve for very large $\alpha$?

e. Draw the Lorenz curve for two values $\alpha_1 > \alpha_2$, and verify that the two Lorenz curves will never cross.

f. Show that the Gini coefficient for the Pareto distribution (with parameter $\alpha$) is $G = \frac{1}{2\alpha - 1}$. How does it compare with your answer in part e?

14.20 How is it possible for a government that pursues a poverty reduction target to increase inequality? Does this possibility invalidate poverty comparisons?

14.21 How are intergenerational persistence and mobility measured?

14.22 (Solon 2004) Consider the following model of intergenerational transmission. Define $Y_{it}$ as the earnings of generation $t$ in family $i$, and $I_{it-1}$ investment by generation $t - 1$ in the human capital of generation $t$. Human capital accumulation is given by $H_{it} = \theta \log(I_{it-1}) + e_{it}$, where $\theta$ is a (uniform) productivity parameter, and $e_{it}$ denotes individual ability and follows an AR(1) process: $e_{it} = \delta + \lambda e_{it-1} + v_{it}$, where $\lambda$ denotes genetic transmission of ability. Human capital translates into earnings on the labor market as follows: $\log(Y_{it}) = \mu + \rho H_{it}$, where $\rho$ denotes the returns to human capital. Family $i$ have Cobb–Douglas preferences over own consumption, $C_{it-1}$, and child’s income, $Y_{it}$, $U(C, Y) = C_{it-1}^{1-\alpha}Y_{it}^\alpha$.

The budget constraint is $C_{it-1} = Y_{it-1} - I_{it-1}$.

a. What is the optimal investment in human capital?

b. What is the resulting equation of intergenerational earnings transmission?

c. What is the resulting intergenerational elasticity?

d. Discuss the solutions.

14.23 To test the predictions of the Solon model described in the exercise above, use the cross-sectional PISA data on education test scores for pupils at 15 years old in OECD countries. (Available from OECD at www.pisa.oecd.org). Pick a country of your choice and assemble
for each pupil of the country of your choice their test score and their ESCS index (i.e., the
economic and sociocultural index of their parent).
a. Plot the pupils test scores and their ESCS index. Compute the regression line.
b. Regroup pupils by ESCS quartile with the first quartile representing the 25 percent of pupils with the lowest ESCS and the fourth quartile representing the 25 percent of pupils with the highest ESCS index. Draw the cumulative distribution function of test scores for each ESCS quartile and check whether the first-order stochastic dominance condition is satisfied.
15 Commodity Taxation

15.1 Introduction

Commodity taxes are levied on transactions involving the purchase of goods. The necessity for keeping accounts ensures that such transactions are generally public information. This makes them a good target for taxation. The drawback, however, is that commodity taxation distorts consumer choices and causes inefficiency. Some striking historical examples can be found in the United Kingdom where there have been window taxes and hearth taxes. The window tax was introduced in 1696 in the reign of William III and lasted until 1851. The tax was paid on any house with more than six windows (increased to eight in 1825), which gave an incentive to brick up any windows in excess of the allowable six. Even today, old houses can be found with windows still bricked up. The hearth tax was levied between 1662 and 1689 at the rate of two shillings (two days’ wages for a ploughman) per annum on each hearth in a building. This induced people to brick up their chimneys and shiver through the winter. In the marketplace, commodity taxes drive a wedge between the price producers receive and the price consumers pay. This leads to inefficiency and reduces the attainable level of welfare compared to what could be achieved using lump-sum taxes. This is the price that has to be paid for implementable taxation.

The effects of commodity taxes are quite easily understood—the imposition of a tax raises the price of a good. On the consumer side of the market, the standard analysis of income and substitution effects predicts what will happen to demand. For producers, the tax is a cost increase, and they respond accordingly. What is more interesting is the choice of the best set of taxes for the government. There are several interesting settings for this question. The simplest version can be described as follows. There is a given level of government revenue to be raised that must be financed solely by taxes on commodities. How must the taxes be set so as to minimize the cost to society of raising the required revenue? This is the Ramsey problem of efficient taxation, first addressed in the 1920s. The insights its study gives are still at the heart of the understanding of setting optimal commodity taxes. More general problems introduce equity issues in addition to those of efficiency.

The chapter begins by discussing the deadweight loss that is caused by the introduction of a commodity tax. A diagrammatic analysis of optimal commodity taxation is then presented. This diagram is also used to demonstrate the Diamond–Mirrlees Production Efficiency result. Following this, the Ramsey rule is derived and an interpretation of
this is provided. The extension to many consumers is then made and the resolution of the equity–efficiency trade-off is emphasized. This is followed by a review of some numerical calculations of optimal taxes based on empirical data.

15.2 Deadweight Loss

Lump-sum taxation was described as the perfect tax instrument because it does not cause any distortions. The absence of distortions is due to the fact that a lump-sum tax is defined by the condition that no change in behavior can affect the level of the tax. Commodity taxation does not satisfy this definition. It is always possible to change a consumption plan if commodity taxation is introduced. Demand can shift from goods subject to high taxes to goods with low taxes, and total consumption can be reduced by earning less or saving more. It is these changes at the margin, which we call substitution effects, that are the tax-induced distortions.

The introduction of a commodity tax raises tax revenue but causes consumer welfare to be reduced. The deadweight loss of the tax is the extent to which the reduction in welfare exceeds the revenue raised. This concept is illustrated in figure 15.1. Before the tax is introduced, the price of the good is $p$ and the quantity consumed is $X^0$. At this price the level of consumer surplus is given by the triangle $abc$. A specific tax of amount $t$ is then levied on the good, so the price rises to $q = p + t$ and quantity consumed falls to $X^1$. This fall in consumption together with the price increase

![Figure 15.1](image_url)

Deadweight loss
reduces consumer surplus to \(aef\). The tax raises revenue equal to \(tX^1\), which is given by the area \(edef\). The part of the original consumer surplus that is not turned into tax revenue is the deadweight loss, \(DWL\), given by the triangle \(bde\).

It is possible to provide a simple expression that approximates the deadweight loss. The triangle \(bde\) is equal to \(\frac{1}{2}tdX\), where \(dX\) is the change in demand \(X^0 - X^1\). This formula could be used directly, but it is unusual to have knowledge of the level of demand before and after the tax is imposed. Accepting this, it is possible to provide an alternative form for the formula. This can be done by noting that the elasticity of demand is defined by \(\varepsilon_d = \frac{dX}{Xdp}\), so it implies that \(dX = \varepsilon_d \frac{X^0}{p} dp\). Substituting this into deadweight loss gives

\[
DWL = \frac{1}{2} \left| \varepsilon_d \right| \frac{X^0}{p} t^2,
\]

since the change in price is \(dp = t\). The measure in (15.1) is approximate because it assumes that the elasticity is constant over the full change in price from \(p\) to \(q = p + t\).

The formula for deadweight loss reveals two important observations. First, deadweight loss is proportional to the square of the tax rate. The deadweight loss will therefore rise rapidly as the tax rate is increased. Second, the deadweight loss is proportional to the elasticity of demand. For a given tax change the deadweight loss will be larger the more elastic is demand for the commodity.

An alternative perspective on commodity taxation is provided in figure 15.2. Point \(a\) is the initial position in the absence of taxation. Now consider the contrast between a
lump-sum tax and a commodity tax on good 1 when the two tax instruments raise the same level of revenue. In the figure the lump-sum tax is represented by the move from point \( a \) to point \( b \). The budget constraint shifts inward, but its gradient does not change. Utility falls from \( U_0 \) to \( U_1 \). A commodity tax on good 1 increases the price of good 1 relative to the price of good 2 and causes the budget constraint to become steeper. At point \( c \) the commodity tax raises the same level of revenue as the lump-sum tax. This is because the value of consumption at \( c \) is the same as that at \( b \), so the same amount must have been taken off the consumer by the government in both cases. The commodity tax causes utility to fall to \( U_2 \), which is less than \( U_1 \). The difference \( U_1 - U_2 \) is the deadweight loss measured directly in utility terms.

Figure 15.2 illustrates two further points to which it is worth drawing attention. Notice that commodity taxation produces the same utility level as a lump-sum tax that would move the consumer to point \( d \). This is clearly a larger lump-sum tax than that which achieved point \( a \). The difference in the size of the two lump-sum taxes provides a monetary measure of the deadweight loss. The effect of the commodity tax can now be broken down into two separate components. First, there is the move from the original point \( a \) to point \( d \). In line with the standard terminology of consumer theory, this is called an *income effect*. Second, there is a *substitution effect* due to the increase in the price of good 1 relative to good 2 represented by a move around an indifference curve. This shifts the consumer’s choice from point \( d \) to point \( c \).

This argument can be extended to show that it is the substitution effect that is responsible for the deadweight loss. To do this, note that if the consumer’s indifference curves
are all L-shaped so that the two commodities are perfect complements, then there is no substitution effect in demand—a relative price change with utility held constant just pivots the budget constraint around the corner of the indifference curve. As shown in figure 15.3, the lump-sum tax and the commodity tax result in exactly the same outcome, so the deadweight loss of the commodity tax is zero. The initial position without taxation is at \( a \) and both tax instruments lead to the final equilibrium at \( b \). Hence the deadweight loss is caused by substitution between commodities.

15.3 Optimal Taxation

The purpose of optimal tax analysis is to find the set of taxes that gives the highest level of welfare while raising the revenue required by the government. The set of taxes that do this are termed \textit{optimal}. In determining these taxes, consumers must be left free to choose their most preferred consumption plans at the resulting prices and firms to continue to maximize profits. The taxes must also lead to prices that equate supply to demand. This section will consider the problem for the case of a single consumer. This restriction ensures that only efficiency considerations arise. The more complex problem involving equity, as well as efficiency, will be addressed in section 15.6.

To introduce a number of important aspects of commodity taxation in a simple way, it is best to begin with a diagrammatic approach. Among the features that this makes clear are the second-best nature of commodity taxes relative to lump-sum taxes. In other words, the use of commodity taxes leads to a lower level of welfare compared to the optimal set of lump-sum taxes. Despite this effect, the observability of transactions makes commodity taxes feasible, whereas optimal lump-sum taxes are generally not, for the reasons explored in chapter 13.

Consider a two-good economy with a single consumer and a single firm (the Robinson Crusoe economy of chapter 2). One of the goods, labor, is used as an input (so it is supplied by the consumer to the firm), and the output is sold by the firm to the consumer. In figure 15.4 the horizontal axis measures labor use and the vertical axis output. The firm’s production set, marked \( Y \) in the figure, is also the production set for the economy. This is displaced from the origin by a distance \( R \) that equals the tax revenue requirement of the government. The interpretation is that the government takes out of the economy \( R \) units of labor for its own purposes. After the revenue requirement has been met, the economy then has constant returns to scale in turning labor into output. The commodity taxes have to be chosen to attain this level of revenue. Normalizing the wage rate to 1, the only output price for the firm that leads to zero profit is shown by \( p \). This is the only level of profit consistent with the assumption of competitive behavior, and \( p \) must
be the equilibrium price for the firm. Given this price, the firm is indifferent to where it produces on the frontier of its production set.

Figure 15.5 shows the budget constraint and the preferences of the consumer. With the wage rate of 1, the budget constraint for the consumer is constructed by setting the consumer’s price for the output to $q$. The difference between $q$ and $p$ is the tax on the consumption good. It should be noticed that labor is not taxed. As will become clear, this is not a restriction on the set of possible taxes. With these prices the consumer’s budget constraint can be written $qx = \ell$, where $x$ denotes units of the output and $\ell$ units of labor. The important properties of this budget constraint are that it is upward sloping and must pass through the origin. The preferences of the consumer are represented by indifference curves. The form of these follows from noting that the supply of labor causes the consumer disutility, so an increase in labor supply must be compensated for by further consumption of output in order to keep utility constant. The indifference curves are therefore downward sloping. Given these preferences, the optimal choice is found by the tangency of the budget constraint and the highest attainable indifference curve. Varying the price, $q$, faced by the consumer gives a series of budget constraints whose slopes increase as $q$ falls. Forming the locus of optimal choices determined by these budget constraints traces out the consumer’s offer curve. Each point on this offer curve can be associated with a budget constraint that runs through the origin and
an indifference curve tangential to that budget constraint. The interpretation given to the offer curve is that the points on the curve are the only ones consistent with utility maximization by the consumer in the absence of lump-sum taxation. It should also be noted that the consumer’s utility rises as the move is made up the offer curve.

Figures 15.4 and 15.5 can be superimposed to represent the production and consumption decisions simultaneously. This is done in figure 15.6, which can be used to find the optimal tax rate on the consumption good. The only points that are consistent with choice by the consumer are those on the offer curve. The maximal level of utility achievable on the offer curve is at the point where it intersects the production frontier. Any level higher than this is not feasible. This optimum is denoted by point $e$, and here the consumer is on indifference curve $I_0$. At this optimum the difference between the consumer price and the producer price for the output, $t^* = q - p$, is the optimal tax rate. That is, it is the tax that ensures that the consumer chooses point $e$. By construction, this tax rate must also ensure that the government raises its required revenue so that $t^*x^* = R$, where $x^*$ is the level of consumption at point $e$.

This discussion has shown how the optimal commodity tax is determined at the highest point of the offer curve in the production set. This is the solution to the problem of finding the optimal commodity taxes for this economy. The diagram also shows why labor can remained untaxed without affecting the outcome. The choices of the
consumer and the firm are determined by the ratio of prices they face or the direction of the price vector (which is orthogonal to the budget constraint). By changing the length (but not the direction) of either $p$ or $q$, one can introduce a tax on labor, but it does not alter the fact that $e$ is the optimum. This reasoning can be expressed by saying that the zero tax on labor is a normalization, not a real restriction on the system.

Figure 15.6 also illustrates the second-best nature of commodity taxation relative to lump-sum taxation. It can be seen that there are points above the indifference curve $I_0$ (the best achievable by commodity taxation) that are preferred to $e$ and that are also productively feasible. The highest attainable indifference curve for the consumer given the production set is $I_1$ with utility maximized at point $e^*$. This point would be chosen by the consumer if they faced a budget constraint that is coincident with the production frontier. A budget constraint of this form would cross the horizontal axis to the left of the origin and would have equation $qx = \ell - R$, where $R$ represents a lump-sum tax equal to the revenue requirement. This lump-sum tax would decentralize the first-best outcome at $e^*$. Commodity taxation can only achieve the second-best at $e$.

![Figure 15.6](image-url)  
Optimal commodity taxation
The diagrammatic illustration of optimal taxation in the one-consumer economy also shows another important result. This result, known as the *Diamond–Mirrlees Production Efficiency Lemma*, states that the optimal commodity tax system should not disrupt production efficiency. In other words, the optimum with commodity taxation must be on the boundary of the production set and all distortions are focused on consumer choice. This section provides a demonstration of the efficiency lemma and discusses its implications.

Production efficiency occurs when an economy is maximizing the output attainable from its given set of resources. This can only happen when the economy is on the boundary of its production possibility set. Starting at a boundary point, no reallocation of inputs among firms can increase the output of one good without reducing that of another (compare this with the conditions for Pareto-efficiency in chapter 2). In the special case where each firm employs some of all the available inputs, a necessary condition for production efficiency is that the marginal rate of substitution (\(MRS\)) between any two inputs be the same for all firms. Such a position of equality is attained, in the absence of taxation, by the profit maximization of firms in competitive markets. Each firm sets the marginal rate of substitution equal to the ratio of factor prices, and since factor prices are the same for all firms, this induces the necessary equality in the \(MRS\)s. The same is true when there is taxation, provided that all firms face the same after-tax prices for inputs, meaning inputs taxes are not differentiated among firms.

To see that the optimum with commodity taxation must be on the frontier of the production set, consider the interior point \(f\) in figure 15.7. If the equilibrium were at \(f\), the consumer's utility could be raised by reducing the use of the input while keeping output constant. Since this is feasible, \(f\) cannot be an optimum. Since this reasoning can be applied to any point that is interior to the production set, the optimum must be on the boundary.

Although figure 15.7 was motivated by considering the input to be labor, a slight re-interpretation can introduce intermediate goods. Assume that there are several industries and that each industry has a production process that uses one unit of labor to produce one unit of an intermediate good. The intermediate goods are then combined by the final goods industry using a production process that has constant returns to scale. The intermediate good production is displayed indirectly in figure 15.7 as the link between the use of labor and the output of the final good. The production efficiency argument
then follows directly as before and now implies that intermediate goods should not be
taxed, since this would violate the equalization of $MRS$s between firms.

The logic of the single-consumer economy can be adapted to show that the efficiency
lemma still holds when there are many consumers. What makes the result so obvious in
the single-consumer case is that a reduction in labor use or an increase in output raises
the consumer’s utility. With many consumers, such a change would have a similar effect
if all consumers supply labor or prefer to have more, rather than less, of the consumption
good. This will hold if there is some agreement in the tastes of the consumers. If this is
so, a direction of movement can be found from an interior point in the production set
to an exterior point that is unanimously welcomed. The optimum must then be on the
boundary.

In summary, the Diamond–Mirrlees Production Efficiency Lemma provides a per-
suasive argument for the nontaxation of intermediate goods and the nondifferentiation
of input taxes among firms. These are results of immediate practical importance, since
they provide a basic property that an optimal tax system must possess. As will become
clear, it is rather hard to make precise statements about the optimal levels of tax, but
what the efficiency lemma provides is a clear and simple statement about the structure
of taxation.
Chapter 15: Commodity Taxation

15.5 Tax Rules

The diagrammatic analysis has shown the general principle behind the determination of the optimal taxes. What is not shown is how the tax burden is allocated across different commodities. The optimal tax problem is to set the taxes on commodities to maximize social welfare subject to raising a required level of revenue. This section looks at tax rules that characterize the solution to this problem.

To derive the rules, it is first necessary to precisely specify a model of the economy. Let there be \( n \) goods, each produced with constant returns to scale by competitive firms. Since the firms are competitive, the price of the commodity they sell must be equal to the marginal cost of production. Under the assumption of constant returns, this marginal cost is also independent of the scale of production. Labor is assumed to be the only input into production.

With the wage rate as numéraire, these assumptions imply that the producer (or before-tax) price of good \( i \) is determined by

\[
p_i = c_i, \quad i = 1, \ldots, n,
\]

(15.2)

where \( c_i \) denotes the number of units of labor required to produce good \( i \). The consumer (or after-tax) prices are equal to the before-tax prices plus the taxes. For good \( i \) the consumer price \( q_i \) is

\[
q_i = p_i + t_i, \quad i = 1, \ldots, n.
\]

(15.3)

Writing \( x_i \) for the consumption level of good \( i \), the tax rates on the \( n \) consumption goods must be chosen to raise the required revenue. With the revenue requirement denoted by \( R \), the revenue constraint can be written as

\[
R = \sum_{i=1}^{n} t_i x_i.
\]

(15.4)

In line with this numbering convention, labor is denoted as good 0, so \( x_0 \) is the supply of labor (labor is the untaxed good, so \( t_0 = 0 \)).

This completes the description of the economy. The simplifying feature is that the assumption of constant returns to scale fixes the producer prices via (15.2) so that equilibrium prices are independent of the level of demand. Furthermore constant returns also implies that whatever demand is forthcoming at these prices will be met by the firms. If the budget constraints are satisfied (both government and consumer), any demand will be backed by sufficient labor supply to carry out the necessary production.
The Inverse Elasticity Rule

Figure 15.6 shows some of the features that the optimal set of commodity taxes will have. What the single-good formulation cannot do is give any insight into how that tax burden should be spread across different goods. For example, should all goods have the same rate of tax or should taxes be related to the characteristics of the goods? The first tax rule considers a simplified situation that delivers a very precise answer to this question. This answer, the inverse elasticity rule, provides a foundation for proceeding to the more general case. The simplifying assumption is that the goods are independent in demand so that there are no cross-price effects between the taxed goods. This independence of demands is a strong assumption, so it is not surprising that a clear result can be derived. The way the analysis works is to choose the optimal allocation and infer the tax rates from this. This was the argument used in the diagram when the intersection of the offer curve and the frontier of the production set was located and the tax rate derived from the implied budget constraint.

Consider a consumer who buys the two taxed goods and supplies labor. The consumer’s preferences are described by the utility function \( U(x_0, x_1, x_2) \), and his budget constraint is \( q_1 x_1 + q_2 x_2 = x_0 \). The utility-maximizing consumption levels of the two consumption goods are described by the first-order conditions \( U'_i = \alpha q_i \), \( i = 1, 2 \), where \( U'_i \) is the marginal utility of good \( i \) and \( \alpha \) is the marginal utility of income. The choice of labor supply satisfies the first-order condition \( U'_0 = -\alpha \).

With taxes \( t_1 \) and \( t_2 \) the government revenue constraint is \( R = t_1 x_1 + t_2 x_2 \). Since producer and consumer prices are related by \( t_i = q_i - p_i \), this can be written as

\[
q_1 x_1 + q_2 x_2 = R + p_1 x_1 + p_2 x_2. \tag{15.5}
\]

The optimal tax rates are inferred from an optimization whereby the government chooses the consumption levels to maximize the consumer’s utility while meeting the revenue constraint. This problem is summarized by the constrained maximization

\[
\max_{\{x_1, x_2\}} L = U(x_0, x_1, x_2) + \lambda [q_1 x_1 + q_2 x_2 - R - p_1 x_1 - p_2 x_2]. \tag{15.6}
\]

In this maximization the quantity of labor supply, \( x_0 \), is determined endogenously by \( x_1 \) and \( x_2 \) from the consumer’s budget constraint, \( x_0 = q_1 x_1 + q_2 x_2 \).

The basic assumption that the demands are independent can be used to write the (inverse) demand function \( q_i = q_i(x_i) \). Using these demand functions and the consumer’s budget constraint to replace \( x_0 \), we write the first-order condition for the quantity of good \( i \):
Chapter 15: Commodity Taxation

\[ U' + U' \left( q_i + x_i \frac{\partial q_i}{\partial x_i} \right) + \lambda \left[ q_i + x_i \frac{\partial q_i}{\partial x_i} - p_i \right] = 0. \quad (15.7) \]

The conditions \( U'_i = \alpha q_i \) and \( U'_0 = -\alpha \) can be used to write this as

\[-\alpha x_i \frac{\partial q_i}{\partial x_i} + \lambda t_i + \lambda x_i \frac{\partial q_i}{\partial x_i} = 0. \quad (15.8)\]

where \( t_i = q_i - p_i \). Now note that \( \frac{\partial q_i}{\partial x_i} = \frac{1}{\varepsilon_i^d} \), where \( \varepsilon_i^d \) is the elasticity of demand for good \( i \). The first-order condition can then be solved to write

\[ \frac{t_i}{p_i + t_i} = -\frac{\lambda - \alpha}{\lambda} \frac{1}{\varepsilon_i^d}. \quad (15.9) \]

Equation (15.9) is the inverse elasticity rule. This is interpreted by noting that \( \alpha \) is the marginal utility of another unit of income for the consumer and \( \lambda \) is the utility cost of another unit of government revenue. Since taxes are distortionary, \( \lambda > \alpha \). Since \( \varepsilon_i^d \) is negative, this makes the tax rate positive.

The inverse elasticity rule states that the proportional rate of tax on good \( i \) should be inversely related to its price elasticity of demand. Furthermore the constant of proportionality is the same for all goods. Recalling the discussion of the deadweight loss of taxation, it can be seen that this places more of the tax burden on goods where the deadweight loss is low. Its implication is clearly that necessities, which by definition have low elasticities of demand, should be highly taxed. It is this latter aspect that emphasizes the fact that the inverse elasticity rule describes an efficient way to tax commodities but not an equitable way. Placing relative high taxes on necessities will result in lower income consumers bearing relatively more of the commodity tax burden than high-income consumers.

15.5.2 The Ramsey Rule

The inverse elasticity rule is restricted by the fact that the demand for each good depends only on the price of that good. This rules out all cross-price effects in demand, meaning that the goods can be neither substitutes nor complements. When this restriction is relaxed, a more general tax rule is derived. The general result is called the Ramsey rule, and it is one of the oldest results in the theory of optimal taxation. It provides a description of the optimal taxes for an economy with a single consumer and with no equity considerations.
To derive the Ramsey rule, it is necessary to change from choosing the optimal quantities to choosing the taxes. Assume that there are just two consumption goods in order to simplify the notation, and let the demand function for good \( i \) be \( x_i = x_i(q) \) where \( q = q_1, q_2 \). The fact that the prices of all the commodities enter this demand function shows that the full range of interactions between the demands and prices are allowed. Using these demand functions, the preferences of the consumer can be written as

\[
U = U(x_0(q), x_1(q), x_2(q)).
\]

(15.10)

The optimal commodity taxes are those that give the highest level of utility to the consumer, while ensuring that the government reaches its revenue target of \( R > 0 \). The government’s problem in choosing the tax rates can then be summarized by the Lagrangean

\[
\max_{\{t_1, t_2\}} L = U(x_0(q), x_1(q), x_2(q)) + \lambda \left[ \sum_{i=1}^{2} t_i x_i(q) - R \right],
\]

(15.11)

where it is recalled that \( q_i = p_i + t_i \). Differentiating (15.11) with respect to the tax on good \( k \), we have the first-order necessary condition

\[
\frac{\partial L}{\partial t_k} = \sum_{i=0}^{2} U'_i \frac{\partial x_i}{\partial q_k} + \lambda \left[ x_k + \sum_{i=1}^{2} t_i \frac{\partial x_i}{\partial q_k} \right] = 0.
\]

(15.12)

This first-order condition needs some manipulation to place it in the form we want. The first step is to note that the budget constraint of the consumer is

\[
q_1 x_1(q) + q_2 x_2(q) = x_0(q).
\]

(15.13)

Any change in price of good \( k \) must result in demands that still satisfy this constraint so that

\[
q_1 \frac{\partial x_1}{\partial q_k} + q_2 \frac{\partial x_2}{\partial q_k} + x_k = \frac{\partial x_0}{\partial q_k}.
\]

(15.14)

In addition the conditions for optimal consumer choice are \( U'_0 = -\alpha \) and \( U'_i = \alpha q_i \). Using these optimality conditions and (15.14), we rewrite the first-order condition for the optimal tax, (15.12), as
\[ \alpha x_k = \lambda \left[ x_k + \sum_{i=1}^{2} t_i \frac{\partial x_i}{\partial q_k} \right]. \]  

(15.15)

Notice how this first-order condition involves quantities rather than the prices that appeared in the inverse elasticity rule. After rearrangement, (15.15) becomes

\[ \sum_{i=1}^{2} t_i \frac{\partial x_i}{\partial q_k} = -\left[ \frac{\lambda - \alpha}{\lambda} \right] x_k. \]  

(15.16)

The next step in the derivation is to employ the Slutsky equation, which breaks the change in demand into the income and substitution effects. The effect of an increase in the price of good \( k \) upon the demand for good \( i \) is determined by the Slutsky equation as

\[ \frac{\partial x_i}{\partial q_k} = S_{ik} - x_k \frac{\partial x_i}{\partial I}, \]  

(15.17)

where \( S_{ik} \) is the substitution effect of the price change (the move around an indifference curve) and \( -x_k \frac{\partial x_i}{\partial I} \) is the income effect of the price change (\( I \) denotes lump-sum income). Substituting from (15.17) into (15.16) gives

\[ \sum_{i=1}^{2} t_i \left[ S_{ik} - x_k \frac{\partial x_i}{\partial I} \right] = -\left[ \frac{\lambda - \alpha}{\lambda} \right] x_k. \]  

(15.18)

Equation (15.18) is now simplified by extracting the common factor \( x_k \), which yields

\[ \sum_{i=1}^{2} t_i S_{ik} = \left[ 1 - \frac{\alpha}{\lambda} - \sum_{i=1}^{2} t_i \frac{\partial x_i}{\partial I} \right] x_k. \]  

(15.19)

The substitution effect of a change in the price of good \( i \) on the demand for good \( k \) is exactly equal to the substitution effect of a change in the price of good \( k \) on the demand for good \( i \) because both are determined by movement around the same indifference curve. This symmetry property implies \( S_{ki} = S_{ik} \), which can be used to rearrange (15.19) to give the expression

\[ \sum_{i=1}^{2} t_i S_{ki} = -\theta x_k, \]  

(15.20)
where \( \theta = \left[ 1 - \frac{\alpha}{\lambda} - \sum_{i=1}^{2} t_i \frac{\partial y_i}{\partial y} \right] \) is a positive constant. Equation (15.20) is the Ramsey rule describing a system of optimal commodity taxes and an equation of this form must hold for all goods, \( k = 1, \ldots, n \).

The optimal tax rule described by (15.20) can be used in two ways. If the details of the economy are specified (the utility function and production parameters), then the actual tax rates can be calculated. Naturally the precise values would be a function of the structure chosen. Although this is the direction that heads toward practical application of the theory (and more is said later), it is not the route that will be currently taken. The second use of the rule is to derive some general conclusions about the determinants of tax rates. This is done by analyzing and understanding the different components of (15.20).

To proceed with this, the focus on the typical good \( k \) is maintained. Recall that a substitution term measures the change in demand with utility held constant. Demand defined in this way is termed compensated demand. Now begin in an initial position with no taxes. From this point the tax \( t_i \) is the change in the tax rate on good \( i \). Then \( t_i S_{ki} \) is a first-order approximation to the change in compensated demand for good \( k \) due to the introduction of the tax \( t_i \). If the taxes are small, this will be a good approximation to the actual change. Extending this argument to take account of the full set of taxes, it follows that \( \sum_{i=1}^{2} t_i S_{ki} \) is an approximation to the total change in compensated demand for good \( k \) due to the introduction of the tax system from the initial no-tax position. In employing this approximation, the Ramsey rule can be interpreted as saying that the optimal tax system should be such that the compensated demand for each good is reduced in the same proportion relative to the before-tax position. This is the standard interpretation of the Ramsey rule.

The importance of this observation is reinforced when it is set against the alternative, but incorrect, argument that the optimal tax system should raise the prices of all goods by the same proportion in order to minimize the distortion caused by the tax system. This is shown by the Ramsey rule to be false. What the Ramsey rule says is that it is the distortion in terms of quantities, rather than prices, that should be minimized. Since it is the level of consumption that actually determines utility, it is not surprising that what happens to prices is secondary to what happens to quantities. Prices only matter so far as they determine demands.

Although the actual tax rates are only implicit in the Ramsey rule, some general comments can still be made. By the approximation interpretation, the rule suggests that as the proportional reduction in compensated demand must be the same for all goods, and those goods whose demand is unresponsive to price changes must bear higher taxes in order to achieve the same reduction. Although broadly correct, this statement can
only be completely justified when all cross-price effects are accounted for. One simple case that overcomes this difficulty is that in which there are no cross-price effects among the taxed goods. This is the special case that led to the inverse elasticity rule.

Returning to the general case, goods that are unresponsive to price changes are typically necessities such as food and housing. Consequently using the Ramsey rule leads to a tax system that bears most heavily on necessities. In contrast, the lowest tax rates would fall on luxuries. If put into practice, such a tax structure would involve low-income consumers paying disproportionately larger fractions of their incomes in taxes relative to high-income consumers. The inequitable nature of this is simply a reflection of the single-consumer assumption: the optimization does not involve equity and the solution reflects only efficiency criteria.

The single-consumer framework is not accurate as a description of reality, and it leads to an outcome that is unacceptable on equity grounds. The value of the Ramsey rule therefore arises primarily through the framework and method of analysis it introduces. This can easily be generalized to more relevant settings. It shows how taxes are determined by efficiency considerations and hence gives a baseline from which to judge the effects of introducing equity.

15.6 Equity Considerations

The lack of equity in the tax structure determined by the Ramsey rule is inevitable given its single-consumer basis. The introduction of further consumers who differ in incomes and preferences makes it possible to see how equity can affect the conclusions. Although the method that is now discussed can cope with any number of consumers, it is sufficient to consider just two. Restricting the number in this way has the merit of making the analysis especially transparent.

Consider then an economy that consists of two consumers. Each consumer \( h, h = 1, 2 \), is described by their (indirect) utility function

\[
U^h = U^h(x_{0}^{h}(q), x_{1}^{h}(q), x_{2}^{h}(q)).
\]  

These utility functions may vary between the consumers. Labor remains the untaxed numéraire, and all consumers supply only the single form of labor service.

The government revenue constraint is now given by

\[
R = \sum_{i=1}^{2} t_{i} x_{i}^{1}(q) + \sum_{i=1}^{2} t_{i} x_{i}^{2}(q),
\]  

(15.22)
where the first term on the right-hand side is the total tax payment of consumer 1 and the second term is the total tax payment of consumer 2. The government’s policy is guided by a social welfare function that aggregates the individual consumers’ utilities. This social welfare function is denoted by

\[ W = W\left(U^1\left(x_1^0, x_1^1, x_1^2\right), U^2\left(x_2^0, x_2^1, x_2^2\right)\right). \]  

Combining (15.22) and (15.23) into a Lagrangean expression (as in equation 15.11), we have the first-order condition for the choice of the tax on good \( k \):

\[
\frac{\partial L}{\partial t_k} = -\frac{\partial W}{\partial U^1} \alpha_1 x_1^k - \frac{\partial W}{\partial U^2} \alpha_2 x_2^k + \lambda \left[ \sum_{h=1}^{2} \left( x_h^k + \sum_{i=1}^{2} t_i \frac{\partial x_i^h}{\partial q_k} \right) \right] = 0. 
\]

The expression in (15.24) has been derived by using (15.13), (15.14), and the first-order conditions for consumer choice \( \frac{\partial U^h}{\partial x^h_1} = \alpha^h q_k \) to deduce that

\[
\frac{\partial U^h}{\partial x^h_0} \frac{\partial x^h_0}{\partial q_k} + \frac{\partial U^h}{\partial x^h_1} \frac{\partial x^h_1}{\partial q_k} + \frac{\partial U^h}{\partial x^h_2} \frac{\partial x^h_2}{\partial q_k} = -\alpha^h x^h_k, \quad h = 1, 2. 
\]

To obtain a result that is easily comparable to the Ramsey rule, define

\[
\beta^h = \frac{\partial W}{\partial U^h} \alpha^h. 
\]

\( \beta^h \) is formed as the product of the effect of an increase in consumer \( h \)'s utility on social welfare and their marginal utility of income. It measures the increase in social welfare that results from a marginal increase in the income of consumer \( h \). Consequently \( \beta^h \) is termed the social marginal utility of income for consumer \( h \). Employing the definition of \( \beta^h \) and the substitutions used to obtain the Ramsey rule, we write the first-order condition (15.24) as

\[
\frac{\sum_{i=1}^{2} t_i S_{ki} x_1^1}{x_1^1 + x_1^k} + \frac{\sum_{i=1}^{2} t_i S_{ki} x_2^1}{x_1^1 + x_1^k} = \lambda \left( \frac{\beta^1 x_1^1}{x_1^1 + x_1^k} + \frac{\beta^2 x_2^1}{x_1^1 + x_1^k} \right) - 1 + \left[ \frac{\sum_{i=1}^{2} t_i \frac{\partial x_i^1}{\partial y_k} x_1^1}{x_1^1 + x_1^k} + \frac{\sum_{i=1}^{2} t_i \frac{\partial x_i^2}{\partial y_k} x_2^2}{x_1^1 + x_1^k} \right]. 
\]

The tax structure that is described by (15.27) can be interpreted in the same way as the Ramsey rule. The left-hand side is approximately the proportional change in aggregate
compensated demand for good \( k \) caused by the introduction of the tax system from an initial position with no taxes. When a positive amount of revenue is to be raised (so that \( R > 0 \)), the level of demand will be reduced by the tax system, so this term will be negative.

The first point to observe about the right-hand side is that unlike the Ramsey rule, the proportional reduction in compensated demand is not the same for all goods. It is therefore necessary to discuss the factors that influence the extent of the reduction, and it is by doing this that the consequences of equity can be seen. The essential component in this regard is the first term on the right-hand side. The proportional reduction in demand for good \( k \) will be smaller the larger is the value of \( \beta^1 \frac{x^1}{x^1 + x^2} + \beta^2 \frac{x^2}{x^1 + x^2} \). The value of this will be large if a high \( \beta^h \) is correlated with \( \frac{x^h}{x^1 + x^2} \). The meaning of this is clear, since a consumer will have a high value of \( \beta^h \) when their personal marginal utility of income, \( \alpha^h \), is large and when \( \frac{\partial W}{\partial U^h} \) is also large so that the social planner gives them a high weight in social welfare. If the social welfare function is concave, both of these will be satisfied by low-utility consumers with low incomes. The term \( \frac{x^h}{x^1 + x^2} \) will be large when good \( k \) is consumed primarily by consumer \( h \). Putting these points together, we have that the proportional reduction in the compensated demand for a good will be smaller if it is consumed primarily by the poor consumer. This is the natural reflection of equity considerations.

The second term on the right-hand side shows that the proportional reduction in demand for good \( k \) will be smaller if its demand comes mainly from the consumer whose tax payments change most as income changes. This term is related to the efficiency aspects of the tax system. If taxation were to be concentrated on goods consumed by those whose tax payments fell rapidly with reductions in income, then increased taxation, and consequently greater distortion, would be required to meet the revenue target.

This has shown how the introduction of equity modifies the conclusions of the Ramsey rule. Rather than all goods having their compensated demand reduced in the same proportion, equity results in the goods consumed primarily by the poor facing less of a reduction. In simple terms, this should translate into lower rates of tax on the goods consumed by the poor relative to those determined solely by efficiency. Equity therefore succeeds in moderating the hard edge of the efficient tax structure.

### 15.7 Applications

At this point in the discussion it should be recalled that the fundamental motive for the analysis is to provide practical policy recommendations. The results that have been
derived do give some valuable insights: the need for production efficiency and the non-uniformity of taxes being foremost among them. Accepting this, the analysis is only of real merit if the tax rules are capable of being applied to data and the actual values of the resulting optimal taxes calculated. The numerical studies that have been undertaken represent the development of a technology for achieving this aim and also provide further insights into the structure of taxation.

Referring back to (15.27), it can be seen that two basic pieces of information are needed in order to calculate the tax rates. The first is knowledge of the demand functions of the consumers. This provides the levels of demand $x^h_k$ and the demand derivatives $\frac{\partial x^h_k}{\partial q^i}$.

The second piece of information is the social marginal utilities of income, $\beta^h$. Ideally these should be calculated from a specified social welfare function and individual utility functions for the consumers. The problem here is the same as that raised in previous chapters: the construction of some meaningful utility concept. The difficulties are further compounded in the present case by the requirement that the demand functions also be consistent with the utility functions.

In practice, the difficulties are circumvented rather than solved. The approach that has been adopted is first to ignore the link between demand and utility and to impose a procedure to obtain the social marginal utilities of income. The demand functions are then estimated using standard econometric techniques. One common procedure is to combine the utility function defined by (14.24) with a utilitarian social welfare function.

Hence assume that the social welfare function is utilitarian ($W = \sum_h U^h$) and that the individual utility functions are isoelastic ($U^h = K \frac{M^h^{1-\varepsilon}}{1-\varepsilon}$). The social marginal utility for $h$ is then given by

$$\beta^h = K \left[ M^h \right]^{-\varepsilon}.$$

The value of the parameter $K$ can be fixed by, for instance, setting the value of $\beta^h$ equal to 1 for the lowest income consumer. With $\varepsilon > 0$ the social marginal utility declines as income rises. It decreases faster as $\varepsilon$ rises, so relatively more weight is given to low-income consumers. This way the value of $\varepsilon$ can be treated as a measure of the concern for equity.

15.7.1 Reform

The first application of the analysis is to consider marginal reforms of tax rates. By a marginal reform it is meant a small change from the existing set of tax rates that moves
the system closer to optimality. This should be distinguished from an optimization of
tax rates that might imply a very significant change from the initial set of taxes.

Marginal reforms are much easier to compute than optimal taxes, since it is only
necessary to evaluate effect of changes not of the whole move. An analogy can be drawn
with hill-climbing: to climb higher, you only need to know which direction leads upward
and do not need to know where the top is. Essentially studying marginal reforms reduces
the informational requirement.

Return to the analysis of the optimal taxes in the economy with two consumers. The
effect on welfare of a marginal increase in the tax on good \( k \) is

\[
\frac{\partial W}{\partial t_k} = - \sum_{i=1}^{2} \beta^i x^i_k, \tag{15.29}
\]

and the effect on revenue is

\[
\frac{\partial R}{\partial t_k} = \sum_{h=1}^{2} \left[ x^h_k + \sum_{i=1}^{2} t_i \frac{\partial x^h_i}{\partial q_k} \right] = X_k + \sum_{i=1}^{2} t_i \frac{\partial X_i}{\partial q_k}, \tag{15.30}
\]

where \( X_i \) is the aggregate demand for good \( i \). The marginal revenue benefit of taxation
of good \( k \) is defined as the extra revenue generated relative to the welfare change of a
marginal increase in a tax. This can be written as

\[
MRB_k = - \frac{\partial R}{\partial t_k} \frac{\partial W}{\partial t_k}. \tag{15.31}
\]

At the optimum all goods should have the same marginal revenue benefit. If that was
not the case, taxes could be raised on those with a high marginal revenue benefit and
reduced for those with a low value. This is exactly the process we can use to deduce
the direction of reform.

From the marginal revenue benefit the economy of information can be clearly seen.
All that is needed to evaluate \( MRB_k \) are the social marginal utilities, \( \beta^h \), the individual
commodity demands, \( x^h_k \), and the aggregate derivatives of demand \( \frac{\partial X_i}{\partial q_k} \) (or, equally, the
aggregate demand elasticities). The demands and the elasticities are easily obtainable
from data sets on consumer demands.

Table 15.1 displays the result of a calculation of the \( MRB_k \) using Irish data for
ten commodity categories in 1987. Two different values of \( \varepsilon \) are given, with \( \varepsilon = 5 \)
representing a greater concern for equity. The interpretation of these figures is that the
Table 15.1

<table>
<thead>
<tr>
<th>Good</th>
<th>$\varepsilon = 2$</th>
<th>$\varepsilon = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other goods</td>
<td>2.316</td>
<td>4.349</td>
</tr>
<tr>
<td>Services</td>
<td>2.258</td>
<td>5.064</td>
</tr>
<tr>
<td>Petrol</td>
<td>1.785</td>
<td>3.763</td>
</tr>
<tr>
<td>Food</td>
<td>1.633</td>
<td>3.291</td>
</tr>
<tr>
<td>Alcohol</td>
<td>1.566</td>
<td>3.153</td>
</tr>
<tr>
<td>Transport and equipment</td>
<td>1.509</td>
<td>3.291</td>
</tr>
<tr>
<td>Fuel and power</td>
<td>1.379</td>
<td>2.221</td>
</tr>
<tr>
<td>Clothing and footwear</td>
<td>1.341</td>
<td>2.837</td>
</tr>
<tr>
<td>Durables</td>
<td>1.234</td>
<td>2.514</td>
</tr>
<tr>
<td>Tobacco</td>
<td>0.420</td>
<td>0.683</td>
</tr>
</tbody>
</table>


The tax levied on the goods toward the top of the table should be raised and the tax should be lowered on the goods at the bottom. Hence services should be more highly taxed and the tax on tobacco should be reduced! The rankings are fairly consistent for both values of $\varepsilon$; there is some movement, but no good moves very far. Therefore a reform based on these data would be fairly robust to changes in the concern for equity.

15.7.2 Optimality

The most developed implementation of the optimal tax rule for an economy with many consumers uses data from the Indian National Sample Survey. Defining $\theta$ to be the wage as a proportion of expenditure, a selection of these results are given in table 15.2 for $\varepsilon = 2$. The table shows that these tax rates achieve some redistribution, since cereals and milk products, both basic foodstuffs, are subsidized. Such redistribution results from the concern for equity embodied in a value of $\varepsilon$ of 2. Interesting as they are, these results are limited, as are other similar analyses, by the degree of commodity aggregation that leads to the excessively general other nonfood category.

The same dataset has been used to analyze the redistributive impact of Indian commodity taxes. The redistributive impact is found by calculating the total payment of commodity tax, $T^h$, by consumer $h$ relative to the expenditure, $\mu^h$, of the consumer. The net gain from the tax system for $h$ is then defined by $-\frac{T^h}{\mu^h}$. The consumer gains from the tax system if $-\frac{T^h}{\mu^h}$ is positive, since this implies that a net subsidy is being received. Contrasting the gain of a consumer from the existing tax system with the gain under the
optimal system provides an indication of both the success of the existing system and the potential gains from the optimal system. The calculations for the existing Indian tax system give the gains shown in table 15.3. The expenditure levels of Rs. 20 and Rs. 50 place consumers with these incomes in the lower 30 percent of the income distribution. The table shows a net gain to consumers at both income levels from the tax system, with the lower expenditure consumer making a proportionately greater gain.

The same calculations can be used to find the redistributive impact of the optimal tax system for a consumer with expenditure level $\mu = 0.5\bar{\mu}$, where $\bar{\mu}$ is mean expenditure, is given in table 15.4. For $\varepsilon = 1$ or more, it can be seen that the potential gains from the tax system, relative to the outcome that would occur in the absence of taxation, are substantial. This shows that with sufficient weight given to equity considerations, the optimal set of commodity taxes can effect significant redistribution and that the existing Indian tax system does not attain these gains.
This section has discussed a method for calculating the taxes implied by the optimal tax rule. The only difficulty in doing this is the specification of the social welfare weights. To determine these, it is necessary to know both the private utility functions and the social welfare function. In the absence of this information, a method for deriving the weights is employed that can embody equity criteria in a flexible way. Although these weights are easily calculated, they are not entirely consistent with the other components of the model. The numbers derived demonstrate clearly that when equity is embodied in the optimization, commodity taxes can secure a significant degree of redistribution. This is very much in contrast to what occurs with efficiency alone.

### 15.8 Efficient Taxation

The tax rules in the previous section have only considered the competitive case. When there is imperfect competition, additional issues have to be taken into account. The basic fact is that imperfectly competitive firms produce less than the efficient output level, so the equilibrium without intervention is not Pareto-efficient. This gives a reason to use commodity taxes to subsidize the output of imperfectly competitive firms relative to that of competitive firms. However, the strength of this argument depends on the degree of tax-shifting, as identified in chapter 9.

The issues involved in tax design can be understood by determining the direction of welfare-improving tax reform starting from an initial position with no commodity taxation. This is undertaken for an economy with a single consumer and a zero-revenue requirement. The fact that no revenue is raised implies that the taxes are used merely to correct for the distortion introduced by the imperfect competition. There are two consumption goods, each produced using labor alone. Good 1 is produced with constant returns to scale by a competitive industry with after-tax price $q_1 = p_1 + t_1$. There is a single household in the economy whose (indirect) utility function is

$$U = U \left( x_0 (q_1, q_2), x_1 (q_1, q_2), x_2 (q_1, q_2) \right).$$

(15.32)
Tax revenue, $R$, is defined by

$$R = t_1 x_1 + t_2 x_2.$$  

(15.33)

Good 2 is produced by a monopolist who chooses its output to maximize profit

$$\pi_2 = [q_2 - c - t_2] x_2 (q_1, q_2),$$  

(15.34)

where $c$ is the constant marginal cost. The profit-maximizing price depends on the tax, $t_2$, and the price of good 1, $q_1$. This relationship is denoted $q_2 = q_2 (q_1, t_2)$. The derivative $\frac{\partial q_2}{\partial t_2}$ measures the rate of shifting of the tax. In the terminology of chapter 9, there is undershifting if $\frac{\partial q_2}{\partial t_2} < 1$ and overshifting if $\frac{\partial q_2}{\partial t_2} > 1$. The dependence of the demand for good 2 on the price of good 1 is reflected in the profit-maximizing price. The derivative $\frac{\partial q_2}{\partial q_1}$ is the cross-price effect of taxation. It can be positive or negative, and since $q_1 = p_1 + t_1$, it follows that $\frac{\partial q_2}{\partial q_1} = \frac{\partial q_2}{\partial t_1}$.

The tax reform problem searches for a pair of tax changes that raises welfare while collecting zero revenue. The initial position is taken to be one where both commodity taxes are zero initially, so the intention is to find a pair of tax changes $dt_1, dt_2$, starting from an initial position with $t_1 = t_2 = 0$, such that $dU > 0$ and $dR = 0$. The formulation ensures that one of the taxes will be negative and the other positive. The aim is to provide a simple characterization of the determination of the relative rates. It should be noted that if both industries were competitive the initial equilibrium would be Pareto-efficient and the solution to the tax problem would have $dt_1 = dt_2 = 0$. So nonzero tax rates will be a consequence of the distortion caused by the imperfect competition.

Totally differentiating the utility function and using the first-order conditions for consumer choice, the effect of the tax changes on utility is

$$dU = -\alpha x_1 \frac{\partial q_1}{\partial t_1} dt_1 - \alpha x_2 \frac{\partial q_2}{\partial t_1} dt_1 - \alpha x_2 \frac{\partial q_2}{\partial t_2} dt_2,$$  

(15.35)

where $\alpha$ is the consumer’s marginal utility of income. Totally differentiating the revenue constraint and using the fact that the initial values of the taxes are $t_1 = t_2 = 0$ gives

$$dR = 0 = x_1 dt_1 + x_2 dt_2.$$  

(15.36)

Solving (15.36) for $dt_1$ and substituting into the derivative of utility determines the utility change as dependent on $dt_2$ alone:

$$dU = \left[-\alpha x_2 \frac{\partial q_2}{\partial t_2} + \alpha x_2 + \alpha x_2 \frac{\partial q_2}{\partial t_1}\right] dt_2.$$  

(15.37)
It is condition (15.37) that provides the key to understanding the determination of the relative tax rates. Since we wish to choose the tax change \( dt_2 \) to ensure that \( dU > 0 \), it follows that the sign of the tax change must be the same as the bracketed term in (15.37). From this observation follows the conclusion that

\[
x_1 \left[ 1 - \frac{\partial q_2}{\partial t_2} \right] + x_2 \frac{\partial q_2}{\partial t_1} < 0 \Rightarrow dt_2 < 0.
\]  

(15.38)

From (15.38) the output of the imperfectly competitive industry should be subsidized and the competitive industry taxed when \( \frac{\partial q_2}{\partial t_2} \) is large, so that overshifting is occurring and \( \frac{\partial q_2}{\partial t_1} \) is negative. These are, of course, sufficient conditions. In general, the greater the degree of tax shifting, the more likely is subsidization. The explanation for this result is that if firms overshift taxes, they will also do the same for any subsidy. Hence a negative \( dt_2 \) will be reflected by an even greater reduction in price. If \( \frac{\partial q_2}{\partial t_1} \) is also negative, the tax on the competitive industry secures a further reduction in the price of good 2.

The conclusion of this analysis is that the rate of tax-shifting is important in the determination of relative rates of taxation. Although the economy is simplified, it does demonstrate that with imperfect competition commodity taxation can be motivated on efficiency grounds alone to mitigate the inefficiency cost of market power.

15.9 Public Sector Pricing

The theory that was developed in the previous sections also has a second application. This arises because there are close connections between the theory of commodity taxation and that of choosing optimal public sector prices. Firms operated by the public sector can be set the objective of choosing their pricing policy to maximize social welfare subject to a revenue target. If the firms have increasing returns to scale, which is often the reason they are operated by the public sector, then marginal cost pricing will lead to a deficit (because marginal cost is below the decreasing average cost). The government will then want to find the optimal deviation from marginal-cost pricing that ensures that the firms break even.

For both commodity taxation and public sector pricing, the government is choosing the set of consumer prices that maximizes welfare subject to a revenue constraint. Under the commodity taxation interpretation, these prices are achieved by setting the level of tax to be included in each consumer price, whereas with public sector pricing, the prices are chosen directly. However, the choice of tax rate is equivalent to the choice of consumer price.
In the context of public sector pricing, the optimal prices are generally known as Ramsey prices. The constraint on the optimization with commodity taxation requires the raising of a specified level of revenue. With public sector pricing this can be reinterpreted as the need to raise a given level of revenue in excess of marginal cost. The tax rates of the commodity taxation problem then translate into the markup over marginal cost in the public sector pricing interpretation. The rules for optimal taxation derived above then characterize the public sector prices.

15.10 Conclusions

This chapter has reviewed the determination of optimal commodity taxes. It has been shown how an efficient system places the burden of taxation primarily on necessities. If implemented, such a system would be very damaging to low-income consumers. When equity is introduced, this outcome is modified to reduce the extent to which goods consumed primarily by those with low incomes are affected by the tax system. These interpretations were borne out by the numerical calculations.

As well as providing these insights into the structure of taxes, the chapter has also been shown that the optimal tax system should ensure production efficiency. The implication of this finding is that there should be no taxes on intermediate goods. This is a very strong and clear prediction. It is also a property that actual value-added tax systems satisfy.

Further Reading

The theory of optimal commodity taxation was given its modern form in:

A simplified version of the optimal tax rule for a many-consumer economy is developed in:

An argument for uniform taxation is presented by:

The welfare effects of tax reform are analyzed in:


The extension of efficient taxation to imperfect competition is described in:


Public sector pricing is described in:


### Exercises

15.1 For the linear demand function \( x = a - bp \), calculate the deadweight loss of introducing a commodity tax \( t \) when the marginal cost of production is constant at \( c \). How is the deadweight loss affected by changes in \( a \) and \( b \)? How does a change in \( b \) affect the elasticity of demand at the equilibrium without taxation?

15.2 A good is traded in a competitive market. The demand function is given by \( X = 75 - 5P \) and supply is perfectly elastic at the price \( P = 10 \).
   a. A specific tax of value \( t = 2 \) is introduced. Determine the tax incidence.
   b. An ad valorem tax at a rate of \( t = 0.2 \) is introduced. Determine the tax incidence.
   c. How do the incidence of the specific tax and the ad valorem tax differ if supply is given by \( Y = 2.5P \)?

15.3 Assume that the demand function is given by \( x = p^{-\varepsilon_d} \) and the supply function by \( y = p^{\varepsilon_s} \). Find the equilibrium price. What is the effect on the equilibrium price of the introduction of a tax \( t = \frac{1}{10} \) if \( \varepsilon_d = \varepsilon_s = \frac{1}{2} \)? Describe how the incidence of the tax is divided between consumers and suppliers.

15.4 The analysis of taxation in the single-consumer economy used labor as an untaxed numéraire. Show that the optimal allocation with commodity taxation is unchanged when the consumption good becomes the untaxed numéraire. Then establish that it does not matter which good is the numéraire and which is taxed.

15.5 The value-added tax system requires that all goods to be sold at a price that includes tax. Any firm purchasing a good to use as an input can reclaim the tax it has paid. Assess this tax structure using the Diamond-Mirrlees Production Efficiency Lemma.

15.6 Consider an economy with a single consumer whose preferences are given by \( U = \log(x) - \ell \), where \( x \) is consumption and \( \ell \) labor supply. Assume that the consumption good is produced using labor alone with a constant-returns-to-scale technology. Units of measurement are chosen so that the producer prices of both the consumption good and the wage rate are equal to 1.
a. Let the consumer’s budget constraint be $q_x = \ell$, where the consumer price is $q = 1 + t$, and $t$ is the commodity tax. By maximizing utility, find the demand function and the labor supply function.

b. Assume the revenue requirement of the government is $\frac{1}{10}$ of a unit of labor. Draw the production possibilities for the economy and the consumer’s offer curve.

c. By using the offer curve and the production possibilities, show that the optimal allocation with commodity taxation has $x = \frac{9}{10}$ and $\ell = 1$.

d. Calculate the optimal commodity tax.

e. By deriving the first-best allocation, show that the commodity tax optimum is second-best.

15.7 Two consumers $A$ and $B$ have an income of $30,000$ and $100,000$ respectively. $A$ and $B$ consume the same bundle of goods with a cost (including tax) of $24,000$. The only tax in the economy is a commodity tax levied uniformly on all goods at a rate of 20 percent.

a. What proportion of income is paid in tax by $A$ and $B$?

b. What implications does such a tax have in terms of equity?

c. Is there any way the commodity tax can be restructured to improve its equity properties?

15.8 For an economy with one consumption good that is produced using only labor, show that at the optimal allocation with commodity taxation tax revenue, $tx$, is equal to the government use of labor, $R$.

15.9 Assume that the production technology is such that each unit of output requires one unit of labor and that the government has a revenue requirement of one unit of labor. Also assume that there is a single consumer.

a. Using a diagram, describe how the optimal tax on the consumption good is determined. Now assume that the consumer has preferences given by $U = \log(x) + \log(10 + \ell)$, where $x$ is consumption and $\ell$ is labor supply (recall that $\ell$ is a supply, so is a negative number).

b. By maximizing utility subject to the budget constraint $q_x + w\ell = 0$, construct the consumer’s offer curve.

c. Treating the equations of the production frontier and the offer curve as a simultaneous system, determine the optimal tax rate.

15.10 An economy has a single consumption good produced using labor and a single consumer. The production process has decreasing returns to scale. Explain the derivation of the optimal commodity tax when profit is not taxed.

15.11 Consider the utility function $U = \alpha \log x_1 + \beta \log x_2 - \ell$ and budget constraint $w\ell = q_1 x_1 + q_2 x_2$.

a. Show that the price elasticity of demand for both commodities is equal to $-1$.

b. Setting producer prices at $p_1 = p_2 = 1$, show that the inverse elasticity rule implies $\frac{\alpha}{\beta} = \frac{q_1}{q_2}$.

c. Letting $w = 100$ and $\alpha + \beta = 1$, calculate the tax rates required to achieve revenue of $R = 10$.

15.12 Let the consumer have the utility function $U = x_1^{\rho_1} + x_2^{\rho_2} - \ell$. 

a. Show that the utility maximizing demands are
\[ x_1 = \left( \frac{\rho_1 w}{q_1} \right)^{1/(1-\rho_1)} \] and
\[ x_2 = \left( \frac{\rho_2 w}{q_2} \right)^{1/(1-\rho_2)}. \]

b. Letting \( p_1 = p_2 = 1 \), use the inverse elasticity rule to show that the optimal tax rates are related by
\[ \frac{1}{t_2} = \left( \frac{\rho_2 - \rho_1}{1-\rho_2} \right) + \left( \frac{1-\rho_1}{1-\rho_2} \right) t_1. \]

c. Setting \( w = 100, \rho_1 = 0.75, \) and \( \rho_2 = 0.5 \), find the tax rates required to achieve revenue of \( R = 0.5 \) and \( R = 10 \). Comment on the results.

d. Calculate the proportional reduction in demand for the two goods comparing the no-tax position with the position after imposition of the optimal taxes for both revenue levels.

15.13 “If all commodities are taxed at the same rate, the distortion in prices is minimized.” Explain why this statement does not act as a guide for setting optimal commodity taxes.

15.14 Consider an economy with a single consumer whose preferences are given by
\[ U = \log(x_1) + \log(x_2) + \ell, \]
where \( x_1 \) and \( x_2 \) are the consumption levels of goods 1 and 2 and \( \ell \) is leisure. Assume that both goods are produced using labor alone, subject to a constant-returns-to-scale technology. Units of measurement are chosen so that the producer prices of both goods and the wage rate are equal to 1.

a. Using \( L \) to denote the consumer’s endowment of time, explain the budget constraint
\[ q_1 x_1 + q_2 x_2 + w \ell = wL. \]

b. Show that the consumer’s demands satisfy the conditions required for the inverse elasticity rule to apply.

c. Use the inverse elasticity rule to conclude that both goods should be subject to the same level of tax.

d. Calculate the tax required to obtain a level of revenue of \( R = 1 \).

e. Show that the commodity taxes are second-best.

15.15 Show how the impact of a commodity tax upon a consumer’s optimal demand can be separated into an “income effect” and a “substitution effect.”

15.16 In the absence of taxation a consumer has the budget constraint
\[ p_1 x_1 + p_2 x_2 - w \ell = 0. \]
Show that an ad valorem tax levied at rate \( t \) on both commodities and on labor raises no revenue. Explain this fact.

15.17 (Ramsey rule) Consider a three-good economy \((k = 1, 2, 3)\) in which every consumer has preferences represented by the utility function \( U = x_1 + g(x_2) + h(x_3) \), where the functions \( g(\cdot) \) and \( h(\cdot) \) are increasing and strictly concave. Suppose that each good is produced with constant returns to scale from good 1, using one unit of good 1 per unit of good \( k \neq 1 \). Let good 1 be the numéraire, and normalize the price of good 1 to equal 1. Let \( t_k \) denote the (specific) commodity tax on good \( k \) so that the consumer price is \( q_k = (1 + t_k) \).

a. Consider two commodity tax schemes \( t = (t_1, t_2, t_3) \) and \( t' = (t'_1, t'_2, t'_3) \). Show that if \( 1 + t'_k = \phi [1 + t_k] \) for \( k = 1, 2, 3 \) for some scalar \( \phi > 0 \), then the two tax schemes raise the same amount of tax revenue.

b. Argue from part a that the government can without cost restrict tax schemes to leave one good untaxed.
c. Set \( t_1 = 0 \), and suppose that the government must raise revenue of \( R \). What are the tax rates on goods 2 and 3 that minimize the welfare loss from taxation?

d. Show that the optimal taxes are inversely proportional to the elasticity of the demand for each good. Discuss this tax rule.

e. When should both goods be taxed equally? Which good should be taxed more?

15.18 Consider a three-good economy \((k = 1, 2, 3)\) in which every consumer has preferences represented by the utility function \( U = x_1 + g(x_2, x_3) \), where the function \( g(\cdot) \) is increasing and strictly concave. Suppose that each good is produced with constant returns to scale from good 1, using one unit of good 1 per unit of good \( k \neq 1 \). Let good 1 be the numéraire and normalize the price of good 1 to equal 1. Let \( t_k \) denote the (specific) commodity tax on good \( k \) so that the consumer price is \( q_k = 1 + t_k \). Suppose that a tax change is restricted to only good 2 so that \( t'_2 = t_2 + \Delta \) with \( \Delta > 0 \).

a. What is the correct measure of the welfare loss arising from this tax increase if \( t_3 = 0 \)?

b. Show that if \( t_3 > 0 \), then the measure of welfare loss in part a overestimates the welfare loss if good 3 is a substitute for good 2. What is then the correct measure of the welfare change?

c. Show that if \( t_3 > 0 \), then the measure of welfare loss in part a underestimates the welfare loss if good 3 is a complement for good 2. What is the correct welfare change?

d. Show that if good 3 is subsidized, \( t_3 < 0 \), then the measure of welfare loss in part a underestimates the welfare loss if good 3 is a substitute for good 2. How can you explain this result?

e. Show that if good 3 is subsidized, \( t_3 < 0 \), then the measure of welfare loss in part a overestimates the welfare loss if good 3 is a complement for good 2.

15.19 The purpose of this exercise is to contrast the incidence of a commodity tax under different market structures. Consider an economy with identical households and identical firms. The representative household receives labor income for its labor supply \( \ell \) and profit income \( \pi \) for its ownership of the firm. The utility function of the household is \( U = 2\sqrt{x} - \ell \). The firm produces one unit of final consumption good \( x \) with one unit of labor input. Labor is the numéraire good: the price of labor is normalized to 1, and labor is the untaxed good. The producer price is \( p \) and the consumer price is \( q = p + t \), where \( t > 0 \) is the (specific) commodity tax.

a. Describe the household’s optimization program treating profit income and the consumer prices in the budget constraint as fixed. Find the demand for good \( x \) as a function of consumer price \( q \).

b. Calculate the elasticity of the slope of the inverse demand function.

c. Suppose that the firms act in unison like a monopolist. Find the supply of the monopoly as a function of \( t \).

d. What is the equilibrium price charged by the monopolist? What is the producer price? What is the division of the tax burden between the producer and the consumer?

e. Suppose that the firms act independently maximizing their own profit-taking prices as given. What is the equilibrium producer price? What is the division of the tax burden between producer and consumer? Compare with the result in part d.
15.20 Consider an economy with two representative households \((h = 1, 2)\) that supply labor \(\ell^h\) to the one representative firm and buy a consumption good \(x^h\). Labor supply is inelastic (with \(\ell^1 = 4\) and \(\ell^2 = 2\)) and perfectly substitutable in production. There is no disutility of labor. The utility function is \(U = x^h\), and the firm produces one unit of \(x\) with one unit of labor. Labor is the numéraire good with its price normalized to 1. The producer price of \(x\) is \(p\). The government can levy individualized commodity tax \(t^h\) on good \(x\). Thus the consumer price facing household \(h\) is \(q^h = p + t^h\). There is no revenue requirement so \(R = t^1 x^1 + t^2 x^2 = 0\).

a. What is the equilibrium producer price?
b. What is the demand for good \(x\) as a function of the tax rate for each household?
c. Use the demand function to express the utility of each household as a function of the price of the consumption good.
d. Show that government budget balance implies that the taxes are related by \(t^2 = -\frac{2t^1}{M^1 - M^2}\).
e. Use the budget balance condition in part d to find the tax rates maximizing the Rawlsian social welfare function \(W = \min\{U^1, U^2\}\).
f. Why individualized commodity taxes are not used in practice?

15.21 Are the following statements true or false?

a. The theory of optimal commodity taxation argues that equal tax rates should be set across all commodities so as to maximize efficiency by “smoothing taxes.”
b. In the United States prescription drugs and CDs are taxed at the same rate of 10 percent. The Ramsey rule suggests that this is the optimal tax policy.
c. Some economists have proposed replacing the income tax with a consumption tax to avoid taxing savings twice. This is a good policy both in terms of efficiency and equity.

15.22 Consider two consumers with preferences

\[
U^1 = \alpha \log(x_1^1) + (1 - \alpha) \log(x_2^1),
\]

\[
U^2 = \beta \log(x_1^2) + (1 - \beta) \log(x_2^2),
\]

and incomes \(M^1 < M^2\). What is the maximum amount of redistribution that can be obtained by levying commodity taxes on goods 1 and 2? Why is it zero if \(\alpha = \beta\)?

15.23 A public sector firm has cost function \(c(x) = F + cx\). It faces the inverse demand curve \(p = a - bx\).

a. What is the first-best price for the firm to charge? What quantity is sold? Does the firm make a loss at this price?
b. What is the break-even price for the firm? What quantity is sold?
c. Compare the difference in consumer surplus between the outcomes in parts a and b. Is there scope for the government to subsidise the firm when it sets the first-best price?
16 Income Taxation

16.1 Introduction

In 1799 an income tax was introduced for the first time in the United Kingdom to pay for the Napoleonic war. The tax was levied at a rate of 10 percent on income above £60 and survived until it was repealed in 1816 following major public opposition. Part of the opposition was due to concerns about privacy, and this was reflected in the decision of Parliament to pulp all documents relating to the income tax. The tax returned in 1842 as a temporary measure (imposed for three years with the possibility of a two-year extension) to cover a major budget deficit. It has remained in place ever since, although it is still temporary and Parliament has to re-apply it every year.

The income tax has remained controversial. As the discussion of chapter 4 showed, the taxation of income is a major source of government revenue. This fact, coupled with the direct observation by taxpayers of income tax payments, explains why the structure of income tax is the subject of much political discussion. The arguments that are aired in such debate reflect the two main perspectives on income taxation. The first views the tax primarily as a disincentive to effort and enterprise. On this ground, it follows that the rate of tax should be kept as low as possible in order to avoid such discouragement. This is essentially the expression of an efficiency argument. The competing perspective is that income taxation is well suited for the task of redistributing income. Hence notions of equity require that high earners should pay proportionately more tax on their incomes than low earners. The determination of the optimal structure of income taxation involves the resolution of these contrasting views.

The chapter begins by conducting an analysis of the interaction between income taxation and labor supply. A number of theoretical results are derived, and these are related to the empirical evidence. This evidence makes clear the extent of the difference between the responses of male and female labor supply to taxation. A model that permits the efficiency and equity aspects of taxation to be incorporated into the design of the optimal tax is then described. A series of results characterizing properties of the optimal tax function are derived using this model, and these properties are interpreted in terms of practical policy recommendations. Calculations of the optimal tax rates that emerge from the model are then reviewed. The chapter is completed by a discussion of political economy aspects of income taxation.
16.2 Equity and Efficiency

There are two major issues involved in the taxation of income. The first is the effect of taxation on the supply of labor. Taxation alters the choices that consumers make by affecting the trade-off between labor and leisure. In this respect a particularly important question is whether an increase in the rate of tax necessarily reduces the supply of labor. If this is the case, support would be provided for the argument that taxes should be kept low to meet the needs of efficiency. Both theoretical and empirical results addressing this question will be discussed. The second issue that has been studied is the determination of the optimal level of income taxation. For reasons that will become clear, this is a complex problem, since it can only be addressed in a model with a meaningful trade-off between efficiency and equity. Having said this, the search for the right trade-off has proved to be a fruitful avenue of investigation.

The essential idea we wish to convey in this chapter is that it is a major mistake to design the income tax structure to meet equity motives without taking into account the impact on work effort. To see why, consider the naïve solution of setting the marginal tax rate at 100 percent for all incomes above some threshold level \( z^\circ \) and at a rate of zero for all incomes below this threshold. We might expect that such a tax structure will maximize the redistribution possible from the rich (those above the income threshold) to the poor (those below). However, this conclusion is incorrect when taxpayers respond to the tax structure. The confiscatory tax above the threshold removes the incentive to earn more than \( z^\circ \), and everyone previously above this level will choose to earn exactly that amount of income. This sets a vicious circle in motion. The government must lower the threshold, inducing everyone above the new level to lower their incomes again, and so forth, until no one chooses to work and income is zero.

It therefore stands to reason that we must analyze the equity of the tax structure in tandem with its effect on work incentives. The idea is to find the tax schedule that meets social objectives, as captured by the social welfare function (see chapter 13), given the adjustment in work effort and labor market participation by taxpayers. Such a tax scheme is said to be optimal conditional on the given objective. The results need to be interpreted with caution, however, because they are very sensitive to the distribution of abilities in the population and to the form of the utility function. More important, they depend on the equity objective as built into the social welfare function.

In this chapter we will only consider welfaristic equity criteria (of which the utilitarian and Rawlsian social welfare functions are noteworthy examples). Hence, insofar as
the social objective is entirely based on individual welfare levels, we are not assessing
the tax structure on the basis of its capacity for either redressing inequality or eliminat-
ing poverty. Neither do we consider egalitarian social objectives like equal sacrifice or
equality of opportunities. There is indeed an interesting literature on “fair” income tax-
ation that examines the distribution of taxes that impose the same loss of utility on
everyone, either in absolute or relative terms. Such arguments are related to the ability-
to-pay principle according to which $1 of tax is less painful for a rich person than for
a poor person (due to the decreasing marginal value of income). This equal sacrifice
approach predicts that the resulting tax structure must be progressive (in the sense that
everyone sacrifices equally if they pay an increasing percentage of their income in tax
as their income rises). It was John Stuart Mill, in his Principles of Political Economy,
who first pointed out this principle of equal sacrifice. He suggested that “Equality of
taxation, therefore as a maxim of politics, means equality of sacrifice. It means apportion-
ing the contribution of each person towards the expenses of government so that he
shall feel neither more nor less inconvenience from his share of the payment than every
other person experiences from his” (bk. V, ch. 2, p. 804).

16.3 Taxation and Labor Supply

The effect of income taxation on labor supply can be investigated using the standard
model of consumer choice. The analysis will begin with the general question of labor
supply and then move on to a series of specific analyses concerning the effect of varia-
tions in the tax system. The major insight this gives will be to highlight the importance
of competing income and substitution effects.

As is standard, it is assumed that the consumer has a given set of preferences over
allocations of consumption and leisure. The consumer also has a fixed stock of time
available that can be divided between labor supply and time spent as leisure. The utility
function representing the preferences can then be defined by

\[ U = U(x, L - \ell) = U(x, \ell), \]

(16.1)

where \( L \) is the total time endowment, \( \ell \) is labor supply, and \( x \) is consumption. Conse-
quently leisure time is \( L - \ell \). Labor is assumed to be unpleasant for the worker, so utility
is reduced as more labor is supplied, implying that \( \frac{\partial U}{\partial \ell} < 0 \). Let each hour of labor earn
the wage rate \( w \) so that income, in the absence of taxation, is \( w\ell \). Letting the (constant)
rate of tax be \( t \), the budget constraint facing the consumer is \( px = [1 - t] w\ell \), where
\( p \) is the price of the consumption good.
The choice problem is shown in panel a of Figure 16.1, which graphs consumption against leisure. The indifference curves and budget constraint are as standard for utility maximization. The optimal choice is at the tangency of the budget constraint and the highest attainable indifference curve. This results in consumption $x^*$ and leisure $L - \ell^*$. There is an alternative way to write the utility function. Let the before-tax income be denoted by $z$, so that $z = w\ell$. Since $\ell = \frac{z}{w}$, utility can then be written in terms of before-tax income as

$$U = U(x, \frac{z}{w}).$$ (16.2)

These preferences can be depicted on a graph of before-tax income against consumption. Expressed in terms of income, the budget constraint becomes $px = [1 - t]z$. This is shown in panel b of Figure 16.1. The optimal choice occurs at the point of tangency between the highest attainable indifference curve and the budget constraint, with consumption $x^*$ and before-tax income $z^*$. The important feature of this alternative representation is that the budget constraint is not affected as $w$ changes, so it is the same whatever wage the consumer earns, but the indifference curves do change, since it is $\frac{z}{w}$ that enters the utility function. How they change is described below.

This standard model can now be used to understand the effects of variations in the wage rate or tax rate. Consider the effect of an increase in the wage rate, which is shown in panel a of Figure 16.2 by the move to the higher budget line and the new
tangency at $c$. The move from $a$ to $c$ can be broken down into a substitution effect ($a$ to $b$) and an income effect ($b$ to $c$). The direction of the substitution effect can always be signed, since it is given by a move around the indifference curve. In contrast, the income effect cannot be signed: it may be positive or negative. Consequently the net effect is ambiguous: an increase in the wage rate can raise or lower labor supply. This is the basic ambiguity that runs throughout the analysis of labor supply. The effect of a wage increase when preferences are written as in (16.2) is shown in panel b of figure 16.2. An increase in the wage rate means that less additional labor is required to achieve any given increase in consumption. This change in the trade-off between labor and consumption causes the indifference curve through a point to pivot round and become flatter. This flattening of the indifference curves causes the optimal choice to move along the budget constraint. The level of before-tax income will rise, but the effect on hours worked is ambiguous.

The effect of a tax increase is now analyzed in the same way. In panel a of figure 16.3 the tax increase rotates the budget line down so that the optimal choice moves from $a$ to $c$. The substitution effect of the tax increase is the move around the indifference curve from $a$ to $b$ and the income effect the move from $b$ to $c$. Using the alternative form of preferences an increase in the tax rate rotates the budget constraint in panel b downward so that the chosen point moves from $a$ to $c$. In neither diagram does the change in tax rate affect the indifference curves.

It is also helpful to consider more complex tax systems using this approach. A common feature of the income tax in many countries is that there is a threshold level of
income below which income is untaxed. This is shown in figures 16.4. The threshold level of income is $z^*$ so at wage rate $w$ this arises at $\frac{z^*}{w}$ hours of labor supply. The economic importance of this threshold is that it puts a kink into the budget constraint. If a set of consumers with differing preferences are considered, some may locate at points such as $a$ and pay no tax, and some may locate at points like $c$. However, it can be expected that a number of consumers will cluster or “bunch” at the kink point $b$. The observation that consumers will bunch at a kink point is a common feature and reflects
the fact that an extra unit of labor will receive net pay \( [1 - t]w \), whereas the previous unit received \( w \). It is therefore helpful to distinguish between interior solutions, such as \( a \) and \( c \), and corner solutions such as \( b \). The consumer at an interior solution will respond to changes in the tax rate in the manner illustrated in figure 16.2. In contrast, a consumer at a corner solution may well be left unaffected by a marginal tax change. The consumer’s choice will only be affected if the change is sufficient to allow the attainment of a utility level higher than at the kink.

More generally, an income tax system may have a number of thresholds with the marginal tax rate rising at each. Such a tax system appears as in figure 16.5. Again, with preferences varying across consumers, the expectation is that there will tend to be collections of consumers at each kink point.

The final issue that is worth investigating in this framework is that of participation in the labor force. The basic assumption so far has been that the worker can continuously vary the number of working hours in order to arrive at the most preferred outcome. In practice, it is often the case that either hours are fixed or else there is a minimum that must be undertaken with the possibility of more. Either case leads to a discontinuity in the budget constraint at the point of minimum hours. The choice for the consumer is then between undertaking no work and working at least the minimum. This is the participation decision: whether or not to join the workforce.

The participation decision and its relation to taxation is shown in figure 16.6 where \( \ell^m \) denotes the minimum working time. The effect of an increase in taxation is to lower the budget constraint. A consumer who was previously indifferent between working
and not (both points being on the same indifference curve) now strictly prefers not to do so. At this margin there is no conflict between income and substitution effects. An increase in taxation strictly reduces participation in the work force.

16.4 Empirical Evidence

The theoretical analysis of section 16.3 has identified the three major issues in the study of labor supply. These are the potential conflict between income and substitution effects that make it impossible to provide any clear-cut results for those consumers at an interior solution, kinks in the budget constraint that make behavior insensitive to taxes, and a participation decision that can be very sensitive to taxation. How important each of these factors is in determining the actual level of labor supply can only be discovered by reference to the empirical evidence.

Empirical evidence on labor supply and the effect of income taxes can be found in both the results of surveys and in econometric estimates of labor supply functions. In considering what evidence is useful, it is best to recall that the labor supply will be insensitive to taxation if working hours are determined by the firm or by a union and firm agreement. When this is the case, only the participation decision is of real interest. The effect of taxation at interior solutions can only be judged when the evidence relates to workers who have the freedom to vary their hours of labor. This is most commonly the position for those in self-employment rather than employment. For those in
employment, variations in hours can sometimes be achieved by undertaking overtime, so this dimension of choice can be considered.

These comments also draw attention to the fact that the nature of labor supply may well be different between males and females, especially married females. It still remains a fact that males continue to be the dominant income earner in most families. This leaves the married female as typically a secondary income earner, and for them there is often no necessity to work. From this position it is the participation decision that is paramount. In contrast, most males consider work to be a necessity, so the participation decision is an irrelevance. It can therefore be expected that the labor supply of males and females will show different degrees of sensitivity to taxation.

Surveys on labor supply have normally arrived at the conclusion that changes in the tax rate have little effect on the labor-supply decision. For instance, a survey of the disincentive effect of high tax rates on solicitors and accountants in the United Kingdom, 63 percent of whom were subject to marginal tax rates above 50 percent, concluded that as many of the respondents were working harder because of the tax rates as were working less hard. Groups such as these are ideal candidates for study for the reasons outlined above: they can be expected to have flexibility in the choice of working hours and should be well informed about the tax system. A similar conclusion was also found in a survey of the effect of income taxation on the level of overtime worked by a sample of weekly paid workers: little net effect of taxation on working hours was found.

These results suggest the conclusion that labor supply does not vary significantly with the tax rate. If this were correct, the labor supply function would be approximately vertical. In terms of the theoretical analysis the survey results point to an income effect that almost entirely offsets the substitution effect. However, the discussion has already suggested that different groups in the population may have different reactions to changes in the tax system. This issue is now investigated by considering evidence from econometric analysis.

Table 16.1 presents some summary econometric estimates of labor-supply elasticities. These are divided into those for married men, married women, and single mothers. Each gives the overall elasticity and its breakdown into substitution and income effects. Estimates for both the United Kingdom and the United States are given.

Since the results in table 16.1 relate to the effect of a wage increase, theory would predict that the substitution effect should be positive. This is what is found in all cases. The income effect, which theoretically can be positive or negative, is found to be always negative. Consequently the negativity offsets the substitution effect, sometimes more
than completely. While there are a range of estimates for each category, some general observations can be made. The estimated elasticity for married men is the lowest and is the only one that is ever estimated to be negative. This implies that the labor-supply curve for married men is close to vertical and may even slope backward. One explanation for this result could be that the working hours of this group are constrained by collective agreements that leave little flexibility for variation.

The labor-supply elasticity of single mothers is, on average, the largest of the three sets. This result is probably a consequence of the participation effect. For single mothers part-time work is an unattractive option, since this usually implies the loss of state benefits. Consequently labor supply becomes an all or nothing decision. Married women represent the intermediate case. For them part-time work is quite common, and this often opens the way to some degree of flexibility in hours of work. As expected, these factors lead to a labor-supply elasticity greater than that of married men but lower than that of single mothers.

Although the estimates vary widely within the groups, indicating some imprecision in the estimates, some general conclusions can still be drawn. First, the elasticity of labor supply is not uniform across the population of workers. It clearly varies among the three groups identified in this discussion and probably varies within these groups. Despite this variation, it is still apparent that the uncompensated labor-supply elasticity for married men is small with estimates grouped around zero. In contrast, the elasticity of women is higher and reflects the participation effect and the greater flexibility they have in the choice of hours.

Table 16.1
Labor-supply elasticities

<table>
<thead>
<tr>
<th></th>
<th>Married women</th>
<th>Married men</th>
<th>Single mothers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>United States</td>
<td>United Kingdom</td>
<td>United States</td>
</tr>
<tr>
<td>Uncompensated wage</td>
<td>0.45</td>
<td>0.43</td>
<td>−0.03</td>
</tr>
<tr>
<td>Compensated wage</td>
<td>0.90</td>
<td>0.65</td>
<td>0.95</td>
</tr>
<tr>
<td>Income</td>
<td>−0.45</td>
<td>−0.22</td>
<td>−0.98</td>
</tr>
</tbody>
</table>

16.5 Optimal Income Taxation

The analysis to this point has considered the positive question of how income taxation affects labor supply. Having understood this, it is now possible to turn to the normative question of how the income tax structure should be determined. This is by nature a complex issue. As has already been noted, in practice, income tax systems generally have a number of thresholds at which the marginal tax rate rises. An investigation of the optimal system must at least be flexible enough to consider such tax systems without limiting the number of thresholds or the rates of tax at each. In fact it must do more much than this. The model of income taxation introduced by Mirrlees (1971) has several important attributes. First, there is an unequal distribution of income, so there are equity motivations for taxation. Second, the income tax affects the labor supply decisions of the consumers so that it has efficiency consequences. Third, in view of the comments above, the structure is sufficiently flexible that no prior restrictions are placed on the optimal tax functions that may arise.

In the model all consumers have identical preferences but differ in their level of skill in employment. The hourly wage received by each consumer is determined by their level of skill. This combines with the labor-supply decision to determine income. The economy is competitive, so the wage rate is also equal to the marginal product of labor and firms price their output at marginal cost. A tax levied on skill would be the first-best policy as it would be a lump-sum tax on the unalterable characteristic that differentiates consumers. But this first-best is not feasible, since the level of skill is assumed to be private information and so unobservable by the government. As the discussion of chapter 13 showed, this makes it impossible to tax skill directly. Since the government cannot observe a consumer’s skill level (which is essentially the initial endowment of the consumer), it employs an income tax as a second-best policy. The income tax function is chosen to maximize social welfare subject to it raising enough revenue to meet the government’s requirements.

There are two commodities: a consumption good and labor. A consumer’s labor supply is denoted by $\ell$ and consumption by $x$. Each consumer is characterized by a skill level $s$. The value of $s$ measures the hourly output of the consumer, and since the economy is competitive, it is equal to the wage rate. If a consumer of ability $s$ supplies $\ell$ hours of labor, that consumer earns income of $s\ell$ before tax. Denote the income of the consumer with skill $s$ by $z(s) = s\ell(s)$. The amount of tax paid on an income $z$ is given by $T(z)$. This is the tax function that the analysis aims to determine. Equivalently,
denote the consumption function by $c(z)$ so that a consumer who earns income $z$ can consume

$$x = c(z) = z - T(z).$$

(16.3)

The relationship among income, the tax function, and consumption is depicted in figure 16.7. In the absence of taxation, income would be equal to consumption and this is depicted by the 45 degree line. Where the consumption function lies above the 45 degree line, the tax payment is negative. It is positive when the consumption function is below the line. For example, the consumer earning $\hat{z}$ in the figure pays an amount of tax $T(\hat{z})$ and can consume $\hat{x}$. The gradient of the consumption function is equal to one minus the marginal rate of tax.

All consumers have the same utility function (so that the possibility of workers displaying different aversion to work is ruled out):

$$U = U(x, \frac{z}{s}).$$

(16.4)

The indifference curves are dependent on the skill level of the consumer, since a high-skill consumer takes less labor time to achieve a given level of income than a low-skill consumer. This is reflected in the fact that at any income and consumption pair $\{\hat{x}, \hat{z}\}$ the indifference curve of a high-skill consumer passing through that point is flatter than the indifference curve of a low-skill consumer. This single-crossing property is termed agent monotonicity and is illustrated in figure 16.8.

---

**Figure 16.7**

Taxation and the consumption function
Chapter 16: Income Taxation

Figure 16.8
Agent monotonicity

Figure 16.9
Income and skill
An immediate consequence of agent monotonicity is that high-skill consumers will never earn less income than low-skill. Generally, they will earn strictly more. This result is shown in figure 16.9. It arises because at the point where the indifference curve of the low-skill consumer is tangential to the consumption function (and so determines the optimal choice for that consumer), the indifference curve of the high-skill consumer is flatter and so cannot be at a tangency. Recall that all consumers face the same tax function and thus the same consumption function no matter what their skills are. The optimal choice for the high-skill consumer therefore has to be to the right of $a$, which implies a higher level of income.

The first property of the optimal tax function relates to the maximum tax rate that will be charged. If the consumption function slopes downward, as shown in figure 16.10, then the shape of the indifference curves ensures that no consumer will choose to locate on the downward-sloping section. This part of the consumption function is therefore redundant and can be replaced by the flat-dashed section without altering any of the consumers’ choices. Economically, along the downward-sloping section, increased work effort is met with lower consumption. Hence there is no incentive to work harder, and such points will not be chosen. Since $c(z) = z - T(z)$, it follows that $c'(z) = 1 - T'(z)$. The argument has shown that $c'(z) \geq 0$, which implies $T'(z) \leq 1$, so the marginal tax rate is less than 100 percent.

It is also possible to put a lower limit on the marginal tax rate. If the gradient of the consumption function is greater than one, meaning $c'(z) > 1$, then $T'(z) < 0$. A
negative tax rate represents a marginal subsidy to the tax payer from the tax system. That is, the after-tax wage for additional work will be greater than the before-tax wage. To demonstrate the logic behind this claim most clearly it is assumed that the social welfare function is utilitarian and that there are only two types of consumers (low-skill and high-skill), and equal numbers of the two types. Begin with the tax function denoted \( c^1(z) \) in figure 16.11, which has gradient greater than one. Along this are located a high-skill consumer at \( h_1 \) and a low-skill consumer at \( l_1 \).

The effect of moving to the new tax function \( c^2(z) \) where the gradient is equal to one is now analyzed in two stages. In the first stage the income levels of the consumers are kept fixed, but the consumption levels change to \( \tilde{x}_l \) and \( \tilde{x}_h \). This is shown in figure 16.12. The position of \( c^2(z) \) is chosen so that the increase in consumption for the low-skill is exactly matched by the decrease in consumption of the high-skill. Because aggregate income and aggregate consumption have not changed, the budget position of the government is unaffected. The change in consumption levels must raise welfare because the marginal utility of consumption for the low-skill is higher than that for the high-skill.

The second stage of the process is to allow the consumers to respond to the tax function \( c^2(z) \) and relocate to utility-maximizing positions. The high-skill consumer will choose \( \{z^*_h, x^*_h\} \) and the low-skill will choose \( \{z^*_l, x^*_l\} \) as shown in figure 16.13.

![Figure 16.11
Initial position and new tax function](image-url)
Figure 16.12
Transfer of consumption

Figure 16.13
Allowing relocation
Being voluntary, the relocation must raise the utilities of both types of consumer. Furthermore the relocations do not affect the budget position of the government, since they are moves along a consumption function with a zero marginal rate of tax.

The fact that social welfare increases in stage 1 and then increases further in stage 2 ensures that consumption function $c^2(z)$ leads to a higher level of social welfare than consumption function $c^1(z)$. The process has also ensured that government revenue is unchanged. Consumption function $c^1(z)$ with a negative marginal rate of tax cannot therefore be optimal. From this it follows that the marginal tax rate must be nonnegative so that $T'(z) \geq 0$. It should be noted that this result does not restrict the average rate of tax, $\frac{T(z)}{z}$, to be nonnegative. The average rate of tax is negative whenever the consumption function is above the 45 degree line, and if the system is redistributive, this will be the case at low incomes.

The final result determines the marginal tax rate faced by the highest skill consumer. Let the consumption function $ABC$ in figure 16.14 be a candidate for optimality. It is now shown that $ABC$ cannot be optimal unless its gradient is one (so the marginal rate of tax is zero) at point $B$ where the highest skill consumer locates. To prove this result, assume that the gradient is less than one (so the marginal tax rate is positive) at point $B$. A better consumption function than $ABC$ will now be constructed. To do this, define $ABD$ by following the old consumption function up to point $B$, and then let the new section $BD$ have gradient of one. The highest ability consumer will now relocate to point $b$. Consequently the highest ability consumer is better off, but their actual tax payment (the vertical distance from the consumption point to the 45 degree line) is

![Figure 16.14](image.png)

**Figure 16.14**

Zero marginal rate of tax
unchanged. So replacing $ABC$ with $ABD$ leaves aggregate tax revenue unchanged, makes one person better off, and makes no one worse off. This must be an improvement for society, so no consumption function, like $ABC$, that has the highest ability person facing a positive marginal rate of tax can be optimal. In other words, the optimal tax function must have a zero marginal rate of tax for the highest skill consumer.

This result is important for assessing the optimality of actual tax schedules. Those observed, in practice, invariably have a marginal rate of tax that rises with income. This leaves the highest income consumers facing the highest marginal tax rate rather than a zero rate. Accordingly such tax systems cannot be optimal. The result also carries implications for discussions about how progressive the income tax system should be. A tax system displays marginal rate progressivity if the marginal rate of tax increases with income. Since it has been shown that the marginal rate of tax should be zero at the top of the income distribution, the optimal tax system cannot be a fully (marginal rate) progressive one.

There has been considerable debate about this result due partly to its contrast with what is observed in practice. There are several points that can be made in this respect. The result is valid only for the highest skill consumer, and it makes no prediction about the tax rate that will be faced by consumer with the second-highest skill. Therefore it does not demonstrate that those close to the top of the skill range should face a tax rate of zero or even close to zero. For them the tax rate may have to be significantly different from zero. If this is the case, observed tax systems may only be “wrong” at the very top, which will not result in too great a divergence from optimality. The result also relies on the fact that the highest skill person can be identified and the tax system adjusted around her needs. Putting this into practice is clearly an impossible task. In summary, the result is important in that it questions preconceptions about the structure of taxes, but it has limited immediate policy relevance.

The results described in this section capture the general properties of an optimal income tax system that can be derived within this framework. They have shown that the marginal tax rate should be between zero and one and that the highest skill consumer should face a zero marginal rate. Moreover they have established that the tax system should not involve marginal rate progressivity everywhere. It is possible to derive further results only by adding further specification. The next section looks at two special cases that give alternative routes for proceeding in this direction. However, even these do not provide entirely transparent insights into the level of optimal tax rates. This can only be done through the use of numerical results, and these are the subject matter of section 16.7.
16.6 Two Specializations

To provide some more insight into the optimal income tax, two specializations of the model are worth considering. These are noteworthy for the very clear view they give into the trade-offs involved in setting the tax rate.

16.6.1 Quasi-Linearity

The first specialization is to consider a special form of the utility function. It is assumed in this section that utility is quasi-linear with respect to labor income,

$$U(x, \frac{s}{s}) = u(x) - \frac{s}{s},$$  \hspace{1cm} (16.5)

so that the marginal disutility of labor $\ell = \frac{s}{s}$ is constant. The utility of consumption, $u(x)$, is increasing and concave (so $u' > 0$ and $u'' < 0$). For this utility function the marginal rate of substitution between consumption and income is $MRS_{x,z} = \frac{1}{su'(x)}$.

Since the marginal rate of substitution is decreasing in $s$, the gradient of the indifference curve through any value of $x$ falls as $s$ rises. This makes the utility function consistent with agent monotonicity.

We simplify further by assuming that there are just two consumers, one with a high level of skill, $s_h$, and the other with a low level, $s_l$. It is assumed that $s_l < s_h < 3s_l$ (the reason for this is explained later). With only two consumers the problem of choosing the optimal tax (or consumption) function can be given the following formulation: whatever consumption function is selected by the government, the fact that there are only two consumers ensures that at most two locations on it will ever be chosen. For example, in figure 16.15, $a_l$ is the allocation chosen by the low-skill consumer and $a_h$ the chosen allocation of the high-skill. Having observed this, it is apparent that selecting the consumption function is equivalent to specifying the two allocations. The rest of the consumption function can then be chosen to ensure that no point on it is better for the consumers than the two chosen allocations. Essentially the consumption function just needs to link the two points, while elsewhere remaining below the indifference curves through the points. Following this reasoning reduces the choice of tax function to a simple maximization problem involving the two locations.

A consumer will only choose the allocation intended for him if he prefers his own location to that of the other consumer. In other words, the allocations must be incentive compatible. Since the high-skill consumer can mimic the low-skill, but not vice versa,
the incentive compatibility constraint must be binding on the high-skill consumer. Denoting the location intended for low-skill consumer by \( \{x_l, z_l\} \) and that for high-skill by \( \{x_h, z_h\} \), the incentive compatibility constraint is

\[
\frac{u(x_h)}{s_h} - z_h = \frac{u(x_l)}{s_l} - z_l. \tag{16.6}
\]

A pair of allocations that satisfy the incentive compatibility constraint (16.6) are shown in figure 16.16. The high-skill consumer is indifferent between the two allocations (\( a_l \) and \( a_h \) are on the same indifference curve for the high-skill) while the low-skill consumer strictly prefers allocation \( a_l \).

The optimization facing a government that maximizes a utilitarian social welfare function is

\[
\max_{\{x_l, x_h, z_l, z_h\}} \left( u(x_l) - \frac{z_l}{s_l} + u(x_h) - \frac{z_h}{s_h} \right) \tag{16.7}
\]

subject to the incentive compatibility constraint (16.6) and the resource constraint

\[
x_l + x_h = z_l + z_h. \tag{16.8}
\]

The resource constraint makes the simplifying assumption that no revenue is to be raised so that the tax system is purely redistributive. What is now shown is that the quasi-linearity of utility allows this maximization problem to be considerably simplified. The simplification then permits an explicit solution to be given.
Given that (16.6) is an equality, it can be solved to write

\[ z_h = s_h[u(x_h) - u(x_l)] + z_l. \]  

(16.9)

Combining this equation with the resource constraint and eliminating \( z_h \) by using (16.9), the income of the low-skill consumer can be written

\[ z_l = \frac{1}{2} [x_l + x_h - s_h[u(x_h) - u(x_l)]]. \]  

(16.10)

Using the resource constraint again gives the income of the high-skill consumer as

\[ z_h = \frac{1}{2} [x_l + x_h + s_l[u(x_h) - u(x_l)]]. \]  

(16.11)

These solutions for the income levels can then be substituted into the objective function (16.7). Collecting terms shows that the original constrained optimization is equivalent to

\[ \max_{\{x_l, x_h\}} \beta_l u(x_l) + \beta_h u(x_h) - \left[ \frac{s_l + s_h}{2s_l s_h} \right] [x_l + x_h], \]  

(16.12)

where \( \beta_l = \frac{3s_l - s_h}{2s_l} \) and \( \beta_h = \frac{s_l + s_h}{2s_l} \). (The assumption \( s_h < 3s_l \) ensures that \( \beta_l \) is greater than zero so that the low-skill consumer has a positive social weight. Without this assumption the analysis becomes more complex.)

Comparing (16.7) and (16.12) allows a new interpretation of the optimal tax problem. The construction undertaken has turned the maximization of the utilitarian social
welfare function subject to constraint into the maximization of a weighted welfare function without constraint. The incentive compatibility and resource constraints have been incorporated by placing a greater weight on the welfare of the high-skill consumer (since $\beta_h > \beta_l$), which in turn ensures that their consumption level must be higher at the optimum. From (16.10) and (16.11) this feeds back into a higher level of income for the high-skill consumer. It can also be seen that as the skill difference between the two consumers increases, so does the relative weight given to the high-skill.

Carrying out the optimization in (16.12), the consumption levels of the consumers satisfy the first-order conditions

$$\beta_i u'(x_i) - \frac{s_l + s_h}{2s_l s_h} = 0, \quad i = l, h, \quad (16.13)$$

so the consumption levels are proportional to the welfare weights. For the high-skill consumer, substituting in the value of $\beta_h$ gives

$$u'(x_h) = \frac{1}{s_h}. \quad (16.14)$$

Consequently the marginal utility of the high-skill consumer is inversely proportional to their skill level. With $u'' < 0$ (decreasing marginal utility) this implies that consumption is proportional to skill. Combining this result with the fact that $MRS_{x,z}^h = \frac{1}{u'(x)}$, it follows that at the optimum allocation $MRS_{x,z}^h = 1$. The finding that the marginal rate of substitution is unity shows that the high-skill consumer faces a zero marginal tax rate. This is the no-distortion-at-the-top result we have already seen. For the low-skill consumer

$$u'(x_l) = \frac{s_l + s_h}{s_h [3s_l - s_h]}, \quad (16.15)$$

and $MRS_{x,z}^l = \frac{s_h [3s_l - s_h]}{s_l [5s_l + 3s_h]} < 1$. These show that consumer $l$ faces a positive marginal rate of tax.

The use of quasi-linear utility allows the construction of an explicit solution to the optimal income tax problem, which shows how the general findings of the previous section translate into this special case. It is interesting to note the simple dependence of consumption levels on the relative skills and the manner in which the constraints become translated into a higher effective welfare weight for the high-skill consumer. This shows that this consumer needs to be encouraged to supply more labor through the reward of additional consumption.
The second specialization restricts the form of the social welfare function. In chapter 13 we introduced the Rawlsian social welfare function, which represents a society that is concerned only with the utility of the worst-off individual. The worst-off person is typically at the bottom of the income distribution and his welfare depends on the extent of redistribution. We now assume that tax revenue is entirely redistributed in the form of lump-sum grants. Consequently, for a Rawlsian government, the optimal income tax is simply that which maximizes the lump-sum grant or, equally, that which maximizes the revenue extracted from taxpayers.

Given a tax schedule $T(z)$, a consumer of skill level $s$ makes the choice of income, $z$ (which is equivalent to choosing labor supply $\ell = \frac{z}{s}$) and consumption $x$ to maximize his utility subject to satisfying the budget constraint $x = z - T(z)$. Let $z = z(s)$ denote the optimal income choice of type $s$ (conditional on the tax function $T$). It has been seen that agent monotonicity implies that high-skill consumers never earn less income than the low-skill. So $z(s)$ is increasing in $s$ and can be inverted to give the increasing inverse function $s = z^{-1}(z)$ that represents the skill level $s$ associated to each income choice. Different tax schemes will induce different relationships between skill and income from the same underlying distribution of skills.

Assume that skill levels are continuously distributed in the population according to a cumulative distribution function $F(s)$ (indicating the proportion of the population below any skill level $s$) with associated density function $f(s) > 0$ (representing the probability associated with a small interval of the continuous skill). The tax scheme $T(z)$ induces the income distribution $G(z) = F(z^{-1}(z))$ with density $g(z) = f(z^{-1}(z))$.

Now we are in a position to derive the optimal income tax associated with a Rawlsian social welfare function following a simple method originally proposed by Piketty (1997). The Rawlsian optimal tax structure maximizes tax revenue, so no alternative tax structure must exist that can raise more revenue from the taxpayers given their optimal labor supply response to that new tax structure. From the first-order condition of the revenue maximization problem, a small deviation from the optimal tax scheme must have no effect on total tax revenue (and larger deviations must lower tax revenue). It follows that a small change of the tax rate at any given income level $z$ must not change total revenue. Using this simple argument, we can derive the optimal tax structure.

Take income level $z$, and consider a small increase in the marginal tax rate at that point of amount $\Delta T'$. This change has two effects on tax revenue. First, holding labor supply constant, it will increase the tax payment by amount $z\Delta T'$ for all those taxpayers with an income above or equal to the level, $z$, at which the higher marginal tax rate
applies. These taxpayers represent a proportion \(1 - G(z)\) of the population. Therefore the revenue gain from this marginal tax change is
\[
[1 - G(z)] z \Delta T'.
\] (16.16)

Obviously the labor supply is not fixed, and it is expected to vary in response to a change in the tax rate. Let \(\varepsilon_s\) denote the elasticity of labor supply with respect to the net price of labor (the percentage change in labor supply in response to a 1 percent reduction in the net price of labor). With perfect competition on the labor market, the price of labor decreases by the amount of the tax rate (i.e., there is no shifting of the tax burden to employers in the form of a higher gross wage). Now the marginal tax rate increase \(\Delta T'\) at income level \(z\) induces a proportional reduction \(\Delta T'\) in the price of labor. Those facing this marginal tax rate change will reduce their labor supply by
\[
\varepsilon_s \frac{\Delta T'}{1 - T'},
\]
inducing a reduction of their taxable income equal to
\[
g(z) \varepsilon_s \frac{\Delta T'}{1 - T'}.
\] (16.17)

The revenue-maximizing tax scheme (Rawlsian tax) is found by equating the revenue loss to the revenue gain from a marginal tax change for every income level. This yields
\[
[1 - G(z)] z \Delta T' = [g(z)] z \varepsilon_s \frac{\Delta T'}{1 - T'},
\] (16.18)
and the Rawlsian tax structure is easily seen to be such that for all income level \(z\),
\[
\frac{T'(z)}{1 - T'(z)} = \frac{1 - G(z)}{\varepsilon_s g(z)}.
\] (16.19)

This expression has the following interpretation. High marginal tax rates over some middle-income interval \([z, z + dz]\) mean that for those middle-income individuals, but also for the upper-income individuals, the government is collecting more taxes. Altogether, they represent a proportion \(1 - G(z)\) that is decreasing with \(z\) and converging to zero for the highest income level (hence the zero marginal tax rate at the top). The cost of the high marginal tax rate over this interval is a greater distortion for those with income in the range \([z, z + dz]\). The total distortion (and revenue loss) will be low, however, if there are relatively few taxpayers in this interval (low \(g(z)\)), or if those in it have a relatively low labor supply elasticity \(\varepsilon_s\).
Interestingly, even though the redistributive motive is maximal under the Rawlsian social welfare function, the optimal tax structure does not require marginal rate progressivity. Indeed we do not really know how the labor supply elasticity changes with income. Suppose that it is constant. Next take the Pareto distribution of income, which is supposed to be a good fit to the empirical distribution of income. For the Pareto distribution, the hazard rate \( \frac{g(z)}{1 - G(z)} \) is increasing almost everywhere. Therefore, from the optimal tax structure given above, it follows that the marginal tax rate must decrease everywhere. Maximal redistribution is better achieved when the tax schedule is regressive (concave) instead of progressive (convex).

### 16.7 Numerical Results

The standard analysis of optimal income taxation was introduced above, and a number of results were derived that provide some characterization of the shape of the tax schedule. The marginal rate was seen to be between zero and one, but as yet no idea was developed, except for the upper end point, of how close it should be to either. Similarly, although equity considerations are expected to raise the marginal rate, this was not demonstrated formally nor was consideration given to how efficiency criteria, particularly the effect of taxation on labor supply, affects the choice of tax schedule. Because of the analytical complexity of the model, these questions are best addressed via numerical analysis.

The numerical results (from Mirrlees 1971) are based on a social welfare function that takes the form

\[
W = \begin{cases} 
\int_0^S s e^{-\xi U f(s)} ds, & \xi > 0, \\
\int_0^S U f(s) ds, & \xi = 0,
\end{cases}
\]

where \( S \) is the maximum level of skill in the population and 0 the lowest, and \( f(s) \) is the density function for the skill distribution. The form of this social welfare function permits variations in the degree of concern for equity by changes in \( \xi \). Higher values of \( \xi \) represent greater concern for equity, with \( \xi = 0 \) representing the utilitarian case. (This is an alternative specification to that of equation 14.24). The individual utility function is the Cobb–Douglas form

\[
U = \log(x) + \log(1 - \ell),
\]

The skill distribution is lognormal with a standard deviation of 0.39. This value of the standard deviation corresponds approximately to a typical value for the income
distribution. If the skill distribution matches the income distribution, then this is a value of particular interest.

A selection of the numerical results obtained from this model are given in tables 16.2 and 16.3. In table 16.2, \(\varepsilon = 0\), so this is the case of a utilitarian social welfare function. Table 16.3 introduces equity considerations by using \(\varepsilon = 1\). In both cases the government revenue requirement is set at 10 percent of national income.

The first fact to be noticed from these results is that the average rate of tax for low-skill consumers is negative. These consumers are receiving an income supplement from the government. This is in the nature of a negative income tax where incomes below a chosen cutoff are supplemented by the government through the tax system. The average rate of tax then increases with skill. The maximum average rate of tax is actually quite small. The value of 34 percent in table 16.3 is not far out of line with the actual rate in many countries.

<table>
<thead>
<tr>
<th>Income</th>
<th>Consumption</th>
<th>Average tax (%)</th>
<th>Marginal tax (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.03</td>
<td>—</td>
<td>23</td>
</tr>
<tr>
<td>0.055</td>
<td>0.07</td>
<td>−34</td>
<td>26</td>
</tr>
<tr>
<td>0.10</td>
<td>0.10</td>
<td>−5</td>
<td>24</td>
</tr>
<tr>
<td>0.20</td>
<td>0.18</td>
<td>9</td>
<td>21</td>
</tr>
<tr>
<td>0.30</td>
<td>0.26</td>
<td>13</td>
<td>19</td>
</tr>
<tr>
<td>0.40</td>
<td>0.34</td>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td>0.50</td>
<td>0.43</td>
<td>15</td>
<td>16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Income</th>
<th>Consumption</th>
<th>Average tax (%)</th>
<th>Marginal tax (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.05</td>
<td>—</td>
<td>30</td>
</tr>
<tr>
<td>0.05</td>
<td>0.08</td>
<td>−66</td>
<td>34</td>
</tr>
<tr>
<td>0.10</td>
<td>0.12</td>
<td>−34</td>
<td>32</td>
</tr>
<tr>
<td>0.20</td>
<td>0.19</td>
<td>7</td>
<td>28</td>
</tr>
<tr>
<td>0.30</td>
<td>0.26</td>
<td>13</td>
<td>25</td>
</tr>
<tr>
<td>0.40</td>
<td>0.34</td>
<td>16</td>
<td>22</td>
</tr>
<tr>
<td>0.50</td>
<td>0.41</td>
<td>17</td>
<td>20</td>
</tr>
</tbody>
</table>
The behavior of the marginal rate of tax is rather different from that of the average rate. It first rises and then falls. The maximum rate is reached around the median of the skill distribution. Except at the extremes of the skill distribution, there is not much variation in the marginal tax rate. To a first approximation, the optimal tax systems reported in these tables have a basically constant marginal rate of tax so that the consumption function is close to being a straight line. This is one of the most surprising conclusions of the analysis of income taxation: the model allows nonconstancy in the marginal tax rate, but this does not feature to a great degree in the optimal solution. Finally the zero tax rate for the highest skill consumer is reflected in the fall of the marginal rate at high skills, but this is not really significant until close to the top of the skill distribution.

These results provide an interesting picture of the optimal income tax function. They suggest that the tax function should subsidize low-skill consumers through a negative income tax but should still face them with a high marginal rate of tax. The maximum marginal rate of tax should not be at the top of the skill distribution but should occur much lower. Generally, the marginal rate should be fairly constant.

16.8 Voting over a Flat Tax

Having identified the properties of the optimal tax structure, we now consider the tax system that emerges from the political process. To do this, we consider people voting over tax schedules that have some degree of redistribution. Because it is difficult to model voting on nonlinear tax schemes given the high dimensionality of the problem, we will restrict attention to a linear tax structure as originally proposed by Romer (1975). We specify the model further with quasi-linear preferences to avoid unnecessary complications and to simplify the analysis of the voting equilibrium.

Assume, as before, that individuals differ only in their level of skill. We assume that skills are distributed in the population according to a cumulative distribution function \( F(s) \) that is known to everyone, with mean skill \( \bar{s} \) and median \( s_m \). Individuals work and consume. They also vote on a linear tax scheme that pays a lump-sum benefit \( b \) to each individual financed by a proportional income tax at rate \( t \). The individual utility function has the quasi-linear form

\[
u(x, z) = x - \frac{1}{2} \left[ \frac{z}{s} \right]^2,
\]

and the individual budget constraint is

\[
x = (1 - t)z + b.
\]
It is easy to verify that in this simple model the optimal income choice of a consumer with skill level \( s \) is
\[
z(s) = [1 - t]s^2. \tag{16.24}
\]
The quasi-linear preferences imply that there is no income effect on the labor supply (i.e., \( z(s) \) is independent of the lump-sum benefit \( b \)). This simplifies the expression of the tax distortion and makes the analysis of the voting equilibrium easier. Less surprisingly, a higher tax rate induces taxpayers to work less and earn less income.

The lump-sum transfer \( b \) is constrained by the government budget balance condition
\[
b = tE(z(s)) = t [1 - t] E(s^2), \tag{16.25}
\]
where \( E(\cdot) \) is the mathematical expectation, and we used the optimal income choice to derive the second equality. This constraint says that the lump-sum benefit paid to each individual must be equal to the expected tax payment \( tE(z(s)) \). This expression is termed the Dupuit–Laffer curve and describes tax revenue as a function of the tax rate. In this simple model the Dupuit–Laffer curve is bell-shaped with a peak at \( t = \frac{1}{2} \) and no tax collected at the ends \( t = 0 \) and \( t = 1 \). We can now derive individual preferences over tax schedules by substituting (16.23) and (16.24) into (16.22). After re-arrangement, (indirect) utility can be written as
\[
v(t, b, s) = b + \frac{1}{2} [1 - t]^2 s^2. \tag{16.26}
\]
Taking the total differential of (16.26) gives
\[
dv = db - [1 - t]s^2 dt, \tag{16.27}
\]
so that along an indifference curve where \( dv = 0 \),
\[
\frac{db}{dt} = [1 - t]s^2. \tag{16.28}
\]

It can be seen from this that for given \( t \), the indifference curve becomes steeper in \((t, b)\) space as \( s \) increases. This monotonicity is a consequence of the single-crossing property of the indifference curves. As we saw in chapter 11, the single-crossing property is a sufficient condition for the Median Voter Theorem to apply. It follows that there is only one tax policy that can result from majority voting: it is the policy preferred by the median voter (half of the voters are poorer than the median and prefer higher tax rates, and the other half are richer and prefer lower tax rates). Letting \( t_m \) be the
tax rate preferred by the median voter, we have $t_m$ implicitly defined by the solution to the first-order condition for maximizing the median voter’s utility. We differentiate (16.26) with respect to $t$, taking into account the government budget constraint (16.25) to obtain

$$\frac{\partial v}{\partial t} = [1 - 2t] E(s^2) - [1 - t] s^2. \quad (16.29)$$

Setting this expression equal to zero for the median skill level $s_m$ yields the tax rate preferred by the median voter

$$t_m = \frac{E(s^2) - s_m^2}{2E(s^2) - s_m^2}, \quad (16.30)$$

or, using the optimal income choice (16.24),

$$t_m = \frac{E(z) - z_m}{2E(z) - z_m}. \quad (16.31)$$

This simple model predicts that the political equilibrium tax rate is determined by the position of the median voter in the income distribution. The greater the income inequality, as measured by the distance between median and mean income, the higher is the tax rate. If the median voter is relatively badly off, with income well below the mean income, then equilibrium redistribution is large. In practice, the income distribution has a median income below the mean income, so a majority of voters would favor redistribution through proportional income taxation. More general utility functions would also predict that the extent of this redistribution decreases with the elasticity of labor supply.

16.9 Conclusions

This chapter introduced the issues surrounding the design of the income tax structure. It was first shown how income and substitution effects left the theoretical impact of a tax increase on labor supply indeterminate. If the income effect is sufficiently strong, it is possible for a tax increase to lead to more labor being supplied. The participation decision was also discussed, and it was argued that taxation could be significant in affecting this choice.

The lack of theoretical predictions places great emphasis on empirical research for determining the actual effects of taxation. The labor-supply responses of different
groups in the population to tax changes were discussed. The observations made were borne out by the empirical results that showed a very small or negative elasticity of supply for married men but a much large positive elasticity for single mothers. The latter can be interpreted as a reflection of the participation decision.

A model that was able to incorporate the important issues of efficiency and equity in income taxation was then introduced. A number of results were derived that capture the general features of an optimal tax system. Notably, the marginal rate of tax facing the highest skill person should be zero and the optimal tax rate is bounded between zero and one. This model was specialized to quasi-linear utility and to a Rawlsian social welfare function, and some further insights were obtained. The numerical simulation results showed the marginal rate of tax to remain fairly constant while the average rate of tax was negative for low-skill consumers. Finally the political economy of taxation was presented by means of a simple model of voting over linear income tax schedules.

Further Reading

For a comprehensive survey of recent income tax policy in the United States see:


The economics of taxation and labor supply are surveyed in:


The initial analysis of the problem of nonlinear income taxation was given in:


Be warned, the analytical parts of this paper are exceptionally complex. Even so, the numerical simulation is easily understood.

The case of Rawlsian taxation is analyzed in:


Further numerical simulations are discussed in:


The zero marginal tax rate at the top was first presented in:

The properties of the quasi-linear model are explored in:


An alternative form of quasi-linearity is used to discuss potential patterns of marginal tax rates in:


Further analysis of the pattern of marginal rates of tax is given in:


Voting over linear income tax was originally developed in:


Voting over a nonlinear income tax schedule is in:


A comprehensive and more advanced presentation of the optimal taxation theory is:


A bargaining approach to the income tax problem is in:


For the relationship between existing income tax systems and the equal sacrifice principle, see:


---

**Exercises**

16.1 Consider the budget constraint \( x = b + (1 - t)w\ell \). Provide an interpretation of \( b \). How does the average rate of tax change with income? Let utility be given by \( U = x - \ell^2 \). How is the choice of \( \ell \) affected by increases in \( b \) and \( t \)? Explain these effects.

16.2 Assume that a consumer has preferences over consumption and leisure described by \( U = x \left[1 - \ell\right] \), where \( x \) is consumption and \( \ell \) is labor. For a given wage rate \( w \), which leads to a higher labor supply: an income tax at constant rate \( t \) or a lump-sum tax \( T \) that raises the same revenue as the income tax?
16.3 Let the utility function be \( U = \log(x) - \ell \). Find the level of labor supply if the wage rate, \( w \), is equal to $10. What is the effect of the introduction of an overtime premium that raises \( w \) to $12 for hours in excess of that worked at the wage of $10?

16.4 An individual has to choose her division of time, \( L \), between labor, \( \ell \), and leisure, \( L - \ell \). Her hourly wage is \( w \).
   a. Use a diagram to show how optimal choice without taxation.
   b. Show how the choice changes when an income tax at rate \( t \) is introduced. Identify the income and substitution effects caused by the income tax. What is the total effect on labor supply?
   c. Express the revenue raised by the income tax in terms of units of consumption.
   d. What happens to consumption and the labor supply if a lump-sum tax is introduced instead of the income tax? Assume that the lump sum tax raises the same revenue (in units of consumption) as the income tax.
   e. Use parts c and d to show the excess of burden of the income tax.

16.5 Assume that utility is \( U = \log(x) + \log(1 - \ell) \). Calculate the labor-supply function. Explain the form of this function by calculating the income and substitution effects of a wage increase.

16.6 (Stern 1976) The utility function of a consumer has the constant elasticity of substitution form
\[
U = \left[ \alpha (T - \ell)^{-\mu} + (1 - \alpha) x^{1/\mu} \right]^{-1/\mu}.
\]
Let the budget constraint be \( x = b + w\ell \), where \( b \geq 0 \) is a lump-sum grant received from the government.
   a. Show that the first-order condition for utility maximization can be written as
\[
\left( \frac{b - \mu w\ell}{b + w\ell} \right)^{\mu+1} = \frac{\alpha}{(1-\alpha)\ell}.
\]
   b. Totally differentiate the first-order condition to find \( \frac{d\ell}{dw} \). Under what conditions is this negative?

16.7 Show that a tax function is average-rate progressive (the average rate of tax rises with income) if \( MRT > ART \).

16.8 Which is better: a uniform tax on consumption or a uniform tax on income?

16.9 Consider the utility function \( U = x - \ell^2 \).
   a. For \( U = 10 \), plot the indifference curve with \( \ell \) on the horizontal axis and \( x \) on the vertical axis.
   b. Now define \( z = s\ell \). For \( s = 0.5 \), 1, and 2 plot the indifference curves for \( U = 10 \) with \( z \) on the horizontal axis and \( x \) on the vertical.
   c. Plot the indifference curves for \( s = 0.5 \), 1, and 2 through the point \( x = 20, z = 2 \).
   d. Prove that at any point \( (x, z) \) the indifference curve of a high-skill consumer is flatter than that of a low-skill.

16.10 Consider an economy with two consumers who have skill levels \( s_1 = 1 \) and \( s_2 = 2 \) and utility function \( U = 10x - \ell^2 \). Let the government employ an income tax function that leads to the
allocation \( x = 4, z = 5 \) for the consumer of skill \( s = 1 \) and \( x = 9, z = 8 \) for the consumer of skill \( s = 2 \).

a. Show that this allocation satisfies the incentive compatibility constraint that each consumer must prefer his allocation to that of the other.

b. Keeping incomes fixed, consider a transfer of 0.01 units of consumption from the high-skill to the low-skill consumer.

i. Calculate the effect on each consumer’s utility.

ii. Show that the sum of utilities increases.

iii. Show that the incentive compatibility constraint is still satisfied.

iv. Use parts i through iii to prove that the initial allocation is not optimal for a utilitarian social welfare function.

16.11 For the utility function \( U = x - \ell^2 \) and two consumers of skill levels \( s_1 \) and \( s_2, s_2 > s_1 \), show that the incentive compatibility constraints imply that the income and consumption levels of the high-skill consumer cannot be lower than those of the low-skill consumer.

16.12 Assume that skill is uniformly distributed between 0 and 1 and that total population size is normalized at 1. If utility is given by \( U = \log(x) + \log(1 - \ell) \) and the budget constraint is \( x = b + (1 - t)\ell \), find the optimal values of \( b \) and \( t \) when zero revenue is to be raised. Is the optimal tax system progressive?

16.13 Consider the tax function \( T(z) = az + bz^2 - c \), with consumption given by \( x = z - az + bz^2 + c \).

a. For a consumer of skill \( s \), find the level of income that maximizes \( U = x - \ell^2 \). (Hint: Substitute for \( x \) using the above and for \( \ell \) using \( z = s\ell \).)

b. Hence calculate labor supply and describe its relation to \( s \).

c. Show that at income level \( z \), the marginal rate of tax is \( a + 2bz \).

d. Substitute the answer from part b into the expression for the marginal rate of tax and calculate the limiting marginal tax rate as \( z \to \infty \).

e. Use the answer to part d to show that the tax function is not optimal.

16.14 Consider an economy with two consumers of skill levels \( s_1 \) and \( s_2, s_2 > s_1 \). Denote the allocation to the low-skill consumer by \( \{x^1, z^1\} \) and that to the high-skill consumer by \( \{x^2, z^2\} \).

a. For the utility function \( U = u(x) - \frac{z}{x} \) show that incentive compatibility requires that \( z^2 = z^1 + \left[ u(x^2) - u(x^1) \right] \).

b. For the utilitarian social welfare function

\[
W = u(x^1) - \frac{z^1}{x^1} + u(x^2) - \frac{z^2}{x^2}
\]

express \( W \) as a function of \( x^1 \) and \( x^2 \) alone.

c. Assuming \( u(x^h) = \log x^h \), derive the optimal values of \( x^1 \) and \( x^2 \) and hence of \( z^1 \) and \( z^2 \).

d. Calculate the marginal rate of substitution for the two consumers at the optimal allocation. Comment on your results.
How is the analysis of section 16.6.1 modified if \( s_2 > 3s_1 \)? (Hint: Think about what must happen to the before-tax income of consumer 1.)

Suppose two types of consumers with skill levels 10 and 20. There is an equal number of consumers of both types. If the social welfare function is utilitarian and no revenue is to be raised, find the optimal allocation under a nonlinear income tax for the utility function \( U = \log(x) - \ell \). Contrast this to the optimal allocation if skill is publicly observable.

Tax revenue is given by \( R(t) = tB(t) \), where \( t \in [0, 1] \) is the tax rate and \( t \in [0, 1] \) is the tax base. Suppose that the tax elasticity of the tax base is \( \varepsilon = -\frac{\gamma}{1+\gamma} \) with \( \gamma \in [\frac{1}{2}, 1] \).

a. What is the revenue-maximizing tax rate?

b. Graph tax revenue as a function of the tax rate both for \( \gamma = \frac{1}{2} \) and \( \gamma = 1 \). Discuss the implications of this Dupuit–Laffer curve.

Consider an economy populated by a large number of workers with utility function \( U = x^\alpha[1 - \ell]^{1-\alpha} \), where \( x \) is disposable income, \( \ell \) is the fraction of time worked \((0 \leq \ell \leq 1)\), and \( \alpha \) is a preference parameter \((0 < \alpha < 1)\). Each worker’s disposable income depends on his fixed “skill” as represented by wage \( w \) and a tax-transfer scheme \((t, B)\) so that \( x = B + [1-t]w\ell \), where \( t \in (0, 1) \) is the marginal tax rate and \( B > 0 \) is the unconditional benefit payment.

a. Find the optimal labor supply for someone with ability \( w \). Will the high-skill person work more than the low-skill person? Will the high-skill person have higher disposable income than the low-skill person? Show that the condition for job market participation is \( w > \frac{1-\alpha}{1+\gamma} B \).

b. If tax proceeds are only used to finance the benefit \( B \), what is the government’s budget constraint?

c. Suppose that the mean skill in the population is \( \overline{w} \) and that the lowest skill is a fraction \( \gamma < 1 \) of the mean skill. If the government wants to redistribute all tax proceeds to finance the cash benefit \( B \), what condition should the tax-transfer scheme \((t, B)\) satisfy?

d. Find the optimal tax rate if the government seeks to maximize the disposable income of the lowest skill worker subject to everyone working.

Consider an economy where each consumer has one of two skill levels. The low skill level is \( s_1 \) and the high skill level is \( s_2 \). There are \( N \) consumers in total and \( n_i \) have skill level \( i \). Each consumer has preferences given by \( U = x - \ell^2 \).

a. State the incentive compatibility constraints for the economy. Which will bind at the optimum?

b. Assuming that no revenue is to be collected, write down the optimization describing the optimal allocation for a utilitarian social welfare function.

c. Show that the incentive compatibility constraint and the production constraint can be solved to express \( x_1 \) and \( x_2 \) in terms of incomes and skills.

d. Use these solutions to obtained a reduced form for the optimization problem and solve.

e. Discuss the effect of changing the relative population proportions.

Under the negative income tax all individuals are guaranteed a minimum standard of living by being awarded a grant, and the grant is reduced as their earnings rise (though by less than one for one). Alternatively, under wage subsidies, for each dollar of earnings up to some
level, the government pays each person a refundable tax credit for each dollar earned up to that level. This tax credit is then phased out after reaching a maximum, so the credit goes to zero for middle-income taxpayers.

a. Compare the work incentives of the wage subsidy and the negative income tax for the entire income distribution. Use a diagram and explain.

b. Assume that the poverty line is fixed at $20,000. Design a negative income tax to combat poverty by choosing a basic grant level and an implicit tax rate at which this grant is reduced as incomes rise. What are the trade-offs involved in setting the grant level and tax rate? What are the efficiency and equity effects of choosing different grant levels and tax rates? How will the program affect people with different incomes?

c. Now consider the possibility of using categorical welfare grants. Under categorical welfare grants all individuals possessing certain characteristics are guaranteed a minimum standard of living, and the grant is taken away one for one as income rises. How should the government choose the right categories for targeting grants to some welfare groups?

d. What are the advantages and disadvantages of categorical grants relative to a negative income tax?

16.21 Consider a single mother with the utility function 

\[ U(x, \ell) = \begin{cases} 
\frac{2}{3} \log(x) + \frac{1}{3} \log(\ell), & x \geq \bar{x}, \\
-\infty, & \text{otherwise} 
\end{cases} \]

where \( x \) is consumption and \( \ell \) is leisure. The mother can work up to 100 hours per month. Any of the 100 hours that are not worked are leisure hours. She earns a wage of $10 per hour and pays no taxes. The consumption price is normalized to $1. To be able to work, she has to incur a child care cost of $5 for every hour worked.

a. Suppose that there are no tax and welfare benefits. How many hours will she work and what will be her consumption level? Draw the graph depicting her budget set, with consumption on the vertical axis and leisure on the horizontal axis.

b. Suppose that the government introduces a negative income tax (NIT) that guarantees an income of $200 per month. The benefit is taken away one for one as earnings increase. Draw the new budget set. Compute the new number of hours worked and consumption level. Has consumption increased and is the mother better off? Why or why not?

c. Now suppose the income guarantee is reduced by one-half to the amount of $100 per month. What is the new number of hours worked and the consumption level? Compare with your result in part b.

d. Consider again the income guarantee in part b of $200 per month, and suppose that the government complements this benefit by offering free child care. Draw the new budget set, and calculate the number of hours worked and consumption level. Calculate the total cost of the program for the government. How does it compare with the program in part b? Define program efficiency as the ratio of the mother’s consumption to government expenditure. Which program dominates on efficiency grounds?

16.22 (Stern 1976) The numerical analysis of income taxation has often assumed that the skill distribution is identical to the income distribution. This need not be the case. Assume that the utility function is defined as

\[ U(x, \ell) = \begin{cases} 
1 - \ell, & x \geq \bar{x}, \\
-\infty, & \text{otherwise} 
\end{cases} \]
a. Provide an interpretation of this utility function.

b. Assume that the skill level of all consumers is strictly positive. What will be the observed distribution of income?

c. Will this distribution be identical to the skill distribution?

d. Comment on the practice of assuming that the income and skill distributions are identical.

16.23 Suppose that income is fixed and distributed according to \( F(y) \) on the interval \([0, Y]\), with the median income below the average income, \( y_m \leq \bar{y} \). Consider the quadratic tax function \( t(y) = -c + by + ay^2 \) with \( c > 0 \) the uniform grant, \( b \) the proportionality parameter, and \( a > 0 \) (\( a < 0 \)) the progressivity (regressivity) parameter. Feasible taxation requires \( t(y) \leq y \) for all \( y \in [0, Y] \). Suppose that taxation is purely redistributive so that budget balance requires \( \int t(y)dF(y) = 0 \). With the quadratic tax schedule \( t(y) = -c + by + ay^2 \), the budget balance requirement implies that \( c = b\bar{y} + a\bar{y}^2 \), where \( \bar{y} \) is the average of square income levels \( \int y^2dF(y) \). Show that for any feasible tax scheme \( t_2 = -c_2 + b_2y + a_2y^2 \) there exists a feasible more progressive (less regressive) tax scheme \( t_1 = -c_1 + b_1y + a_1y^2 \) with \( a_1 > a_2 \), \( b_1 < (>)b_2 \) and \( c_1 > c_2 \) such that more than half the voters would prefer \( t_1 \) to \( t_2 \).

16.24 (Joint taxation) In some countries the tax unit is the individual. In other countries, such as France, the tax unit is the household. Consider a two-person household where the first earner earns \( w_f \) and the second earner earns \( w_s \). In an individual system, the household pays tax of amount \( T_I(w_f, w_s) = t(w_f) + t(w_s) \). In a joint tax system, the household pays \( T_J(w_f, w_s) = 2t(\frac{w_f + w_s}{2}) \).

a. Bill and Jody live together. Bill earns 30 and Jody earns 10. The tax system \( t(w) \) has two brackets with a tax rate of 10 percent for income between 0 and 20 and a tax rate of 40 percent for income above 20. What would be the average tax rate of the household if the tax system is individual? What if it is a joint system?

b. Show that for all \((w_f, w_s)\), \( T_J(w_f, w_s) \leq T_I(w_f, w_s) \). Explain why this is true for any convex and increasing tax function.

c. Taxation schemes are often based on the idea of horizontal equity: households having the same ability to pay taxes (represented by taxable income) should pay a similar amount of tax. John is single and earns 40. What is John’s average tax rate? Do you think that the individual taxation system respect the horizontal equity principle? And what about the joint taxation system?

16.25 (Kleven, Kreiner, and Saez 2009) Consider the previous exercise and assume that the first earner works full time and earns \( w_f = \bar{w} \). The second earner chooses her labor supply \( l \in [0, 1] \) and earns \( w_s = wLf \), with \( \bar{w} \leq \bar{w} \).

a. Show that for all increasing and weakly convex tax function \( t(w) \), the marginal tax rate of the second earner is greater with a joint tax system than with an individual tax system.

b. What is the expected impact of the joint tax system on the labor supply of the second earner compared to the individual tax system? Is it possible to deduce the impact of joint taxation on the well-being of the household?
17  Tax Evasion

17.1 Introduction

It is not unusual to be offered a discount for payment in cash. This is almost routine in the employment of the services of builders, plumbers, and decorators. It is less frequent, but still occurs, when major purchases are made in shops. While the expense of banking checks and the commissions charged by credit card companies may explain some of these discounts, the usual explanation is that payment in cash makes concealment of the transaction much easier. Income that can be concealed need not be declared to the tax authorities.

The same motivation can be provided for exaggeration in claims for expenses. By converting income into expenses that are either exempt from tax or deductible from tax, the total tax bill can be reduced. Second jobs are also a lucrative source of income that can be concealed from the tax authorities. A tax return that reports no income, or at least a very low level, is likely to attract more attention than one that declares earnings from primary employment but fails to mention income from secondary employment.

In contrast to these observations on tax evasion, the analysis of taxation in the previous chapters assumed that firms and consumers reported their taxable activities honestly. Although acceptable for providing simplified insights into the underlying issues, this assumption is patently unacceptable when confronted with reality. The purpose of this chapter can be seen as the introduction of practical constraints on the free choice of tax policy. Tax evasion, the intentional failure to declare taxable economic activity, is pervasive in many economies as the evidence given in the following section makes clear and is therefore a subject of practical as well as theoretical interest.

The chapter begins by considering how tax evasion can be measured. Evidence on the extent of tax evasion in a range of countries is reviewed. The chapter then proceeds to try to understand the factors involved in the decision to evade tax. Initially this decision is represented as a choice under uncertainty. The analysis predicts the relationship among the level of evasion, tax rates, and punishments. Within this framework the optimal degree of auditing and of punishment is considered. Evidence that can be used to assess the model’s predictions is then discussed. In the light of this, some extensions of the basic model are considered. A game-theoretic approach to tax compliance is presented where taxpayers and governments interact strategically. The last section emphasizes the importance of social interaction on compliance decisions.
17.2 The Extent of Evasion

Tax evasion is illegal, so those engaging in it have every reason to seek to conceal what they are doing. This introduces a fundamental difficulty into the measurement of tax evasion. Even so, the fact that the available estimates show that evasion constitutes a significant part of total economic activity underlines the importance of measurement. The lost revenue due to tax evasion also confirms the value of developing a theory of evasion that can be used to design a tax structure that minimizes evasion and ensures that policy is optimal given that evasion occurs. Before proceeding, it is worth making some distinctions. First, *tax evasion* is the failure to declare taxable activity. Tax evasion should be distinguished from *tax avoidance*, which is the reorganization of economic activity, possibly at some cost, to lower tax payment. Tax avoidance is legal, whereas tax evasion is not. In practice, the distinction is not as clear-cut, since tax avoidance schemes frequently need to be tested in court to clarify their legality. Second, the terms *black*, *shadow*, or *hidden economy* refer to all economic activities for which payments are made but are not officially declared. Under these headings would be included illegal activities, such as the drug trade, and unmeasured activity, such as agricultural output by smallholders. Added to these would also be the legal, but undeclared, income that constitutes tax evasion. Finally the *unmeasured* economy would be the shadow economy plus activities such as do-it-yourself jobs that are economically valuable but do not involve economic transaction.

This discussion reveals that there are several issues concerning how economic activity should be divided between the regular economy and the shadow economy. For instance, most systems of national accounts do not include criminal activity (Italy, however, does make some adjustment for smuggling). In principle, the UN System of National Accounts includes both legal and illegal activities, and it has been suggested that criminal activity should be made explicit when the system is revised. Although this chapter is primarily about tax evasion, an attempt at the measurement of tax evasion may include some or all of the components of the shadow economy.

The essential problem involved in the measurement of tax evasion is that its illegality provides an incentive for individuals to keep the activity hidden. Furthermore, by its very nature, tax evasion does not appear in any official statistics. This implies that the extent of tax evasion cannot be measured directly but must be inferred from economic variables that can be observed.

A first method for measuring tax evasion is to use survey evidence. This can be employed either directly or indirectly as an input into an estimation procedure. The
obvious difficulty with survey evidence is that respondents who are active in the hidden economy have every incentive to conceal the truth. There are two ways in which the problem of concealment can be circumvented. First, information collected for purposes other than the measurement of tax evasion can be employed. One example of this is the use that has been made of data from the Family Expenditure Survey in the United Kingdom. This survey involves consumers recording their incomes and expenditures in a diary. Participants have no reason to falsely record information. The relation between income and expenditure can be derived from the respondents whose entire income is obtained in employment that cannot escape tax. The expenditures recorded can then be used to infer the income of those who do have an opportunity to evade. Although these records are not surveys in the normal sense, studies of taxpayer compliance conducted by revenue collection agencies, such as the Internal Revenue Service in the United States, can be treated as survey evidence and have some claim to accuracy.

The second general method is to infer the extent of tax evasion, or the hidden economy generally, from the observation of other economic variables. This is done by determining total economic activity and then subtracting measured activity, which gives the hidden economy. The direct input approach observes the use of an input to production and from this predicts what output must be. An input that is often used for this purpose is electricity, since this is universally employed and accurate statistics are kept on energy consumption. The monetary approach employs the demand for cash to infer the size of the hidden economy on the basis that transactions in the hidden economy are financed by cash rather than checks or credit cards. Given a relationship between the quantity of cash and the level of economic activity, this allows estimation of the hidden economy. What distinguishes alternative studies that fall under the heading of monetary approaches is the method used to derive the total level of economic activity from the observed use of cash. One way to do this is to assume that there was a base year in which the hidden economy did not exist. The ratio of cash to total activity is then fixed by that year. This ratio allows observed cash use in other years to be compounded up into total activity. An alternative has been to look at the actual use of banknotes. The issuing authorities know the expected lifespan of a note (i.e., how many transactions it can finance). Multiplying the number of notes used by the number of transactions gives the total value of activity financed.

Table 17.1 presents estimates of the size of the hidden economy for a range of countries. These figures are based on a combination of the direct input (actually use of electricity as a proxy for output) and money demand approaches. Further details can be found in the source reference. The table clearly indicates that the hidden economy is a significant issue, especially in the developing and transition economies. Even for Japan


Table 17.1
Hidden economy as percentage of GDP, average over 1990 to 1993

<table>
<thead>
<tr>
<th>Developing</th>
<th>Transition</th>
<th>OECD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Egypt 68–76%</td>
<td>Georgia 28–43%</td>
<td>Italy 24–30%</td>
</tr>
<tr>
<td>Thailand 70%</td>
<td>Ukraine 28–43%</td>
<td>Spain 24–30%</td>
</tr>
<tr>
<td>Mexico 40–60%</td>
<td>Hungary 20–28%</td>
<td>Denmark 13–23%</td>
</tr>
<tr>
<td>Malaysia 38–50%</td>
<td>Russia 20–27%</td>
<td>France 13–23%</td>
</tr>
<tr>
<td>Tunisia 39–45%</td>
<td>Latvia 20–27%</td>
<td>Japan 8–10%</td>
</tr>
<tr>
<td>Singapore 13%</td>
<td>Slovakia 9–16%</td>
<td>Austria 8–10%</td>
</tr>
</tbody>
</table>

Source: Schneider and Enste (2000).

and Austria, which have the smallest estimated size of hidden economy, the percentage figure is still significant.

As already noted, these estimates are subject to error and must be treated with a degree of caution. Having said this, there is a degree of consistency running through them. They indicate that a value for the hidden economy of at least 10 percent is not an unreasonable approximation. Therefore the undeclared economic activity is substantial relative to total economic activity. Tax evasion is clearly an important phenomenon that merits extensive investigation.

17.3 The Evasion Decision

The estimates of the hidden economy have revealed that tax evasion is a significant part of overall economic activity. We now turn to modeling the decision to evade in order to understand how the decision is made and the factors that can affect that decision.

The simplest model of the evasion decision considers it to be just a gamble. If taxpayers declare less than their true income (or overstate deductions), there is a chance that they may do so without being detected. This leads to a clear benefit over making an honest declaration. However, there is also a chance that they may be caught. When they are, a punishment is inflicted (usually a fine but sometimes imprisonment) and they are worse off than if they had been honest. In deciding how much to evade, the taxpayer has to weigh up these gains and losses, taking account of the chance of being caught and the level of the punishment.

A simple formal statement of this decision problem can be given as follows. Let the taxpayer have an income level $Y$, which they know but is not known to the tax collector. The income declared by the taxpayer, $X$, is taxed at a constant rate $t$. The amount of
unreported income is \( Y - X \geq 0 \) and the unpaid tax is \( r(Y - X) \). If the taxpayer evades without being caught, their income is given by \( Y^\text{nc} = Y - rX \). When the taxpayer is caught evading, all income is taxed and a fine at rate \( F \) is levied on the tax that has been evaded. This gives an income level \( Y^c = [1 - r]Y - Fr(Y - X) \). If income is understated, the probability of being caught is \( p \).

Assume that the taxpayer derives utility \( U(Y) \) from an income \( Y \). After making declaration \( X \), the income level \( Y^c \) occurs with probability \( p \) and the income level \( Y^\text{nc} \) with probability \( 1 - p \). In the face of such uncertainty the taxpayer should choose the income declaration to maximize expected utility. Combining these facts, the declaration \( X \) solves

\[
\max_{X} \mathbb{E}[U(X)] = (1 - p)U(Y^\text{nc}) + pU(Y^c).
\]  

(17.1)

The solution to this choice problem can be derived graphically. To do this, observe that there are two states of the world. In one state of the world, taxpayers are not caught evading and have income \( Y^\text{nc} \). In the other state of the world, they are caught and have income \( Y^c \). The expected utility function describes preferences over income levels in these two states. The choice of a declaration \( X \) determines an income level in each state, and by varying \( X \), the taxpayer can trade off income between the two states. A high value of \( X \) provides relatively more income in the state where the taxpayer is caught evading and a low value of \( X \) relatively more where the taxpayer is not caught.

The details of this trade-off can be identified by considering the two extreme values of \( X \). When the maximum declaration is made so that \( X = Y \), the taxpayer’s income will be \([1 - r]Y\) in both states of the world. Alternatively, when the minimum declaration of \( X = 0 \) is made, income will be \([1 - r(1 + F)]Y\) if caught and \( Y \) if not. These two points are illustrated in figure 17.1, which graphs income when not caught against income when caught. The other options available to the consumer lie on the line joining \( X = 0 \) and \( X = Y \); this is the opportunity set showing the achievable allocations of income between the two states. From the utility function can be derived a set of indifference curves—the points on an indifference curve being income levels in the two states that give the same level of expected utility. Including the indifference curves of the utility function completes the diagram and allows the taxpayer’s choice to be depicted. The taxpayer whose preferences are shown in figure 17.1 chooses to locate at the point with declaration \( X^* \). This is an interior point with \( 0 < X^* < Y \)—some tax is evaded but some income is declared.

Besides the interior location of figure 17.1 it is possible for corner solutions to arise. The consumer whose preferences are shown in panel a of figure 17.2 chooses to declare
Figure 17.1
Interior choice: $0 < X^* < Y$

Figure 17.2
Corner solutions
his entire income, so \( X^* = Y \). In contrast the consumer in panel b declares no income, so \( X^* = 0 \).

The interesting question is what condition guarantees that evasion will occur rather than the no-evasion corner solution with \( X = Y \). Comparing the figures it can be seen that evasion will occur if the indifference curve is steeper than the budget constraint where it crosses the dashed 45 degree line. The condition that ensures that this occurs is easily derived. Totally differentiating the expected utility function (17.1) at a constant level of utility gives the slope of the indifference curve as

\[
\frac{dY^c}{dY^{nc}} = -\frac{[1 - p] U'(Y^{nc})}{p U'(Y^c)},
\]

where \( U'(Y) \) is the marginal utility of income level \( Y \). On the 45 degree line \( Y^{nc} = Y^c \), so the marginal utility of income is the same whether or not the tax evader is caught. This implies that

Slope of indifference curve  =  \(- \frac{1 - p}{p}\).  \hspace{1cm} (17.3)

What this expression suggests is that all the indifference curves have the same slope, \(- \frac{1 - p}{p}\), where they cross the 45 degree line. The slope of the budget constraint is seen in figure 17.1 to be given by the ratio of the penalty \( Ft [Y - X] \) to the unpaid tax \( t[Y - X] \). Thus

Slope of budget constraint  =  \(- F\).  \hspace{1cm} (17.4)

Because of these features the indifference curve is steeper than the budget constraint on the 45 degree line if \( \frac{1 - p}{p} > F \), or

\[ p < \frac{1}{1 + F}. \]  \hspace{1cm} (17.5)

This result shows that evasion will arise if the probability of detection is too small relative to the fine rate.

Several points can be made about this condition for evasion. First, this is a trigger condition that determines whether or not evasion will arise, but it does not say anything about the extent of evasion. Second, the condition is dependent only on the fine rate and the probability of detection, so it applies for all taxpayers regardless of their utility-of-income function \( U(Y) \). Consequently, if one taxpayer chooses to evade, all taxpayers should evade. Third, this condition can be given some practical evaluation. Typical punishments inflicted for tax evasion suggest that an acceptable magnitude for
$F$ is between 0.5 and 1. In the United Kingdom, the Taxes Management Act specifies the maximum fine as 100 percent of the tax lost, which implies the maximum value of $F = 1$. This makes the ratio $\frac{1}{1+F}$ greater or equal to $\frac{1}{2}$. Information on $p$ is hard to obtain, but a figure between 1 in a 100 or 1 in a 1,000 evaders being caught is probably a fair estimate. Therefore $p < \frac{1}{2} < \frac{1}{1+F}$, and the conclusion is reached that the model predicts all taxpayers should be evading. In the United States, taxpayers who understate their tax liabilities may be subjected to penalties at a rate between 20 to 75 percent of the underreported taxes, depending on the gravity of the offence. The proportion of all individual tax returns that are audited was 1.7 percent in 1997. This is clearly not large enough to deter cheating, and everyone should be underreporting taxes. In fact the Taxpayer Compliance Measurement Program reveals that 40 percent of US taxpayers underpaid their taxes. This is a sizable minority but not the widespread evasion the theoretical model would predict. So taxpayers appear to be more honest than might be expected.

The next step is to determine how the amount of tax evasion is affected by changes in the model’s variables. There are four such variables that are of interest: the income level $Y$, the tax rate $t$, the probability of detection $p$, and the fine rate $F$. These effects can be explored by using the figure depicting the choice of evasion level.

Take the probability of detection first. The probability of detection does not affect the opportunity set but does affect preferences. The effect of an increase in $p$ is to make the indifference curves flatter where they cross the 45 degree line. As shown in figure 17.3, this moves the optimal choice closer to the point $X = Y$ of honest declaration. The amount of income declared rises, so an increase in the probability of detection reduces the level of evasion. This is a clearly expected result, since an increase in the likelihood of detection lowers the payoff from evading and makes evasion a less attractive proposition.

A change in the fine rate only affects income when the taxpayer is caught evading. The consequence of an increase in $F$ is that the budget constraint pivots round the honest report point and becomes steeper. Since the indifference curve is unaffected by the penalty change, the optimal choice must again move closer to the honest declaration point. This is shown in figure 17.4 by the move from the initial choice of $X^{old}$ when the fine rate is $F$ to the choice $X^{new}$ when the fine rate increases to $\hat{F}$. An increase in the fine rate therefore leads to a reduction in the level of tax evasion. This result, and the previous result, show that the effects of the detection and punishment variables on the level of evasion are unambiguous.

Now consider the effect of an increase in income from the initial level $Y$ to a higher level $\hat{Y}$. This income increase causes the budget constraint to move outward. As already
Chapter 17: Tax Evasion

Figure 17.3
Increase in detection probability

Figure 17.4
Increase in the fine rate
noted, the slope of the budget constraint is equal to $-F$, which does not change with income, so the shift is a parallel one. The optimal choice then moves from $X^{old}$ to $X^{new}$ in figure 17.5. How the evasion decision is affected depends on the degree of absolute risk aversion, $R_A(Y) = -\frac{U''(Y)}{U'(Y)}$, of the utility function. What absolute risk aversion measures is the willingness to engage in small bets of fixed size. If $R_A(Y)$ is constant as $Y$ changes, the optimal choices will be on a locus parallel to the 45 degree line. There is evidence, though, that in practice, $R_A(Y)$ decreases as income increases, so wealthier individuals are more prone to engage in small bets, in the sense that the odds demanded to participate diminish. This causes the locus of choices to bend away from the 45 degree line, so that the amount of undeclared income rises as actual income increases. This is the outcome shown in figure 17.5. Hence, with decreasing absolute risk aversion, an increase in income increases tax evasion.

The final variable to consider is the tax rate. An increase in the tax rate from $t$ to $\hat{t}$ moves the budget constraint inwards. As can be seen in figure 17.6, the outcome is not clear-cut. However, when absolute risk aversion is decreasing, the effect of the tax increase is to reduce tax evasion. This final result has received much discussion because it is counter to what seems intuitive. A high tax rate is normally seen as providing a motive for tax evasion, whereas the model predicts precisely the converse. Why the result emerges is because the fine paid by the consumer is determined by $t$ times $F$. An increase in the tax rate thus has the effect of raising the penalty. A tax rate increase takes income away from the taxpayer when they are caught—the state in which they
Figure 17.6
Tax rate increase

have least income. It is through this mechanism that a higher tax rate can reduce evasion.

This completes the analysis of the basic model of tax evasion. It has been shown how the level of evasion is determined and how this is affected by the parameters of the model. The next section turn to the issue of determining the optimal levels of auditing and punishment when the behavior of taxpayers corresponds to the predictions of this model. Some empirical and experimental evidence is then considered and used to assess the predictions of the model.

17.4 Auditing and Punishment

The analysis of the tax evasion decision assumed that the probability of detection and the rate of the fine levied, when caught evading, were fixed. This is a satisfactory assumption from the perspective of the individual taxpayer. From the government’s perspective, though, these are variables that can be chosen. The probability of detection can be raised by the employment of additional tax inspectors, and the fine can be legislated or set by the courts. The purpose of this section is to analyze the issues involved in the government’s decision.

It has already been shown that an increase in either $p$ or $F$ will reduce the amount of undeclared income. The next step is to consider how $p$ and $F$ affect the level of
revenue raised by the government. Revenue in this context is defined as taxes paid plus the money received from fines. From a taxpayer with income \( Y \), the expected value (it is expected, since there is only a probability the taxpayer will be fined) of the revenue collected is

\[
R = tX + p [1 + F] t \{ Y - X \}. \tag{17.6}
\]

Differentiating with respect to \( p \) shows that the effect on revenue of an increase in the probability of detection is

\[
\frac{\partial R}{\partial p} = [1 + F] t \{ Y - X \} + t \{ 1 - p - pF \} \frac{\partial X}{\partial p} > 0, \tag{17.7}
\]

whenever \( pF < 1 - p \). Recall from (17.5) that if \( pF \geq 1 - p \), there is no evasion, so \( p \) has no effect on revenue. Carrying out the differentiation for the fine rate shows that if \( pF < 1 - p \)

\[
\frac{\partial R}{\partial F} = pt \{ Y - X \} + t \{ 1 - p - pF \} \frac{\partial X}{\partial F} > 0. \tag{17.8}
\]

An increase in the fine will therefore raise revenue if tax evasion is taking place. Again, the fine has no effect if \( pF \geq 1 - p \) and there is no evasion. These expressions show that if evasion is taking place, an increase in the probability of either detection or the fine will increase the revenue the government receives.

The choice problem of the government can now be addressed. It has already been noted that an increase in the probability of detection can be achieved by the employment of additional tax inspectors. Tax inspectors require payment; as a consequence an increase in \( p \) is costly. In contrast, there is no cost involved in raising or lowering the fine. Effectively, increases in \( F \) are costless to produce. From these observations the optimal policy can be determined.

Since \( p \) is costly and \( F \) is free, the interests of the government are best served by reducing \( p \) close to zero while raising \( F \) toward infinity. Serge Kolm has termed this the policy of “hanging taxpayers with probability zero.” Expressed in words, the government should put virtually no effort into attempting to catch tax evaders but should severely punish those it apprehends. This is an extreme form of policy, and nothing like it is observed in practice. Surprising as it is, it does follow from the logical application of the model.

Numerous comments can be made about this conclusion. The first begins with the objective of the government. In previous chapters it was assumed that the government is guided in its policy choice by a social welfare function. There will be clear
differences between a policy designed to maximize revenue and one that maximizes welfare. For instance, inflicting an infinite fine on a taxpayer caught evading will have a significantly detrimental effect on welfare. Even if the government does not pursue welfare maximization, it may be constrained by political factors such as the need to ensure re-election. A policy of severely punishing tax evaders may be politically damaging especially if tax evasion is a widely established phenomenon.

One could think that such an argument is not relevant because, if the punishment is large enough to deter cheating, it should not matter how dire it is. If fear keeps everyone from cheating, the punishment never actually occurs and its cost is irrelevant. The problem with this argument is that it ignores the risk of mistakes. The detection process may go wrong, or the taxpayer can mistakenly understate taxable income. If the punishment is as large as possible, even for small tax underpayments, then mistakes will be very costly. To reduce the cost of mistakes, the punishment should be of the smallest size required to deter cheating. Minimal deterrence accomplishes this purpose.

A further observation, and one where the consequences will be investigated in detail, concerns the policy instruments under the government’s control. The view of the government so far is that it is a single entity that chooses the level of all its policy instruments simultaneously. In practice, the government consists of many different departments and agencies. When it comes to taxation and tax policy, a reasonable breakdown would be to view the tax rate as set by central government as part of a general economic policy. The probability of detection is controlled by a revenue service whose objective is the maximization of revenue. Finally the punishment for tax evasion is set by the judiciary. This breakdown shows why the probability and fine may not be chosen in a cohesive manner by a single authority. What it does not do is provide any argument for why the fine should not be set infinitely high to deter evasion. An explanation can be found by applying reasoning from the economics of crime. This would view tax evasion as just another crime, and the punishment for it should fit with the scheme of punishments for other crimes. The construction of punishments relies on the argument that they should provide incentives that lessen the overall level of crime. To see what this means, imagine that crimes can be ranked from least harmful to most harmful. Naturally, if someone is going to commit a crime, the authorities wish that they commit a less harmful one rather than a more harmful one. If more harmful ones are also more rewarding (think of robbing a bank while armed compared to merely attempting to snatch cash), then a scheme of equal punishments will not provide any incentive for committing the less harmful crime. What will provide the right incentive is for the more harmful crime to also have heavier punishment. So the extent of punishment should be related to the harmfulness of the crime: punishment should fit the crime.
This framework has two implications. First of all, the punishment for tax evasion will not be varied freely in order to maximize revenue. Instead, it will be set as part of a general crime policy. The second implication is that the punishment will also be quite modest, since tax evasion is not an especially harmful crime. Arguments such as these are reflected in the fact that the fine rate on evasion is quite low—a value between 1.5 and 2 would not be unrealistic. As already noted, the maximum fine in the United Kingdom is 100 percent of the unpaid tax, but the Inland Revenue may accept a lesser fine depending on the “size and gravity” of the offense.

Putting all of these arguments together suggests adopting a different perspective on choosing the optimal probability of detection. With the tax rate set as a tool of economic policy and the fine set by the judiciary, the only instrument under the control of the revenue service is the probability of detection. As already seen, an increase in this raises revenue but only does so at a cost. The optimal probability is found when the marginal gain in revenue just equals the marginal cost—and this could occur at a very low value of the probability of detection.

17.5 Evidence on Evasion

The model of tax evasion has predicted the effect that changes in various parameters will have on the level of tax evasion. In some cases, such as the effect of the probability of detection and the fine, these are unambiguous. In others, particularly the effect of changes in the tax rate, the effects depend on the precise specification of the tax system and on assumptions concerning attitudes toward risk. These uncertainties make it valuable to investigate further evidence to see how the ambiguities are resolved in practice. The analysis of evidence also allows the investigation of the relevance of other parameters, such as the source of income, and other hypotheses on tax evasion, such as the importance of social interaction.

There have been two approaches taken in studying tax evasion. The first was to collect survey or interview data and use econometric analysis to provide a quantitative determination of the relationships. The second was to use experiments to provide an opportunity for designing the environment to permit the investigation of particular hypotheses.

When income levels ascertained from interviews were contrasted to those given on the tax returns of the same individuals, a steady decline of declared income as a proportion of reported income appeared with income rises. This finding is in agreement with the comparative statics analysis. Table 17.2 provides a sample of data to illustrate
Table 17.2
Declaration and income

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Midpoint</td>
<td>18.5</td>
<td>22.5</td>
<td>27.5</td>
<td>32.5</td>
<td>37.5</td>
</tr>
<tr>
<td>Assessed income</td>
<td>17.5</td>
<td>20.6</td>
<td>24.2</td>
<td>28.7</td>
<td>31.7</td>
</tr>
<tr>
<td>Percentage</td>
<td>94.6</td>
<td>91.5</td>
<td>88.0</td>
<td>88.3</td>
<td>84.5</td>
</tr>
</tbody>
</table>

Source: Mork (1975).

Table 17.3
Explanatory factors

<table>
<thead>
<tr>
<th>Variable</th>
<th>Propensity to evade</th>
<th>Extent of evasion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inequity</td>
<td>0.34</td>
<td>0.24</td>
</tr>
<tr>
<td>Number of evaders known</td>
<td>0.16</td>
<td>0.18</td>
</tr>
<tr>
<td>Probability of detection</td>
<td>−0.17</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>−0.29</td>
<td></td>
</tr>
<tr>
<td>Experience of audits</td>
<td>0.22</td>
<td>0.29</td>
</tr>
<tr>
<td>Income level</td>
<td>−0.27</td>
<td></td>
</tr>
<tr>
<td>Income from wages and salaries</td>
<td>0.20</td>
<td></td>
</tr>
</tbody>
</table>

Source: Spicer and Lundstedt (1976).

this. Interviewees were placed in income intervals according to their responses to interview questions. The information on their tax declaration was then used to determine assessed income. The percentage is found by dividing the assessed income by the midpoint of the income interval.

Econometrics and survey methods have been used to investigate the importance of attitudes and social norms in the evasion decision. The study reported in table 17.3 shows that the propensity to evade taxation is reduced by an increased probability of detection and an increase in age. An increase in income reduces the propensity to evade. With respect to attitude and social variables, both an increase in the perceived inequity of taxation and of the number of other tax evaders known to the individual make evasion more likely. The extent of tax evasion is increased by attitude and social variables but also by the experience of the taxpayer with previous tax audits. The social variables are clearly important in the decision to evade tax.

As far as the effect of the tax rate is concerned, data from the US Internal Revenue Services Taxpayer Compliance Measurement Program survey of 1969 shows that tax evasion increases as the marginal tax rates increases but is decreased when wages are a
significant proportion of income. This result is supported by employing the difference between income and expenditure figures in National Accounts as a measure of evasion. In contrast, a study of Belgian data found precisely the converse conclusion, with tax increases leading to lower evasion. Therefore these studies do not resolve the ambiguity about the relation between marginal tax rates and tax evasion.

Turning now to experimental studies, tax evasion games have shown that evasion increases with the tax rate and that evasion falls as the fine is increased and the detection probability reduced. Further results have shown that women evade more often than men but evade lower amounts and that purchasers of lottery tickets, presumed to be less risk averse, are no more likely to evade than nonpurchasers but evade greater amounts when they do evade. Finally the very nature of the tax evasion decision has been tested by running two sets of experiments. One was framed as a tax evasion decision and the other as a simple gamble with the same risks and payoffs. For the tax evasion experiment some taxpayers chose not to evade even when they would under the same conditions with the gambling experiment. This suggests that tax evasion is not viewed as just a gamble.

There are two important lessons to be drawn from this brief review of the empirical and experimental results. First, the theoretical predictions are generally supported except for the effect of the tax rate. The latter remains uncertain with conflicting conclusions from the evidence. Second, it appears that tax evasion is more than the simple gamble portrayed in the basic model. In addition to the basic element of risk, there are social aspects to the evasion decision.

17.6 Effect of Honesty

The evidence discussed in the previous section has turned up a number of factors that are not explained by the basic model of tax evasion. Foremost among these are that some taxpayers choose not to evade even when they would accept an identical gamble and that there are social aspects of the evasion decision. The purpose of this section is to show how simple modifications to the model can incorporate these factors and can change the conclusions concerning the effect of the tax rate.

The feature that distinguishes tax evasion from a simple gamble is that taxpayers submitting incorrect returns feel varying degrees of anxiety and regret. To some, being caught would represent a traumatic experience that would do immense damage to their self-esteem. To others, it would be only a slight inconvenience. The innate belief in honesty of some taxpayers is not captured by representing tax evasion as just a gamble.
nor are the nonmonetary costs of detection and punishment captured by preferences defined on income alone. The first intention of this section is to incorporate these features into the analysis and to study their consequences.

A preference for honesty can be introduced by writing the utility function as

\[ U = U(Y) - \chi E, \]  

(17.9)

where \( \chi \) is the measure of the taxpayers honesty and, with \( E = Y - X \) the extent of evasion, \( \chi E \) is the utility (or psychic) cost of deviating from complete honesty. To see the consequence of introducing a psychic cost of evasion, suppose that taxpayers differ in their value of \( \chi \) but are identical in all other respects. Taxpayers with higher values of \( \chi \) will suffer from a greater utility reduction for any given level of evasion. In order for them to evade, the utility gain from evasion must exceed the utility reduction. The population is therefore separated into two parts: some taxpayers choosing not to evade (those with high values of \( \chi \)) and others who evade (those with low \( \chi \)). It is tempting to label taxpayers who do not evade as honest, but this is not really appropriate, since they will evade if the benefit is sufficiently great.

Let the value of \( \chi \) that separates the evaders from the nonevaders be denoted \( \hat{\chi} \). A change in any of the parameters of the model \( (p, F, \text{and } t) \) now has two effects. First, it changes the benefit from evasion, which alters the value of \( \hat{\chi} \). For instance, an increase in the rate of tax raises the benefit of evasion and increases \( \hat{\chi} \) with the consequence that more taxpayers evade. Second, the change in the parameter affects the evasion decision of all existing tax evaders. Putting these effects together, it becomes possible for an increase in the tax rate to lead to more evasion in the aggregate. This is in contrast to the basic model where the tax rate increase would reduce evasion.

The discussion of the empirical evidence has drawn attention to the positive connection between the number of tax evaders known to a taxpayer and the level of that taxpayer’s own evasion. This observation suggests that the evasion decision is not made in isolation by each taxpayer but is made with reference to the norms and behavior of the general society of the taxpayer. Given the empirical significance of such norms, the second part of this section focuses on their implications.

Social norms have been incorporated into the model of the evasion decision in two distinct ways. One approach is to introduce social norms as an additional element of the utility cost to evasion. The additional utility cost is assumed to be an increasing function of the proportion of taxpayers who do not evade. This formulation captures the fact that more utility will be lost, in terms of reputation, the more out of step the taxpayer is with the rest of society. The consequence of this modification is to reinforce the separation of the population into evaders and nonevaders.
An alternative approach is to explicitly impose a social norm on behavior. One such social norm can be based on the notion of Kantian morality and, effectively, on having individuals assess their fair contribution in tax payments toward the provision of public goods. This calculation then provides an upper bound on the extent of tax evasion. To calculate the actual degree of tax evasion, each taxpayer performs the expected utility maximization calculation, as in (17.1), and evade whichever is the smaller out of this quantity and the previously determined upper bound. This formulation is also able to provide a positive relation between the tax rate and evaded tax for some range of taxes and to divide the population into those who evade tax and those who do not.

The introduction of psychic costs and of social norms is capable of explaining some of the empirically observed features of tax evasion that are not explained by the standard expected utility maximization hypothesis. This is achieved by modifying the form of preferences, but the basic nature of the approach is unchanged. The obvious difficulty with these changes is that there is little to suggest precisely how social norms and utility costs of dishonesty should be formalized.

### 17.7 Tax Compliance Game

An initial analysis of the choice of audit probability was undertaken in section 17.4. It was argued there that the practical situation involves a revenue service that chooses the probability to maximize total revenue, taking as given the tax rate and the punishment. The choice of probability in this setting requires an analysis of the interaction between the revenue service and the taxpayers. The revenue service reacts to the declarations of taxpayers, and taxpayers make declarations on the basis of the expected detection probability.

Such interaction is best analyzed by formalizing the structure of the game that is being played between the revenue service and the taxpayers. The choice of a strategy for the revenue service is the probability with which it chooses to audit any given value of declaration. This probability need not be constant for declarations of different values and is based on its perception of the behavior of taxpayers. For the taxpayers, a strategy is a choice of declaration given the audit strategy of the revenue service. At a Nash equilibrium of the game, the strategy choices must be mutually optimal: the audit strategy must maximize the revenue collected, net of the costs of auditing, given the declarations; the declaration must maximize utility given the audit strategy.

Even without specifying further details of the game, it is possible to make a general observation: predictability in auditing cannot be an equilibrium strategy. First, no
auditing at all cannot be optimal because it encourages maximal tax evasion. Second, auditing of all declarations cannot be a solution either because no revenue service incurs the cost of auditing where full enforcement induces everyone to comply. Finally, pre-specified limits on the range of declarations to be audited are also flawed. Taxpayers tempted to underreport income will make sure to stay just outside the audit limit, and those who cannot avoid being audited will choose to report truthfully. Exactly the wrong set of taxpayers will be audited. This establishes that the equilibrium strategy must involve randomization.

But how should the probability of audit depend on the information available on the tax return? Since the incentive of a taxpayer is to understate income to reduce their tax liabilities, it seems to require that the probability of an audit should be higher for low-income reports. More precisely, the probability of an audit should be high for an income report that is low compared to what one would expect from someone in that taxpayer’s occupation or given the information on previous tax returns for that taxpayer. This is what theory predicts and what is done in practice.

A simple version of the strategic interaction between the revenue service and a taxpayer is depicted in figure 17.7. The taxpayer with true income \( Y \) can either evade (reporting zero income) or not (truthful income report). By reporting truthfully, the taxpayer pays tax \( T \) to the revenue service (with \( T < Y \)). The revenue service can either audit the income report or not audit. An audit costs \( C \) for the revenue service to conduct but provides irrefutable evidence as to whether the taxpayer has misreported income. If the taxpayer is caught evading, he pays the tax due, \( T \), plus a fine \( F \) (where the fine includes the cost of auditing and a tax surcharge so that \( F > C \)). If the taxpayer is not caught evading, then he pays no tax at all. The two players choose their strategies simultaneously, which reflects the fact that the revenue service does not know whether
the taxpayer has chosen to evade when it decides whether to audit. To make the problem interesting, we assume that $C < T$, so the cost of auditing is less than its potential gain, which is to recover the tax due.

There is no pure strategy equilibrium in this tax compliance game. On the one hand, if the revenue service does not audit, the agent strictly prefers evading, and therefore the revenue service is better off auditing as $T + F > C$. On the other hand, if the revenue service audits with certainty, the taxpayer prefers not to evade as $T + F > T$, which implies that the revenue service is better off not auditing. Therefore the revenue must play a mixed strategy in equilibrium, with the audit strategy being random (i.e., unpredictable). Similarly for the taxpayer the evasion strategy must also be random.

Let $e$ be the probability that the taxpayer evades, and $p$ the probability of audit. To obtain the equilibrium probabilities, we solve the conditions that the players must be indifferent between their two pure strategies. For the government to be indifferent between auditing and not auditing, it must be the case that the cost from auditing, $C$, equals the expected gain in tax and fine revenue, $e[T + F]$. For the taxpayer to be indifferent between evading and not evading, the expected gain from evading, $(1 - p)T$ equals the expected penalty $pF$. Hence in equilibrium the probability of evasion is

$$e^* = \frac{C}{T + F}, \quad (17.10)$$

and the probability of audit is

$$p^* = \frac{T}{T + F}, \quad (17.11)$$

where both $e^*$ and $p^*$ belong to the interval $(0, 1)$ so that both evasion and audit strategies are random.

The equilibrium probabilities are determined by the strategic interaction between the taxpayer and the revenue service. For instance, the audit probability declines with the fine, although a higher fine may be expected to make auditing more profitable. The reason is that a higher fine discourages evasion, thus making auditing less profitable. Similarly evasion is less likely with a high tax because a higher tax induces the government to audit more. Note that these results are obtained without specifying the details of the fine function, which could be either a lump-sum amount or something proportional to evaded tax. Evasion is also more likely the more costly it is auditing, since the revenue service is willing to audit at a higher cost only if the taxpayer is more likely to have evaded tax.
The equilibrium payoffs of the players are

\[ u^* = Y - T + e^*[T - p^*[T + F]], \]  

(17.12)

for the taxpayer and

\[ v^* = (1 - e^*)T + p^*[e^*[T + F] - C], \]  

(17.13)

for the revenue service. Substituting into these payoffs the equilibrium probabilities of evasion and audit gives

\[ u^* = Y - T, \]  

(17.14)

\[ v^* = T - \frac{C}{T + F} T. \]  

(17.15)

Because the taxpayer is indifferent between evading and not evading, his equilibrium payoff is equal to his truthful payoff \( Y - T \). This means that the unpaid taxes and the fine cancel out in expected terms. Increasing the fine does not affect the taxpayer. However, a higher fine increases the payoff of the revenue service, since it reduces the amount of evasion. Hence increasing the penalty is Pareto-improving in this model. The equilibrium payoffs also reflect the cost from evasion. Indeed, for any tax \( T \) paid by the taxpayer, the revenue service effectively receives \( T - \Delta \), where \( \Delta = \frac{C}{T + F} T \) is the deadweight loss from evasion. Thus evasion involves a deadweight loss that is increasing with the tax rate.

### 17.8 Behavioral Models

The analysis of the evasion decision based on the expected utility function produced very clear and precise results. It provided a sufficient condition for a taxpayer to conceal some income and determined the response of the taxpayer to changes in the key variables \((f, p, t, \text{ and } Y)\). It is rare for a model to provide such unambiguous conclusions. Unfortunately, not all the results it produces are in agreement with the evidence.

When evaluated at the levels of audit probability and fine rate the model predicts that all taxpayers will engage in evasion whenever the opportunity arises. The extent of evasion varies across countries, but there are many countries for which honest payment of taxes is the predominant behavior. For these countries the model does not predict the correct level of evasion. The model also predicts that the level of evasion falls when the tax rate rises. This is a consequence of the fine being a multiple of the tax.
evaded so that the "effective" punishment is $EfE$, where $E$ is the amount of evasion. The effective punishment for a given amount of evasion increases as the tax rate increases so discouraging evasion. Intuition suggests a high tax rate should provide an incentive to evade, whereas the model predicts the opposite. The empirical and experimental evidence on this point is mixed. This has not prevented a presumption developing in the literature that the result is wrong, from which it follows that the model is inadequate and must be improved.

The natural response to the perceived failings of the model is to consider the range of alternative choice models proposed by behavioral economics. These alternative models can be understood as relaxing the very restrictive assumptions of expected utility theory. The behavioral models can be viewed as special cases of the general formulation of the value function

$$V = \sum_{i=1}^{n} w_i(p_1, \ldots, p_n)v(x_i),$$

(17.16)

where $w_i(p_1, \ldots, p_n)$ are decision weights and $v(x_i)$ the payoff function. The interpretation is that the decision maker (the taxpayer for an evasion decision) knows that the true probabilities of the events are $p_1, \ldots, p_n$. However, in their evaluation of the alternatives they transform these probabilities into the decision weights. The usual assumption in behavioral economics is that unlikely events are typically assigned a decision weight greater than the probability. So, if the chance of auditing is small, it is given much greater weight in the subjective assessment of the taxpayer. The logic is of the kind “I know only 4 percent of people are audited but if I evade I will almost certainly be audited.” A utility function, denoted $U(x_i)$, is a special case of a payoff function that satisfies the additional assumptions $U'(\cdot) > 0$ and $U''(\cdot) < 0$. Expected utility fits into this formulation with $w_i(p) = p_i$, and $v(x_i) = U(x_i)$.

Rank dependent expected utility begins by ranking the outcomes from worst to best. For tax evasion the worst outcome is $Y_c$ and the best outcome is $Y_n$. It then assigns a weight $\sigma(Y_c)$ to the worst outcome and $1 - \sigma(Y_c)$ to the best outcome, where it is assumed that $\sigma(0) = 0$ and $\sigma(1) = 1$. The value function can then be written as

$$\max_{\{E\}} V(E) = \sigma(p)U(Y_c) + [1 - \sigma(p)]U(Y_n).$$

(17.17)

Furthermore it is assumed that the weighting function is inverse S-shaped (concave, then convex). This has the effect of inflating low probabilities and deflating large
probabilities. In practice, the probability of audit is low, so it is natural to take \( \sigma (p) > p \). By comparing (17.17) to the expected utility function of the standard model, we can see that the sufficient condition for evasion to occur with rank dependent expected utility is \( 1 - \sigma (p) - \sigma (p) f > 0 \). Since the weighting inflates the probability of detection, this provides a stricter sufficient condition, and so \( p \) must be lower than in the standard model before evasion occurs.

It should be noted that although this formulation changes the prediction about the extent of evasion, it does not change the comparative static effect of a change in the tax rate, \( t \). Given a probability, \( p \), the change of variables \( 1 - \sigma (p) \equiv 1 - q \), \( \sigma (p) \equiv q \) can be made, and the optimization becomes that of the standard model but with \( p \) replaced by \( q \). The qualitative properties of the comparative statics are therefore unchanged.

**Prospect Theory** is based on three assumptions. First, payoffs are based on gains and losses relative to a reference point. Second, the payoff function is concave in gains but convex in losses. Third, the additional utility from a small gain is less than the utility reduction from a small loss. Two issues arise in applying this theory to the evasion decision. Since there is no natural reference point or functional form for the payoff, there is considerable flexibility within Prospect Theory for the representation of the choice problem. The combination of concavity and convexity means that there is often no interior solution for the choice problem so that the evasion choice is all-or-nothing.

The latter point is now illustrated for a commonly used version of Prospect Theory. The payoff function is assumed to take the power function form

\[
v(z) = \begin{cases} 
  z^\beta, & z \geq 0, \\
  -\gamma (-z)^\beta, & z < 0.
\end{cases}
\]

The effect of the parameter \( \gamma \) is to give the payoff function a kink at zero when \( \gamma \neq 1 \). It is usually assumed that \( \gamma > 1 \), which ensures that the utility fall from a small loss is greater than the utility rise from a small gain. Assume that the reference level of income, \( R \), is income after the correct tax payment, so that \( R = Y [1 - t] \). Then, if the taxpayer is caught, the loss relative to the reference level is

\[
Y^c - R = -ft[Y - X]. 
\tag{17.18}
\]

If the taxpayer is not caught the gain is

\[
Y^n - R = t[Y - X]. 
\tag{17.19}
\]
Combining these assumptions determines the payoff function

\[ V = p \left[ -\gamma \left( Y - X \right)^\beta \right] + \left[ 1 - p \right] \left( t Y - X \right)^\beta. \]  

(17.20)

Observe that this payoff function can be factored as

\[ V = E^\beta \left[ \left( 1 - p \right) t^\beta - p \gamma f^\beta t^\beta \right]. \]  

(17.21)

where the level of evasion is \( E = Y - X \). It can be seen directly from (17.21) that the optimal choice of evasion is \( E = Y \) if \( 1 - p - p \gamma f^\beta > 0 \) but \( E = 0 \) if \( 1 - p - p \gamma f^\beta < 0 \). The choice of reference point and payoff function combines to ensure that there will always be a corner solution (except for when \( 1 - p - p \gamma f^\beta = 0 \), in which case the taxpayer is indifferent about the value of \( E \)). The taxpayer either declares honestly or declares nothing. The tax rate does not affect this decision, so the model also predicts that changes in the tax rate will not affect the amount of evasion.

The central feature of Disappointment Theory is the comparison of the outcome that is achieved with the outcomes that could have been achieved. When the realized outcome is poor, an amount of disappointment reduces the payoff. This provides a disincentive to undertake risky actions that may lead to poor outcomes. In the context of tax evasion this will, ceteris paribus, reduce the chosen level of evasion. We now consider whether the inclusion of disappointment modifies the sufficient condition for evasion or the tax effect.

Let \( \tilde{Y}(E) \) denote the random level of income after evading an amount \( E \). \( \tilde{Y}(E) \) is equal to \( Y^c \) with probability \( p \) and \( Y^n \) with probability \( 1 - p \). For Disappointment Theory the payoff with an amount of evasion \( E \) is written as

\[ V = E^\beta \left[ \left( 1 - p \right) t^\beta - \gamma f^\beta t^\beta \right]. \]  

(17.22)

where \( D(\tilde{Y}(E) - E[\tilde{Y}(E)]) \) is the disappointment function and the \( E \) the expectations operator. It is assumed that \( D(\tilde{Y}(E) - E[\tilde{Y}(E)]) > 0 \) if \( \tilde{Y}(E) < E[\tilde{Y}(E)] \) and \( D(\tilde{Y}(E) - E[\tilde{Y}(E)]) < 0 \) if \( \tilde{Y}(E) > E[\tilde{Y}(E)] \). The interpretation of (17.22) is that the payoff is composed of two parts. The first part arises from expected utility of income as in the standard model. The second part subtracts the disappointment that arises. From the definitions of \( Y^n \) and \( Y^c \) it follows that

\[ E[\tilde{Y}(E)] = pY^c + \left[ 1 - p \right] Y^n = Y[1 - t] + \left[ 1 - p - pf \right] tE. \]  

(17.23)

If the taxpayer is audited, then income is \( Y^c \), which is less than \( E[\tilde{Y}(E)] = Y[1 - t] + \left[ 1 - p - pf \right] tE \), so the disappointment is positive. Conversely, \( Y^c > Y[1 - t] + \left[ 1 - p - pf \right] tE \), so the disappointment is negative if the taxpayer is not caught.
Substituting for $\tilde{Y} (E)$ (17.22) can be written

$$V = pU(Y [1 - t] - ftE) + [1 - p] U(Y [1 - t] + tE)$$


(17.24)

From the objective in (17.24) the necessary condition for evasion to occur is that

$$\frac{\partial V}{\partial E} \bigg|_{E=0} > 0.$$ 

(17.25)

Calculating the derivative gives

$$\frac{\partial V}{\partial E} = -pft U'(Y^c) + [1 - p] t U'(Y^n)$$

$$+ pD'(- [1 - p] [1 + f] tE) [1 - p] [1 + f] t$$


(17.26)

It can be seen immediately that when $E = 0$, the third and fourth terms cancel. The sufficient condition for evasion to occur is then $1 - p - pf > 0$, which is identical to that for the standard model. In addition the payoff is dependent on $t E$, so the solution for $E$ will involve a term in $\frac{1}{t}$. Hence, as $t$ increases, $E$ will fall. Therefore the model does not provide a route to reversing the direction of the tax effect.

Ambiguity Theory applies when there is a lack of precise knowledge of the audit probability so the taxpayer forms a second-order probability distribution over possible audit probabilities. An increase in ambiguity means an increase in dispersion of this probability distribution. Experimental evidence shows people are heterogeneous with respect to preferences over ambiguity. This implies that the sufficient condition for evasion to occur will differ across taxpayers.

Ambiguity can be represented in the payoff by putting weight $\gamma$ on the worst outcome, $\lambda$ on the best outcome, and $[1 - \lambda - \gamma]$ on the expected value given the belief, $p$, about probability of audit. This gives the payoff

$$V = \gamma \min_{x_1, \ldots, x_n} v(x) + \lambda \max_{x_1, \ldots, x_n} v(x) + [1 - \lambda - \gamma] \sum_{i=1}^n p_i v(x_i).$$ 

(17.27)

The sum $\gamma + \lambda$ represents the amount of perceived ambiguity and $1 - \lambda - \gamma$ is the degree of confidence in the belief $p$. Individuals can differ in their personal values of $\gamma$ and $\lambda$. Preferences with $\gamma$ relatively large describe a pessimistic individual who is concerned about what will happen if the very worst outcome arises: as $\gamma$ tends to 1, the preferences are entirely focused on the minimum possible payoff.
In the case of tax evasion there are only two outcomes: $Y^n$ and $Y^c$, with $Y^c < Y^n$. The payoff function for ambiguity then reduces to

$$V = \gamma v(Y^c) + \lambda v(Y^n) + \left[ 1 - \lambda - \gamma \right] \left[ pv(Y^c) + \left[ 1 - p \right] v(s_h) \right]$$

$$= \left[ \gamma + \left[ 1 - \lambda - \gamma \right] p \right] v(Y^c) + \left[ \lambda + \left[ 1 - \lambda - \gamma \right] \left[ 1 - p \right] \right] v(Y^n) \quad (17.28)$$

The form of (17.28) reveals that in this context ambiguity is equivalent to a particular weighting of the probabilities. As the degree of confidence in the belief goes down ($1 - \gamma - \lambda \to 0$), the weights tend to $\lambda$ and $\gamma$. These are individual-specific attributes and are not related to the objective probability, $p$. They are instead entirely determined by the individual attitude toward ambiguity.

When the payoff function is a utility function ($v(Y) = U(Y)$), the implications of the ambiguity model can be compared directly to the standard model. The sufficient condition for evasion to take place can be read directly from (17.28), since the weighting replaces $p$ in the standard model by $\gamma + \left[ 1 - \lambda - \gamma \right] p$. Provided that $\gamma + \left[ 1 - \lambda - \gamma \right] p > p$, the sufficient condition for evasion to take place will be stricter. As far as the effect of the tax rate is concerned the ambiguity model does not change the outcome if the payoff function is a utility. It only changes the weighting, so the comparative statics of the tax rate remain the same.

The idea of more or less ambiguity is an interesting one for the tax compliance decision because it raises important policy questions about the likely effect of making auditing processes more or less transparent. It is possible that increasing ambiguity by making policy less transparent will increase compliance. The model will also yield a different necessary condition for evasion that will be individual-specific, and it can predict higher levels of compliance than the standard model. It does not change the direction of the tax effect.

The models discussed above have a variety of different implications for the evasion decision. Most can affect the sufficient condition for evasion. For the standard model this condition is the same for all taxpayers because they all use the same objective probability in the formation of expected utility. Behavioral models permit the weights to vary between individuals. The weights can also differ from the true probability, so the sufficient condition can be stricter and reduce the predicted number of taxpayers engaged in evasion. What these models cannot easily do is change the direction of the tax effect. This comparative static result is a consequence of the fine being determined by the tax rate. As long as the payoffs depend on the product of $t$ and $E$, an increase in the tax rate will reduce the amount of evasion.
17.9 Compliance and Social Interaction

It has been assumed so far that the decision by any taxpayer to comply with the tax law is independent of what the other taxpayers are doing. This decision is based entirely on the enforcement policy (penalty and auditing) and economic opportunities (tax rates and income). In practice, however, we may expect that someone is more likely to break the law when noncompliance is already widespread than when it is confined to a small segment of the population. This observation is supported by the evidence in table 17.3, which shows that tax compliance is susceptible to social interaction.

The reasoning behind this social interaction can be motivated along the following lines: the amount of stigma or guilt I feel if I do not comply may depend on what others do and think. Whether they also underpay taxes may determine how I feel if I do not comply. As we now show, this simple interdependence among taxpayers can trigger a dynamic process that moves the economy toward either full compliance or no compliance at all.

To see this, consider a set of taxpayers. Each taxpayer has to decide whether to evade taxes. Fixing the enforcement parameters, the payoff from evading taxes depends on the number of noncompliers. On the one hand, the payoff from noncompliance is increasing with the number of noncompliers because then the chance of getting away with the act of evasion increases. On the other hand, the payoff from compliance decreases with the number of noncompliers. The reason can be that you suffer some resentment cost from abiding with the law when so many are breaking the law. Therefore individuals care about the overall compliance in the group when choosing to comply themselves.

It follows from this that the choice of tax evasion becomes more attractive when more taxpayers make the same choice of breaking the law. Because of the way interactions work, the choice of tax evasion becomes more attractive when more taxpayers make the same choice of breaking the law. The aggregate compliance tendency is toward one of the extremes: only the worst outcome of nobody complying or the best outcome of full compliance are possible. This is illustrated in figure 17.8 depicting the payoff from compliance and noncompliance (vertical axis) against the noncompliance rate in the group (horizontal axis). At the intersection of the two payoff functions, taxpayers are indifferent between compliance or noncompliance. Starting from this point, a small reduction in noncompliance will break the indifference in favor of compliance and trigger a chain reaction toward increasing compliance. Alternatively, a small increase in noncompliance triggers a chain reaction in the opposite direction making noncompliance progressively more attractive.
In this situation, how do we encourage taxpayers to abide by the law when the dynamic is pushing in the opposite direction? The solution is to get a critical mass of individuals complying to reverse the dynamic. This requires a short but intense audit policy backed by a harsh punishment in order to change the decisions of enough taxpayers that the dynamics switch toward full compliance. When at this new full-compliance equilibrium, it is possible to cut down on audit costs because compliance is self-sustained by the large numbers of taxpayers who comply. It follows from this simple argument that a moderate enforcement policy with few audits and light penalties over a long period is ineffective. Another interesting implication of this model is that two countries with similar enforcement policies can end up with very different compliance rates. Social interaction can be a crucial explanation for the astoundingly high variance of compliance rates across locations and over time that are much higher than can be predicted by differences in local enforcement policies.

17.10 Conclusions

Tax evasion is an important and significant phenomenon that affects both developed and developing economies. Although there is residual uncertainty surrounding the accuracy of measurements, even the most conservative estimates suggest the hidden economy in the United Kingdom and United States to be at least 10 percent of the measured economy. There are many countries where it is very much higher. The substantial size
of the hidden economy, and the tax evasion that accompanies it, require understanding so that the effects of policies that interact with it can be correctly forecast.

The predictions of the standard representation of tax evasion as a choice with risk were derived and contrasted with empirical and experimental evidence. This showed that although it is valuable as a starting point for a theory of evasion, the model did not incorporate some key aspects of the evasion decision, notably the effects of a basic wish to avoid dishonesty and the social interaction among taxpayers. The analysis was then extended to incorporate both of these issues.

Further Reading

The literature on tax evasion is summarized in:


For a comprehensive survey of the extent of tax evasion in a wide range of countries, see:


The earliest model of tax compliance is:


The economic approach to crime is developed in:


Empirical and experimental evidence is found in:


On the effect of social interaction on law compliance, see:


A game-theoretic model of the strategic interaction between taxpayer and enforcement agency, including honest taxpayers, is in:


On the effect of fiscal corruption and the desirability of a flat tax, see:


The literature on behavioral models of evasion and social interaction is surveyed in:


### Exercises

17.1 Economic efficiency requires that consumers exploit all opportunities to increase their welfare. Does this argument legitimize tax evasion?

17.2 Should the welfare of tax evaders be included in an assessment of social welfare? What if their inclusion implied that tax evasion should not be punished? Would you provide the same arguments for violent crimes?

17.3 Consider a consumer with utility function $U = Y^{1/2}$.

   a. Defining the coefficient of absolute risk aversion by $R_A(Y) = -\frac{\partial^2 U}{\partial Y^2} \frac{\partial U}{\partial Y}$, show that this is a decreasing function of $Y$.

   The consumer is faced with a gamble that results in a loss of 1 with probability $p = 0.5$ and a gain of 2 with probability $1 - p$.

   b. Show that there is a critical value of income $Y^*$ at which the consumer is indifferent between participating in this gamble and receiving income $Y^*$ with certainty. Hence show that the gamble will be undertaken at any higher income but will not at any lower income.

17.4 A consumer with utility function $U = Y^{1/2}$ has to choose between a job that provides income $Y_0$ but no possibility of evading tax and a job that pays $Y_1$ but makes evasion possible. What value of $Y_1$ makes the consumer indifferent between these two jobs? How does a change in the tax rate affect this level of income?

17.5 Given utility function $U = -e^{-Y}$.

   a. Show that the coefficient of absolute risk aversion $R_A(Y) = -\frac{U''}{U'}$ is constant (where $U'$ and $U''$ denote the first and second derivative of $U$ with respect to $Y$, respectively). Show that $U' > 0$ (positive marginal utility of income) and that $U'' < 0$ (diminishing marginal utility of income).

   b. Show that the undeclared income, $Y - X$, is independent of $Y$ for a consumer with this utility function. (*Hint:* For a function $y = e^{f(x)}$, the derivative is $\frac{dy}{dx} = e^{f(x)} \frac{df(x)}{dx}$.)

17.6 For the utility function $U = \log(Y)$:

   a. Show that the coefficient of relative risk aversion $R_A(Y) = -\frac{U''}{U'}$ is constant (where $U'$ and $U''$ denote the first and second derivative of $U$ with respect to $Y$, respectively).

   b. Show that the proportion of income not declared, $\frac{X}{Y}$, is independent of $Y$ for a consumer with this utility function. (*Hint:* Let $X = aY$ in the first-order condition and show that $Y$ can be eliminated.)
Chapter 17: Tax Evasion

17.7 In a state-contingent income space, each person’s indifference curves are generally bowed in toward the origin. Why does this imply that the person is risk averse?

17.8 Does it matter whether $p$ and $F$ are interpreted as subjective or objective quantities? If the revenue service chooses to prosecute celebrities for evasion while fining noncelebrities, does it believe in the objective interpretation?

17.9 A consumer with utility function $U = Y^{1/2}$ determines the amount of income to declare to the tax authority.
   a. Denoting the probability of detection by $p$, the tax rate by $t$, and the fine by $F$, provide an expression for the optimal value of $X$.
   b. For $F = \frac{1}{2}$ and $p = \frac{1}{2}$ show that the declaration $X$ is an increasing function of $t$.
   c. Assume that the revenue authority aims to maximize the sum of tax revenue plus fines less the cost of auditing. If the latter is given by $c(p) = p^2$, graph the income of the revenue authority as a function of $p$ for $Y = 10$, $F = \frac{1}{2}$, and $t = \frac{3}{4}$. Hence derive the optimal value of $p$.

17.10 A consumer has a choice between two occupations. One occupation pays a salary of $80,000 but gives no chance for tax evasion. The other pays $75,000 but does permit evasion. With the probability of detection $p = 0.3$, the tax rate $t = 0.3$, and the fine rate $F = 0.5$, which occupation will be chosen if $U = Y^{1/2}$?

17.11 Use the parameter values from the previous exercise with the modification that pay in the occupation permitting evasion is given by $40,000(1 - n)$, where $n$ is the proportion of the population choosing this occupation. What is the equilibrium value of $n$? How is this value affected by an increase in $t$?

17.12 Consider the simultaneous move game between a taxpayer and a tax inspector. The taxpayer chooses whether or not to underreport his taxable income. The tax inspector chooses whether or not to audit the income report. The cost of auditing is $c > 1$, and the fine (including tax payment) imposed if the taxpayer is caught cheating is $F$ (with $F > c > 1$). With a truthful report the taxpayer has to pay a tax of 1 unit of income. The payoffs are given in the matrix where the first number in each cell denotes the tax inspector’s payoff and the second number is the taxpayer’s payoff. Find the Nash equilibria of this game, considering both pure and mixed strategies.

<table>
<thead>
<tr>
<th></th>
<th>Under report</th>
<th>Truthful report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audit</td>
<td>$F - c, -F$</td>
<td>$1 - c, -1$</td>
</tr>
<tr>
<td>No audit</td>
<td>0, 0</td>
<td>0, −1</td>
</tr>
</tbody>
</table>

17.13 A revenue service announces that it will only audit income declarations below a critical level $Y^*$. If you had an income in excess of $Y^*$, what level of income would you announce? Once declarations are made, will the revenue service act according to its announced auditing plans?
17.14 Consider the game between taxpayer and revenue service described in the payoff matrix below.

<table>
<thead>
<tr>
<th></th>
<th>Audit</th>
<th>No audit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Honest</td>
<td>100, −10</td>
<td>100, 10</td>
</tr>
<tr>
<td>Evade</td>
<td>150, T</td>
<td>5</td>
</tr>
</tbody>
</table>

a. For what value of \( T \) is (Evade, No audit) a Nash equilibrium?
b. Can (Evade, Audit) ever be a Nash equilibrium? What does this imply about the punishment structure?
c. Does a simultaneous move game capture the essence of the auditing problem?

17.15 Is tax evasion just a gamble?

17.16 “Those who follow social customs are fools.” True or false?

17.17 (Crackdowns 1) Police often engage in “crackdowns” on crime, which are intermittent periods of high-intensity policing. Some examples include drunk driving interdictions accomplished using sobriety checkpoints, crackdowns on speeding achieved through a greater announced police presence on certain highways, or crackdowns on drug traffickings. Two features characterize our notion of random crackdowns. First, they are arbitrary, in the sense that they subject certain groups (identified by presence in a particular time or place, or by other observable characteristics), which are not notably different from other groups in criminal propensities, to higher intensity police monitoring. Second, they are publicized; that is, those who are subjected to crackdowns are informed about them before they engage in criminal activity aimed at particular neighborhoods. Discuss the potential costs and benefits of a similar crackdown policy to combat tax evasion. (Hint: You can also use behavioral arguments for nonrational choices.)

17.18 (Crackdowns 2) To illustrate the main idea behind a crackdown policy, consider an example where the tax enforcement agency seeks to minimize tax evasion subject to a fixed budget constraint. Consider a population of 100 taxpayers, half of whom would never evade taxes and half of whom would evade taxes unless it is sufficiently likely that they will be audited. A taxpayer’s propensity to evade tax is unobservable to the tax enforcement agency. The tax enforcement budget is such that only 50 taxpayers can be audited.

a. Random audit: Suppose that taxpayers are audited at random, so that each taxpayer has a probability \( \frac{1}{2} \) of being audited. Then what will be the compliance rate?
b. Crackdown audit: Suppose now that half of the taxpayers are selected based on observable characteristics (e.g., eye color) known to be independent of the propensity to evade taxes (so it is arbitrary). Nevertheless, suppose that all taxpayers in this group are audited and the rest of the taxpayers are not audited. Overall, 50 taxpayers are audited as required by the budget constraint. Then what will be the compliance rate? Compare with part a.
c. What objections would you make to this arbitrary crackdown tax enforcement?

17.19 Assume that 10 percent of the population will always evade paying taxes no matter what anyone else does. Equally 10 percent of the population will never evade paying taxes. The remaining 80 percent are more likely to evade when the proportion evading increases. Prove
that there will be at least one equilibrium level of tax evasion. Show how multiple equilibria can occur. Which equilibria are stable?

17.20 Assume that all taxpayers have the utility function

\[ U(Y^j) = \left( \frac{Y^j}{1 + \alpha} \right)^{1-\alpha} - 1, \]

where \( Y^j \) is income in state \( j \), and \( j = c, n \). When the taxpayer has true income \( Y \), declares income level \( X \), and evasion is detected, the fine is given by

\[ F(X, Y) = (1 + s) t (Y - X). \]

Detection occurs with probability \( p \).

a. Determine the optimal declaration and the resulting level of expected utility.

b. What is the optimal announcement under the cutoff rule?

c. Derive the condition that determines whether a taxpayer prefers the random audit policy or the cutoff rule. Will a taxpayer with true income below \( c \) ever prefer the cutoff rule?

17.21 Consider the optimal audit strategy of a tax authority. All taxpayers have either a low income \( Y_L \) or a high income \( Y_H \), with \( Y_L < Y_H \). They file a tax return, but the rich taxpayers may attempt to underreport. The proportions of taxpayers with high and low incomes are known, but a personal tax return can only be verified through an audit that costs \( c \). There is a constant tax rate on income \( t \) and a fine consisting of a surcharge \( F \) on any underpaid tax. The parameters \( c, t, \) and \( F \) are not chosen by the tax authority.

a. Suppose that the tax authority can pre-commit to its audit policy. What is the optimal audit strategy for the tax authority? Is such a policy credible? Why or why not?

b. If there is a fixed fraction of high-income taxpayers who are known to report truthfully, what could be a credible audit strategy? What is the impact on the equilibrium audit strategy of an increase in the cost of auditing?

17.22 Tax evasion is sometimes described as “contagious,” meaning that an increase in evasion encourages yet further evasion. In such circumstances, is the only equilibrium to have everyone evading?

17.23 An experiment is conducted in three different countries. Participants are told their income levels (in units of the local currency), the tax rate, the probability of detection, and the fine structure. These parameters are the same for all participants. It is found that the amount of income not declared is, on average, different among the countries. Discuss possible explanations for this finding.
In this chapter we step back from the specific models used in chapters 15 and 16 in order to consider taxation from a broader perspective. We have already stressed that the role of taxation is to allow the government to achieve an allocation of resources that is preferred to that which would arise in the absence of intervention. The mixed economy approach we have adopted, by which is meant the combination of competitive trading alongside intervention by the government, is not the only means of organizing economic activity. Many alternatives exist, such as the command economy with direct government resource allocation or economies based on workers’ cooperatives. The analysis we describe contrasts the success of the mixed economy at delivering redistribution with other forms of economic organization. It is quite remarkable that this highly abstract analysis delivers clear and precise conclusions.

All forms of economy can be interpreted as *allocation mechanisms*: they provide a mechanism for allocating the economy’s resources among competing uses. From this viewpoint, the mixed-economy model of taxation on which we have focused represents just one form of an allocation mechanism from among a very broad set of potential allocation mechanisms. This line of reasoning leads to some important questions. For instance, how effective is taxation in achieving the government’s aims? Expressed differently, are there any allocation mechanisms that can achieve those aims better than the mixed economy with taxation? If there are, what is the nature of these alternatives? By considering the issue of the choice between allocation mechanisms, with the mixed economy as one option among many, this chapter will provide clear answers to these questions.

To provide the motivation for the approach we take, it is necessary to clarify some basic issues. As we have noted many times already, government policy aims to resolve the trade-off between efficiency and equity in a way that maximizes its objective function. Improving efficiency is uncontroversial since it implies a Pareto improvement, and therefore will be unanimously supported. In contrast, attaining equity aims implies redistribution, so there will be some consumers that lose from this (as others gain). Those that stand to lose will have an incentive to take actions to reduce the loss. Such changes in action provide limits on the amount of redistribution that can be undertaken. Different allocation mechanisms will face different limits, so the effectiveness of taxation as an
allocation mechanism can be measured against its success in achieving redistribution. More generally, it is possible to determine the limit to which redistribution can be undertaken and to inquire whether taxation can achieve this limit.

The prime obstacle to redistribution is its deleterious effect on the incentive to create wealth. A greater incentive to create and accumulate wealth could be linked with a greater inequality of income. The cause of inequality can take either of two forms. First, inequality can follow simply from luck. Some people are born with higher innate talent than others—such talents are assigned as the outcome of the “genetic lottery.” In addition, among those who start equal, fortune favors the endeavors of some more than others. Many people may think it perfectly legitimate to tax away the economic benefits that arise from luck as a source of inequality. But the ability of the government to do so depends on the possibility of it observing the innate talents and inducing their owner to bring them to full fruition despite the knowledge that the resulting income will be taxed away. For example, even if born with the talent to become a concert pianist, a person might only choose to endure the hours of necessary practice if he can foresee future benefits from doing so. The second cause of inequality is the effort that is applied to gain wealth, given the level of ability. Some people choose to work harder than others, and allowing them to keep the fruits of their own effort seems legitimate. Put another way, there can be little justification for redistributing from those whose incomes are high because of their efforts to those whose incomes are low solely because they choose to supply little effort.

If the government could determine what part of inequality was due to differences in effort and what part due to luck, these differences would present no problems for redistribution. In such circumstances the tax system could be finely adjusted to elicit the correct effort level from all consumers by imposing a harsh punishment for realizing an income that is too low. In practice, effort is not observable and people can pretend to work when they do not. A high income may result from hard work or be due to chance (better natural talent), but without detailed information, the government cannot infer which. Redressing inequality with respect to one dimension but not the other then becomes a difficult task. Higher innate talent must be judged, faute de mieux, indirectly by income that depends on effort. Taxing income is then a blunt tool for attacking the symptom but not the cause. There is as a result a conflict between the redistributive and the incentive effects of the tax system that determines the scope for redistribution. The resolution of this trade-off was the basis of the analysis of optimal income taxation in chapter 16 and the requirement that a tax system has to be incentive-compatible.

The purpose of this chapter is to describe how incentive-compatibility, which is the requirement that the allocation mechanism must be consistent with the incentives for
Chapter 18: Limits to Redistribution

the truthful revelation of information, places limits upon the ability of the government to undertake redistribution. The fact that an optimal allocation mechanism will be consistent with truth-telling is called the revelation principle. We will then explain why taxation is typically the best mechanism, from among the set of incentive-compatible mechanisms, for achieving redistribution. This conclusion is termed the tax principle. This result is interesting given that the notion of taxing income directly was for a long time rejected on the grounds that the collection of the necessary information would be costly and constitute a risk of invasion of privacy. The United Kingdom did not adopt a permanent income tax until the 1870s and the US Congress established the same form of taxation in the 1890s. Even today, the use of income taxation is not widespread and has a negligible role in many developing countries. While income tax is an important instrument of redistribution, there are additional instruments that can be used. Among them are commodity taxes tilted toward the poor. We will then describe under what circumstances differential commodity taxation can usefully supplement the income tax and expand the possibilities for redistribution. This is the issue of the tax mix.

18.2 Revelation Principle

Any economic system is essentially a general allocation mechanism that determines the economic allocation for each possible economic environment. The economic allocation includes the bundle of commodities that each consumer receives, the production activities that are undertaken in the economy, and the public goods that are provided. The economic environment is a specification of each person’s exogenous characteristics (their endowments, needs, tastes, talents, etc.) as well as the production possibilities.

The competitive economy is one form of allocation mechanism. In chapter 2 the competitive economy was described by the preferences and endowments of consumers and the firms’ production technologies; this was the economic environment. Given this information, resources were allocated via the competitive trading mechanism: taking prices as given, each consumer chose demand to maximize utility, and each firm chose output to maximize profit. Prices were adjusted until supply and demand were equal, and then trading took place to allocate commodities. We have also discussed some alternative allocation mechanisms; two examples were the personalized pricing model and the Clarke–Groves mechanism in chapter 6.

These allocation mechanisms have been judged by their success in allocating resources. The First Theorem of Welfare Economics says that the competitive economy is a successful allocation mechanism because it leads (under certain conditions) to a
Pareto-efficient allocation. We have also seen the circumstances under which the competitive economy fails to achieve efficiency, so government intervention can achieve a Pareto improvement. But this is not the end of the story, since an allocation can be efficient and at the same time extremely unfair. The competitive economy is indeed blind to injustice, and the allocation achieved by the mechanism can involve extreme manifestations of inequity in the form of poverty and starvation.

In response to this problem, the Second Theorem of Welfare Economics says that (under certain conditions) any Pareto-efficient allocation can be reached through competitive trading by means of lump-sum redistribution. Lump-sum redistribution does not interfere with the working of the competitive market because it is based on the unalterable characteristics of the economic agents. The problem is that the personal characteristics relevant for redistribution need not be publicly observable (tastes, talents, needs, abilities, etc.). Therefore redistribution has to be based on information that people choose to reveal. These issues were introduced in chapter 13.

The basis of a mechanism is that the agents in the economy make a report of information. This report could, for example, be a statement of their preferences (e.g., the valuation reported in the Clarke–Groves mechanism) or of their demands (e.g., in the competitive economy). Alternatively, it could take the form of a direct statement of their characteristics. Given these announcements, the mechanism determines the economic allocation. For the Clarke–Groves mechanism, the allocation is a choice to provide the public good or not and a set of side payments. In the competitive economy the allocation is determined by the market equilibrium given the announced demands.

A desirable economic allocation is defined as an achievable allocation that maximizes a social choice function. For example, a utilitarian social choice function will select the allocation that maximizes the sum of utilities. The implementation problem is to design the allocation mechanism so that it produces a desirable economic allocation, even when agents attempt to manipulate it to their advantage. Manipulation means reporting false characteristics in order to obtain an advantage—in chapter 6 such manipulation arose in connection with the Lindahl equilibrium. The scope for manipulation arises because some (or all) of the characteristics that are relevant for the proper allocation of resources are not publicly observable.

To be more explicit, consider an economy populated by a set of agents \( I = \{1, \ldots, i, \ldots, n\} \). Each agent, \( i \), is described by \( \theta_i \), which is a list of all the personal characteristics that are economically relevant. Let \( \Theta_i \) be the set of possible characteristics for agent \( i \), so that \( \theta_i \) has to be drawn from the alternatives in \( \Theta_i \). The corresponding profile of characteristics for the set of agents in the economy is denoted \( \theta = (\theta_1, \ldots, \theta_n) \).
Finally let $X$ be the set of economic allocations that are technically possible using the economy’s resources.

Given the characteristics for the agents the optimal allocation, that which is best according to the social choice functions is selected. The social choice function determines an allocation for every possible profile of the agents’ characteristics. Specifically, if the profile of characteristics is $\theta = (\theta_1, \ldots, \theta_n)$, the social choice function selects the allocation $x = f(\theta)$. It must be the case that $x$ is one of the allocations in $X$.

The central idea of mechanism design is the construction of an allocation mechanism that implements the social choice function. That is, if the social choice function selects the allocation $x = f(\theta)$, then given the characteristics $\theta$, the mechanism will result in the allocation $x$ being achieved. It can be seen that this was exactly the process described in chapter 2. The social welfare function selected the Pareto-efficient allocation that maximized welfare, and then competitive activity was combined with lump-sum taxation to decentralize this allocation.

A mechanism is a strategy set for each agent and an outcome function that determines the allocation given a choice of strategies for each agent. The strategy set of $i$ is denoted $S_i$ and describes the choices that are open to agent $i$. The collection of strategy sets for the set of agents is $(S_1, \ldots, S_n)$. Letting the chosen strategy of agent $i$ be $s_i$, the outcome function $g(\cdot)$ determines the allocation $x = g(s_1, \ldots, s_n)$. A mechanism, $M$, is described by the strategy sets and the outcome function so that $M = (S_1, \ldots, S_n, g(\cdot))$. Each agent makes her choice of strategy to maximize her payoff. Often there will be strategic interaction among the agents in the choice of strategies. The type of strategic action involved will determine the appropriate equilibrium concept.

The mechanism $M = (S_1, \ldots, S_n, g(\cdot))$ is said to implement the social choice function $f(\cdot)$ if there is an equilibrium strategy profile $(s_1^*, \ldots, s_n^*)$ for the mechanism $M$ such that $g(s_1^*, \ldots, s_n^*) = x = f(\theta_1, \ldots, \theta_n)$. The strategy profile is an equilibrium if $s_i^*$ is optimal for agent $i$ given the strategy choices of the other agents. This definition applies whether or not there is strategic interaction. These definitions are illustrated in figure 18.1.

Among the very large set of possible mechanisms, there is a particularly interesting set of mechanisms, called direct revelation mechanisms, in which each agent is asked to reveal his characteristics directly (the characteristics can also be called the agent’s type). Given the announcements of characteristics, the allocation is chosen according to the social choice function $f(\cdot)$. More precisely, a direct revelation mechanism is a mechanism in which the strategy set is $S_i = \Theta_i$ for all $i$, so a strategy is an announcement of characteristics, and $g(\theta_1, \ldots, \theta_n) = x = f(\theta_1, \ldots, \theta_n)$ for all possible characteristics $(\theta_1, \ldots, \theta_n)$.
As an example of the distinction between direct and indirect mechanisms, consider the taxation of income. In the model of chapter 16 agents were distinguished by their ability levels. A direct mechanism would involve each agent making an announcement of his ability. In contrast, income taxation is an indirect mechanism in which the announcement is an income level. Naturally, the income level is related to ability but is not directly ability.

The reason for the interest in direct mechanisms is that any indirect mechanism (where an announcement is made that need not be directly about characteristics) can always be replaced with a direct mechanism that achieves the same outcome. This can be easily demonstrated. Consider an indirect mechanism where the announcement of $i$ is $\mu_i$ and the outcome $x = \phi(\mu_1, \ldots, \mu_n)$. Since $\mu_i$ is chosen as the announcement by $i$, there must be a relationship among their characteristics and the value chosen, we write this relationship as $\mu_i = m(\theta_i)$. But then the outcome of the indirect mechanism is $x = \phi(\mu_1, \ldots, \mu_n) = \phi(m(\theta_1), \ldots, m(\theta_n))$, which ultimately depends on the characteristics. A direct mechanism with $f(\theta_1, \ldots, \theta_n)$ chosen to be identical to $\phi(m(\theta_1), \ldots, m(\theta_n))$ will then provide precisely the same outcome. This is illustrated in figure 18.2. In brief, there is nothing an indirect mechanism can achieve that can’t be achieved with a direct mechanism. Only direct mechanisms need therefore be studied.
The interesting issue here is whether mechanisms can be designed such that it is optimal for the agents to announce their true characteristics. In this respect the social choice function $f(\cdot)$ is truthfully implementable in dominant strategies (or incentive-compatible) if the direct revelation mechanism $M = (\Theta_1, \ldots, \Theta_n, f(\cdot))$ has a dominant strategy equilibrium $(s^*_1, \ldots, s^*_n)$ in which truth-telling is optimal, so that $s^*_i = \theta_i$ for any value of $\theta_i$ in the set $\Theta_i$. By a dominant strategy equilibrium, it is meant that truth-telling is optimal for all agents regardless of what the other agents choose to do.

We are now almost in a position to state the fundamental result of this section but one final definition is necessary before this can be done. A truly feasible allocation is defined as one which is both technically feasible (given the endowments and technology of the economy) and which is informationally feasible given that the characteristics on which redistribution is based are not publicly observable. The following theorem is the Revelation Principle that is key to analyzing allocation mechanisms.

**Theorem 18.1**  (Revelation principle Dasgupta, Hammond, and Maskin 1979 and Myerson 1979) The set of allocations that are truly feasible when agents’ characteristics are not observable is the set of allocations that are incentive compatible (or truthfully implementable in dominant strategies).
There are two aspects to this theorem. First, it tells us that there is nothing to be gained by constructing mechanisms in which agents are deliberately led to reveal false information. We can instead confine attention to mechanisms in which the best strategy is to tell the truth. Second, it reveals that we need to only study mechanisms where the equilibrium is found in dominant strategies. There were several previous occasions in the text where we analyzed the Nash equilibrium of games, and it may seem that restricting attention to dominant strategies would reduce the set of allocations that could be achieved. This is not true for the following simple reason: the mechanism must be designed so that truth is the Nash equilibrium strategy for each agent whatever are the characteristics of the other agents. Now consider agent \( i \). Since truth is his Nash equilibrium strategy for all possible characteristics of agents other than \( i \), this implies that it must be chosen in response to any announcement by the others (whether these are truthful or not). Therefore, if it is always a Nash strategy, truth must be a dominant strategy.

### 18.3 The Tax Principle

If optimal lump-sum taxes could be employed, the first-best allocation would be achieved. There is no allocation mechanism that can achieve more than this. However, optimal lump-sum taxes are not incentive-compatible, so they do not function as an allocation mechanism. Other forms of taxation, such as income and commodity taxation, are incentive-compatible but achieve only a second-best allocation. There is therefore a gap between what the economy could achieve at the first-best and what taxation achieves at the second-best. This observation raises the question that this section answers: Can we improve upon the allocation achieved with taxation? In other words, is there any allocation mechanism that can locate the economy somewhere between the second-best achieved by taxation and the first-best?

The answer to this question is provided by the Taxation Principle. Loosely speaking, the Taxation Principle says that we cannot improve upon taxation in the sense that it is not possible to implement additional allocations by using a more complicated mechanism. Taxation allocates goods to individuals with different characteristics by designing the budget set to achieve the desired outcome. The appeal of taxation is that it requires the government to possess only limited information: the government only needs to know the aggregate distribution of characteristics in the population and not the characteristics of any individual. Every individual faces the same tax system and hence the same consumption opportunities, so that false revelation of characteristics
does not change the budget set. Furthermore the point in the budget set chosen by an
individual is, by definition, the best that is available to him and there is no incentive to
cheat.

The Taxation Principle determines what allocations can be reached under information
and incentive constraints, namely all the allocations that can be implemented using a
suitable tax system. Since this principle is the fundamental justification for employing
taxation, it is important to state it precisely. To do this, we need some preliminary
definitions.

Following Hammond (1979), an allocation mechanism is admissible whenever it is
both individually and collectively physically feasible. It is anonymous if the allocation
that results from the permutation of the characteristics of two individuals is just the
permutation of the original allocation (i.e., the allocation mechanism does not depend on
the name of the agent). An anonymous allocation mechanism makes each individual’s
allocation a function only of his own announcement and the announcements of all other
players. A mechanism is straightforward whenever it implements a given allocation in
dominant truth-telling strategies. A dominant strategy is a strategy that an individual,
given his information, is willing to play regardless of what he believes others know
and the way he believes others behave. The Tax Principle can now be stated.

**Theorem 18.2** (Tax principle Guesnerie 1995 and Hammond 1979) In a large econ-
omy, an allocation is implementable by an admissible tax mechanism if and only if it
is implementable by an admissible straightforward allocation mechanism.

The intuition behind this result is quite simple. The first step is to show that if an
allocation is implementable by taxation, then it can be implemented by an admissible
straightforward allocation mechanism. The justification for this result is that a tax
system is an indirect mechanism that has a dominant strategy solution consisting of
the announcements of each individual’s trades. In fact a tax schedule confronts each
individual with the same budget set so that the bundle allocated to them depends only
on their strategy. Thus, by appealing to the Revelation Principle, there exists a direct
mechanism (for individuals to announce their characteristics) that is equivalent to the
(indirect) tax mechanism.

The second step requires the demonstration that any allocation that is implementable
by a straightforward allocation mechanism is implementable by taxation. Since the allo-
cation must be incentive-compatible, each individual must find the consumption bundle
allocated to that individual at least as good as that allocated to any other individual (if
they did not, they would make a false statement of characteristics in order to secure the
alternative). A budget set can then be constructed that contains the consumption bundle allocated to each of the individuals but excludes any preferable bundles. This budget set implicitly defines a tax system. Faced with this budget set, the individuals will always choose the consumption bundle allocated by the mechanism. Taxation therefore achieves the same outcome as the alternative allocation mechanism. Hence tax systems yield, in general, the best allocations that can be achieved in a private information world! The only truly feasible allocations are those that can be decentralized by some nonlinear income and commodity taxes (with everybody facing the same nonlinear tax schedule).

We now illustrate this Taxation Principle in our economy composed of two ability types ($\theta_l < \theta_h$) and two goods ($x$ and $\ell$). This example allows us to clarify the extent to which incentive considerations limit redistribution. We assume, for simplicity, an equal number of individuals of both ability types. Also we assume that the government observes only income $z = \theta \ell$ and consumption $x$. From the preferences over consumption and labor can be derived corresponding preferences over consumption and income given by

$$U(x, \ell) = U\left(x, \frac{z}{\theta}\right).$$

(This construction was explored in detail in chapter 16.) The indifference curves of both types are depicted in figure 18.3 in income/consumption space. They satisfy the single-crossing assumption in the sense that the high-ability individuals have a flatter indifference curve than the low-ability individuals through any point. This is because the high-ability type needs to undertake less work to earn any given income. Utility is increasing when moving in the north–west direction. With no redistribution, both types choose an income-consumption bundle on the 45 degree line tangent to their indifference curve; these are denoted $\{x^n_l, z^n_l\}$ and $\{x^n_h, z^n_h\}$. Each type strictly prefers her own bundle.

We now consider the scope for redistribution when only income and consumption are observable. The impossibility of observing ability opens up the possibility that the high-ability individuals will, if redistribution is pushed too far, choose the allocation intended for the low-ability individuals. If this happens the high-ability individuals are said to mimic the low-ability. To avoid this happening, the allocations are restricted to be incentive-compatible. In this case incentive-compatibility means that the allocations of income and consumption should be such that the high-ability individuals do not prefer the bundle intended for the low-ability. This implies that the allocation for the low-ability type, $\{x_l, z_l\}$, must lie on or below the indifference curve through the allocation
Figure 18.3
Allocation with no redistribution

\{x_h, z_h\} for the high-ability type. Expressed in terms of utility, the allocation must satisfy

\[ U(x_h, z_h) \geq U(x_l, z_l) \]

It can be seen immediately that this incentive constraint rules out complete equalization of utilities. Since \( \theta_h > \theta_l \) and the supply of labor causes disutility, it has to be the case that

\[ U\left(x_h, \frac{z_h}{\theta_h}\right) > U\left(x_l, \frac{z_l}{\theta_l}\right) \quad (18.1) \]

Therefore any incentive allocation must give the high-ability greater utility than the low-ability, so the equality of utility is not feasible when abilities are not observable. This conclusion confirms that the existence of private information places a limit on potential redistribution.

But then what exactly will be the maximum redistribution and the form of the optimal redistribution scheme? To answer this we analyze the problem of choosing the allocations \{x_l^*, z_l^*\} for the low-ability type and \{x_h^*, z_h^*\} for the high-ability type that maximize the utility of the low-ability type subject to the constraints that the allocation is feasible (meaning total income is equal to total consumption) and that the high-ability will not choose to pretend to be low-ability. The solution is displayed in figure 18.4.

The optimal allocation is characterized by the fact that the incentive constraint is binding: the high-ability individuals are just indifferent between their own bundle and the bundle of the low-ability. This means that the allocations \{x_l^*, z_l^*\} and \{x_h^*, z_h^*\}
lie on the same indifference curve for the high-ability type. With an equal number of each type, feasibility requires that the amount taken off each high-ability individual be equal to the amount given to each low-ability individual. In the figure, redistribution away from the high-ability means that the allocation must lie below the 45 degree line, with the vertical distance from the allocation to the 45° line measuring the quantity of their output redistributed to the low-ability. Correspondingly, the allocation of the low-ability must be that same distance above the 45 degree line. The amount taken from the high-ability type is maximized along a given indifference curve when the indifference curve is parallel to the 45 degree line. Thus the allocation \( \{x_h^*, z_h^*\} \) in figure 18.4 maximizes the amount of redistribution along indifference curve \( I_h^* \). The allocation \( \{x_l^*, z_l^*\} \) is then determined by the point at which the distances \( a \) and \( b \) are equal.

According to the Tax Principle, this optimal allocation is decentralizable by a nonlinear income tax schedule. A nonlinear tax function \( T(z) \) induces a nonlinear consumption function relating consumption and income defined by \( x(z) = z - T(x) \). The consumption function that decentralizes the optimal allocation is depicted in figure 18.5 by the dashed curve. This function must have the property that it passes through the two optimal allocations but must not go above either the indifference curve \( I_h^* \) of the low-ability individual or the indifference curve \( I_h^* \) of the high-ability. If the consumption function satisfies these conditions, it will ensure that the optimal allocations are in the budget set but that preferred alternatives are excluded. Provided that the consumption function...
Chapter 18: Limits to Redistribution

Figure 18.5
Decentralization by taxation

is kinked at \( \{x^*_l, z^*_l\} \), this can be achieved. The nonlinear tax function constructed in this way ensures the decentralization.

In the absence of taxation, income would be equal to consumption and this is depicted by the 45 degree line. Where the consumption function lies above the 45 degree line, the tax payment is negative. It is positive where the consumption function is below the line. It is clear from the figure 18.5 that both ability types face the same tax function and will be induced to choose exactly the bundle designated for them. In addition the construction ensures that the gradient of the indifference curve of the high-ability individual is equal to 1. Referring back to chapter 16, this implies that high-ability individuals face a marginal income tax rate of 0 percent so that their income/consumption choice is not distorted at the margin. This was exactly the result we derived for income taxation using a different argument. However, because the bundle designated for the low-ability must lie on the indifference curve of the high-ability, the choice of the low-ability type has to be distorted in order to relax the incentive-compatibility constraint. This is easily seen from the fact that such distortion shifts the bundle of the low-ability further away from the bundle of the high-ability and thereby reduces the incentive of the high-ability to mimic the low-ability.

It has been shown how the allocation arising from any admissible straightforward allocation mechanism can be decentralized by means of a nonlinear tax schedule. More important, a nonlinear tax schedule can attain the maximum redistribution that is possible given the limitations on information. There are no further redistributive instruments
that can expand the possibilities of redistribution. Therefore, having analyzed taxation in earlier chapters, we have exhausted the possibilities for achieving redistribution.

18.4 Tax Mix: Separation Principle

What are the relative merits of taxing (total) income and taxing consumption? Historically the debate over the appropriate base for annual taxation and redistribution was based on the definition of the best taxable capacity with two opposite approaches. The *Haig–Simons approach* claims that total income, defined as labor income plus accrued income from capital, is the best measure of capacity to pay and so should serve as the tax base. Conversely, the *Kaldor approach* claims that annual consumption is the best measure of capacity to pay and so should serve as the tax base. The latter approach was further supported by the moral argument that it is better to tax what people take out of the economy (consumption) rather than what people put into the economy (labor supply generating income). The literature has evolved away from this distinction to assess the merits of different taxes on the sole basis of their impact on the equilibrium outcome and the welfare properties of this equilibrium outcome.

To be sure, tax systems must be based on observable variables. In practice, governments use income and consumption as the basis of taxation, even if they are imperfect measures of individual earning ability. It is important to observe that the distinction between the taxation of consumption and income is diminished when viewed from a lifetime perspective. This is because current savings are future consumption, and thus, as consumption must equal income over the life cycle, a tax on the value of consumption is equivalent to a tax on income. This can be seen from a simple two-period life-cycle model. Labor income in the first period is allocated between consumption and saving:

\[ z_1 = x_1 + s_1. \]  

(18.2)

Consumption in the second-period is financed by labor income and the value of saving with accumulated interest:

\[ z_2 + (1 + r)s_1 = x_2. \]  

(18.3)

Eliminating saving between these budget constraints gives the lifetime budget constraint:

\[ z_1 + \frac{z_2}{1 + r} = x_1 + \frac{x_2}{1 + r}. \]  

(18.4)
Chapter 18: Limits to Redistribution

This demonstrates that the value of lifetime income (left-hand side) is equal to the value of lifetime consumption (right-hand side). From this perspective, it does not matter whether taxes are levied on income or on the value of consumption.

Further insight can be obtained by directly comparing an income tax and a consumption tax. Let the tax rate on labor income be $t_\ell$ and the tax rate on interest (capital income) be $t_c$. The budget constraints in the two-period model with this income tax system are

$$ (1 - t_\ell) z_1 = x_1 + s_1, \quad (18.5) $$

and

$$ (1 - t_\ell) z_2 + (1 + (1 - t_\ell) r) s_1 = x_2. \quad (18.6) $$

Combining these budget constraints over the life cycle obtains

$$ (1 - t_\ell) z_1 + \frac{(1 - t_\ell) z_2}{1 + (1 - t_\ell) r} = x_1 + \frac{x_2}{1 + (1 - t_\ell) r}. \quad (18.7) $$

The lifetime budget constraint in (18.7) shows why an income tax distorts the intertemporal consumption decision. With taxation the price of period 1 consumption relative to period 2 consumption is $1 + r$. With the income tax the relative price is $1 + (1 - t_\ell) r$, so, if $t_\ell > 0$, the relative price of period 1 consumption is reduced. This is why it is often said that an income tax provides a disincentive to save, but whether saving in fact falls will depend on the resolution of income and substitution effects. Now consider a consumption tax at rate $\tau$. The two budget constraints are

$$ z_1 = (1 + \tau) x_1 + s_1, \quad (18.8) $$

$$ z_2 + (1 + r) s_1 = (1 + \tau) x_2, \quad (18.9) $$

which combine to the life-cycle budget constraint

$$ z_1 + \frac{z_2}{1 + r} = (1 + \tau) x_1 + \frac{(1 + \tau) x_2}{1 + r}. \quad (18.10) $$

As can be seen from (18.10), the consumption tax does not affect the relative price of consumption, so it creates no distortion in the intertemporal consumption decision. The key point we wish to make is that the income tax is distortionary only if the tax on interest income is positive. There is no reason why this must be so, and many tax systems allow for tax-free saving vehicles. Now set $t_c = 0$. It can be directly observed from (18.7) and (18.10) that if the labor income tax and the consumption tax satisfy

$$ 1 - t_\ell = \frac{1}{1 + \tau}, \quad (18.11) $$

then.


then the two lifetime budget constraints are identical. This argument demonstrates an important equivalence result: an income tax with exemption for interest on savings is equivalent to a consumption tax. The result is important because it shows that the issue at the heart of debate over the choice of tax base is not really about taxing income or consumption but instead is about the tax treatment of interest income (or capital income more generally). Many of the arguments about the choice of tax base obscure this basic point.

In the simple models we have used so far in this chapter, we have assumed that there is a single consumption good. When there are two or more consumption goods, the commodity taxes levied on them need not be uniform. As we showed in chapter 15, the Ramsey rule says that, when there is no income tax the tax on each commodity should be inversely related to the elasticity of demand. We now consider how this conclusion can be modified when income and commodity taxes can be simultaneously employed.

The central question is whether commodity taxes should be differentiated when combined with a nonlinear income tax. If they are not, a uniform commodity tax can be easily replicated by the income tax (provided there is an adequate deduction for savings). There is a sense in which differentiated commodity taxation can usefully supplement income taxation by reducing the distortion in the labor/consumption choice induced by income tax. If we tax commodities that are substitutes for work and subsidize those that are complements, we can encourage people to work more and thus reduce the work-discouraging effect of the income tax. The optimal differentiation depends on how the preferences for the goods vary with labor supply. Turning this argument around, if the preference between commodities does not vary with labor supply, then there seems to be no argument for differential commodity taxes.

Preferences over commodities are independent of labor supply if the utility function is separable. What separable means is that utility can be written as \( U = U(u(x), \ell) \), so that the marginal rate of substitution between any pair of goods depends only on the subutility \( u(x) \) and is independent of the labor supply. We now demonstrate that if labor and goods are separable in utility, then there is no need to supplement an optimal nonlinear income tax with differential commodity taxation.

The result that commodity taxation is not needed with separability between labor and goods is easily demonstrated in the two-ability model we have been using in this chapter. We now interpret \( x_i \) as a vector describing the consumption levels of \( n \) goods, \( x_i = (x_{i1}, \ldots, x_{in}) \). We already know that the optimal allocation \( \{x_i^*, \ell_i^*\} \) is constrained by the incentive of high-ability individuals to mimic the low-ability. That is,
the downward incentive constraint, \( U \left( u(x^*_h), \frac{z^*_l}{p} \right) \geq U \left( u(x^*_l), \frac{z^*_l}{p} \right) \), is binding at the optimum because it requires less work for a high-ability individual to earn the income \( z^*_l \) of a low-ability individual, \( \ell_h = \frac{z^*_l}{p} < \ell_l = \frac{z^*_l}{p} \). The only difference between the two types at any given level of income arises from the difference in the amount of work. This feeds into a utility difference, \( U \left( u(x^*_l), \ell_l \right) \neq U \left( u(x^*_h), \ell_h \right) \), when the high-ability mimic the low-ability. With separable preferences, such a difference does not affect the indifference curves between commodities (since \( u(x^*_h) \) is the same for both types), so we cannot use differential taxation to separate the types (i.e., to make the consumption bundle of the low-ability individual less attractive to high-ability individual). The fact that differential taxes do not help in relaxing the incentive constraint renders their use unnecessary.

With a nonseparable utility function, \( U(x, \ell) \), the indifference curves between commodities differ, and we can tax more heavily the good that the high-ability person values more in the low-ability consumption bundle. This reduces the incentive for the high-ability to mimic the low-ability. Figure 18.6 illustrates how differential commodity taxation (changing prices from \( p \) to \( p' \)) can be used to make the consumption bundle of the low-ability type less attractive for the high-ability type. The change in prices from the budget constraint labeled \( p \) to the budget constraint labeled \( p' \) does not affect the utility of the low-ability (the budget constraint pivots around their indifference curve \( I_l \)), but it causes a reduction in utility of the high-ability (shown by the new budget

\[ I^1_h \]
constraint changing the location of the choice from initial indifference curve $I^1_h$ to the lower indifference curve $I^2_h$). This reduction in utility for the high-ability type relaxes the incentive constraint.

Finally, there is a compliance argument for using commodity taxes over income taxes. The indirect consumption taxes are a more effective method of collecting taxes from individuals who do not (fully) report income or pay tax under the income tax. The reason is that those who do not report their income pay no income tax, but eventually will pay consumption taxes on their purchases.

### 18.5 Capital Income Tax

A capital income tax is a tax on the return to savings. It results in future consumption being taxed more heavily than current consumption. Thus the merits of capital income tax boil down to determining if it is optimal to tax consumption in different periods at different rates.

The separation principle of the previous section can be used to answer this question. If one interprets the different consumption goods as consumption in different periods, then the separation result can be interpreted as recommending exclusively a progressive tax on labor income or, assuming no initial assets and interpersonal transfers, a progressive tax on lifetime consumption. Even if the separability condition is not satisfied, it does not necessarily follow that future consumption should be more heavily taxed than current consumption, as would be the case with a capital income tax. In fact a capital income subsidy (negative tax) to encourage saving could well be the optimal policy instrument.

The most striking result in the literature on capital income taxation is that with an infinite horizon (no end period to the economy) and a labor income tax, the capital income tax should converge to zero in the long run. We demonstrate this result in chapter 24. The result is quite general and the intuition also very simple. As already noted, taxing capital income is, in effect, taxing future consumption more heavily than current consumption. A positive capital income tax in every year imposes a very high effective tax rate on consumption in the long run, say 20 or 30 years into the future. It is implausible that in the very long run we would want to continue to tax capital income as it would imply that we should tax consumption more and more heavily as we move into later, and later, periods. Thus, no matter what the individual preferences are, the capital income tax will have to converge to zero in the long run. This finding has led some economists to favor taxing labor income but not capital income, or taxing consumption
by taxing labor income minus net savings. However, the conditions required for the result are too restrictive to provide a basis for abandoning the capital income tax.

The static separability principle with multiple persons was derived on the basis that each household has just one form of labor supply (leaving aside that there are different labor supplies and different labor incomes in different periods of life). Even if consumption were separable from leisure at any given point in time, there is no presumption that consumption and leisure are separable at different points in time. So one might, for example, believe that some age-related labor income tax schedule would make sense. But, more fundamentally, the static separability principle ignores the fact that we observe labor and consumption decisions early in life before those later in life. Thus we might choose to adopt a tax schedule in future periods that is conditioned on current decisions. Alternatively, a government may have an incentive to announce that there is to be no capital income tax and then deviate from the announcement later on in order to tax capital income once it has been accumulated.

In a dynamic setting with a population that has differing earnings abilities, a capital income tax can improve redistributive possibilities. The logic is that the existence of capital income reduces the possibility of redistributing through the labor income tax because those of high ability have greater incentive to earn less in response to redistributive taxation in the presence of additional capital income. Taxing capital income weakens the incentive to earn less, and so relaxes the incentive compatibility constraint and expands the scope for redistribution.

Even if capital income taxes distort consumption choices over time, there are some good redistributive motives for including them in the tax base. Nevertheless, it is not possible to use the theory to claim that both labor and capital income should be taxed equally. There are practical arguments why this might be beneficial, such as the prevention of tax avoidance through the transformation of labor income into capital income, but no theoretical demonstration confirms that this is optimal. So, returning to the historical Haig–Simons claim, there is no reason why the tax base should be total income defined as the sum of labor income and capital income. In the United States, the marginal tax rates on capital and labor income are related to each other, but in Europe, the Nordic countries have provided a lead with a dual income tax where there is a progressive income tax on labor and a uniform tax rate on capital income. Also deductions from labor income for pension saving is a widespread practice used to encourage people to save more for their retirement. In chapter 3 we reviewed the many reasons for people to make the “wrong” savings choice.
18.6 Non–Tax Redistribution

The principal implication of the previous analysis, as reflected in the tax principle, is that society cannot improve redistributive possibilities by using nonfiscal instruments. The question is then why nonfiscal forms of redistribution are so widely used. Governments frequently provide goods such as education or health services at less than their costs, which may be viewed as redistributional policies. One may expect that a cash transfer of the same value would have more redistributional power than such in-kind transfer programs. This is mistaken. There are three reasons why such transfers in-kind may be superior to cash transfers as achieved through standard tax-transfer programs.

One reason is political. Rather than voting for a tax-transfer scheme that benefits the poor, I might impose my own preferences and vote for providing certain services such as education, even though the recipient would have preferred another use. Voters may support redistributional policies if they are in-kind but not if in cash with use at the discretion of the recipient. Political considerations may dictate that many government provision programs like education, pensions, and basic health insurance ought to be universal. Without universality the programs would not have the political support required to be adopted or continued. For instance, public pensions and health care would be far more vulnerable politically if they were targeted to the poor and not available to others. It should be noted that it does not follow that if a government program is universal, then there is no redistribution. First, if the program is financed by proportional income taxation, the rich will contribute more than the poor to its cost. Second, even if everyone contributes the same amount to the program, it is possible that the rich will not make use of the publicly provided good. To take an example, consider the public provision of basic health care made available to everyone for free and financed by a uniform tax on all households. There exists a private health care alternative with higher quality but at some cost. Since the rich can afford better quality health care, they will choose to use private health care even though free public health care is available. These rich households still pay their contribution to the public program, and thus the poor households derive a net benefit from this cross-subsidization.

Provision in-kind can also be justified by a self-selection argument. We have shown that what ultimately sets the limit to redistribution is the possibility of it becoming advantageous for higher ability people to earn lower incomes by expending less effort and thereby paying taxes (or receiving transfers) intended for the lower ability people. The limit to redistribution is reached when a person of a given ability would be just as well off earning the income of someone with lesser ability. The self-selection argument
is that anything that makes it less attractive for people to mimic those with lesser ability will extend the limit to redistribution. As seen in the previous section, if the government could supplement a nonlinear income tax with differential commodity taxes, it will do so in certain cases (i.e., if the shares of income devoted to the consumption of each good are not independent of the amount of leisure).

The use of in-kind transfers to supplement standard taxation can be optimal. This is an efficiency-based argument. Basically a given degree of redistribution can be more efficiently achieved by using in-kind transfers as a supplement to income taxes. The argument, unlike the separability argument just used for differential commodity taxes, relies on differences in preferences among different income groups. Consider two individuals who differ not only in their ability but also in their health status. Suppose that lower ability also means poorer health, so that the less able also spend more on health. Then both income and health expenditures act as a signal of ability. It follows that the limits to redistribution can be relaxed if transfers are made partly in the form of provision of health care (or, equivalently, by full subsidization of health expenditures). The reason is simply that the more able individuals (with less tendency to become ill) are less likely to claim in-kind benefits in the form of health care provision, than they would be to claim cash benefits. To take another example, suppose that the government is considering redistributing either in cash or in the form of low-income housing. All households, needy or otherwise, would prefer the cash transfer. However, few non-needy households would want to live in low-income housing since they can afford better housing. Thus self-selection arises through which the non-needy drop out of the housing program and only the needy choose to take it up. In short, transfers in-kind invite people to self-select in a way that reveals their neediness. When need is correlated to income-earning ability, the in-kind transfers can relax incentive compatibility and selection constraints, thereby improving the government’s ability to redistribute income.

A final justification for transfers in-kind is time consistency. In this case the argument relies on the inability of a government to commit to its future actions. This is different from the earliest Strotz (1956) argument for government time inconsistency as it does not arise from a change in the government’s objective over time (e.g., because of elections) nor from the fact that the government is not welfaristic or rational. The time consistency problem arises from a perfectly rational government that fully respects individual preferences but does not have the power to commit to its policy in the long run. The time-consistency problem is obvious with pensions. If households expect the government to provide a basic pension to those with too little savings, then their incentive to save for retirement consumption and provide for themselves is reduced. Anticipating this fact, the government may prefer to provide public pensions instead.
A related time consistency problem can explain why transfer programs, such as social security, education, and job training are provided in-kind. If a welfaristic government cannot commit not to come to the rescue of those in need in the future, potential recipients will have little reason to invest in education or to undertake job training because, if they do not, the government will help them out anyway. Again, the government can improve both economic efficiency and redistribution by making education and job training available at less than their cost, rather than making cash transfers of equivalent value.

18.7 Conclusions

We addressed in this chapter the important question of what limits the extent of redistribution in a society. We began with the very general concept of selecting the allocation mechanism that achieved the best possible outcome given the constraint that some relevant individual characteristics are private information. This information restriction requires that the characteristics be reported, either directly or indirectly, as part of the allocation mechanism. The competitive economy with taxation is just one example of such a mechanism. Then we explored whether there are any better mechanisms.

When there are no informational problems, lump-sum transfers combined with competitive economic activity can take the economy to the first-best outcome. No allocation mechanism can perform better than this. When there is private information, the situation is considerably changed: optimal lump-sum taxes are not incentive-compatible. All other forms of taxation (income taxes, commodity taxes etc.) are distortionary and so only allow the attainment of allocations that are strictly worse than the first-best. These observations raise the possibility that there may be allocation mechanisms that can achieve outcomes that are strictly better than those arising from distortionary taxation. Perhaps surprisingly, the Tax Principle shows this is not the case. The taxation of observable variables is as good as any other allocation mechanism.

These are very general and very powerful results. What they do is offer a fresh perspective as to how successful is the competitive economy as an allocation mechanism. When combined with intervention via taxation, there is no other allocation mechanism that can better it. But it should always be recalled that the intervention has to be well-intended and that politics can also shape the outcome and further constrain the possibilities for redistribution. For example, with political determination of redistribution, increasing the number of the rich can actually leave the poor worse off: as the rich become politically more influential, the extent of redistribution from each rich person...
may fall by so much as to more than offset the increase in the relative number of rich redistributing to the poor.

The arguments developed in the chapter are also based on welfaristic objectives (i.e., social objectives that fully respect individual preferences). Recently there has been growing emphasis on nonwelfaristic objectives that include equality of chance and capability, fairness, poverty alleviation, and inequality. The real challenge is then not just the value judgments associated with the different objectives but also that welfaristic and nonwelfaristic objectives are typically in conflict. As the social choice theory has convincingly shown, resolving the conflict involves a value judgment that ultimately is left to the political process.

Further Reading

For a more complete theoretical treatment of the information constraint to redistribution, see:


Two excellent reviews of the central issues that arise with redistribution are:


Two excellent reviews of the central issues that arise with taxation and the choice of the best tax base are:


The relevance of indirect taxation to supplement income tax is in:

Part VI: Taxation


The self-selection argument for in-kind redistribution is in


Estimation of the incentive effect of welfare programmes is in:


Government time consistency problem is in:


Strotz, R. H. 1956. Myopia and inconsistency in dynamic utility

---

**Exercises**

**18.1** Distinguish between a direct revelation mechanism and an indirect revelation mechanism. Briefly describe how a labor income tax schedule can be represented by a direct revelation mechanism.

**18.2** Describe briefly the Tax Principle and its main implications for redistribution.

**18.3** Provide arguments for departing from the tax principle and complementing taxes by in-kind transfers. Give a specific example for each special case.

**18.4** The separation principle claims that consumption taxes are redundant in the presence of nonlinear labor income tax. What are the restrictions imposed on preferences for this result to apply? Discuss the implication of preferences that are not separable between consumption and leisure (e.g., for some goods that are complements to leisure).
18.5 Explain why consumption tax may be a better instrument for redistribution than labor income tax when labor income is imperfectly observable (e.g., because of self-report and tax evasion).

18.6 Under what circumstances can we apply the Atkinson–Stiglitz separation result to support the claim that there must be zero taxation on capital income?

18.7 Consider a consumer who lives for two periods. The consumer’s preferences over current consumption \( C_0 \) and future consumption \( C_1 \) are given by the utility function \( U(C_0, C_1) = C_0 C_1 \). The consumer’s income in the current and future periods are given by \( w_0 = 40 \) and \( w_1 = 45 \), respectively. Assume the interest rate is 50 percent (i.e., \( r = 0.5 \)).

a. Suppose that initially there is no tax.
   i. Determine the consumer’s lifetime wealth.
   ii. Write the lifetime budget constraint. What is the slope of the budget constraint?
   iii. Graph the lifetime budget constraint putting \( C_0 \) on the horizontal axis and \( C_1 \) on the vertical axis. Show the consumer’s endowment point \((w_0, w_1) = (40, 45)\) on the graph.
   iv. Solve the utility maximization problem to find the optimal current and future consumption \( (C^*_0, C^*_1) \) and saving \( s^* \). Show the optimal consumption bundle on the same graph. Is the consumer a borrower or a lender?

b. Suppose that interest income is subject to a tax at the rate of 75 percent (i.e., \( t = 0.75 \)), and interest payments by borrowers are deductible from taxable income. Repeat i–iv in part a.

18.8 Consider the following two-period utility: \( V(C_1, C_2, L) = U_1(C_1, L) + \delta U_2(C_2) \), where \( C_1 \) and \( C_2 \) are first- and second-period consumption, \( L \) is labor in the first period, \( \delta \) is a discount factor (with \( \delta < 1 \) reflecting either time preference or survival probability or both), and \( V, U_1, \) and \( U_2 \) are utility functions with standard properties (i.e., increasing and concave in consumption and decreasing and convex in labor). With no taxation, the individual \( i \) with productivity \( \theta_i \) and initial endowment \( \omega_i \) faces the following intertemporal budget constraint: \( \omega_i + \theta_i L = C_1 + (1 + r)^{-1} C_2 \), where \( r > 0 \) is the interest rate, \( \omega \) is an initial endowment, and \( \theta \) is the ability level or the wage rate.

a. Suppose that \( \omega = 0 \) and consider only two ability levels \( \{\theta_l, \theta_h\} \) with \( \theta_l < \theta_h \) and relative proportions \( n_l + n_h = 1 \). Assume that individuals only differ in their ability, and suppose further that the government can only observe labor income \( Y = \theta L \). Show that the optimal income taxation requires zero marginal tax on the high-ability income.

b. Assuming further weak separability between consumption and labor, show that the optimal taxation requires no taxation of interest (on savings).

18.9 Consider the previous exercise and assume weak separability between consumption and labor. We now modify this setting by introducing some additional features. Consider differential endowments \( \omega_i \).

a. If initial endowments \( \omega_i \) are observables what is the implication for the optimal income tax?

b. If initial endowments \( \omega_i \) are observables and positively correlated with ability \( \theta_i \), what is the implication for the optimal income tax?

18.10 We now modify the previous exercise by assuming differential discount rates \( \delta_i \) (but equal endowments).
a. Suppose that discount rates are observables. What is the implication for the optimal income tax?
b. Suppose that observable discount rates are correlated with ability. What is the implication for the optimal income tax? What is the implication for capital income tax?

18.11 Airlines do practice overbooking flights (number of passengers who have booked is greater than the number of seats available on the flight). When there is an overbooked flight, the airlines offer passengers a reward in money and free travel for giving up their seats. What are the effects of these arrangements on redistribution?

18.12 "Bottle laws" require a payment at the purchase of every bottle or can of drink for environmental motives. The payment is a refundable deposit. Consumers get a refund if they return the bottle for recycling. In practice, many consumers do not collect the refund. They prefer to dispose their bottles or cans for others to take for recycling and the refund. In effect, nonrecyclers treat the payment as a green tax for recycling and cleanliness. What is the distribution effect of bottle laws?

18.13 (Brandeis tax) The “Brandeis ratio” is the average income of the richest 1 percent (which includes the billions earned by the lucky few) relative to the median income. In the United States the Brandeis ratio has grown more than proportionated. Ian Ayres has a post in 2011 at Freakonomics with a chart showing that in 1980, one-percenters, on average, made 12.5 medians, but in 2006 (the latest year for which data are available) the average income of the richest 1 percent was a whopping 36 medians. Ayres then makes the case for a "Brandeis tax." His recommendation is to tax inequality instead of taxing income and to increase taxes on the richest 1 percent if and only if the Brandeis ratio is going up. Discuss this proposal in the light of the Tax Principle.

18.14 (Reciprocity) Consider the following variation of the ultimatum game. The proposer offers a wage \( w \in [0, \pi] \) to the responder who can reject or accept the offer \( w \). If the offer is rejected, both players receive zero payoff. If the offer is accepted, the responder can make a costly effort \( e \in [0, \ell] \). The payoff for the proposer is then \( v^P = \mu e - w \) while the responder payoff is \( v^r = w - c(e) \), where \( \mu \) is the (constant) marginal utility of effort for the proposer and \( c(e) \) is the strictly increasing cost of effort for the responder.

a. Under the standard assumptions of rationality and selfishness, what is the subgame perfect equilibrium of this game?
b. Suppose now that the payoff to the proposer is given by \( v^P = (\mu - w) e \), so that the proposer cannot make losses by proposing high wages (unlike in part a). What is the (subgame perfect) equilibrium under the assumptions in part a?
c. Experimental evidence reports a positive correlation between the effort choice and the wage offer. This suggests some reciprocal effort in response to a generous wage effort. Discuss the differences from part a and part b.
VII  MULTIPLE JURISDICTIONS
19 Fiscal Federalism

19.1 Introduction

Fiscal federalism is the division of revenue collection and expenditure responsibilities among different levels of government. Most countries have a central (or federal) government, state or county governments, town councils, and, at the lowest level, parish councils. Each level has restrictions on the tax instruments it can employ and the expenditures that it can make. Together they constitute the multi-leveled and overlapping administration that governs a typical developed country.

The central government can usually choose whatever tax instruments it pleases, and although it has freedom in its expenditure, it usually focuses on national defense, the provision of law and order, infrastructure, and transfer payments. The taxation powers of state governments are more restricted. In the United Kingdom, local governments can levy only property taxes; in the United States, both commodity and local income taxes are allowed. Their responsibilities include education, local infrastructure, and the provision of health care. Local governments provide services such as rubbish collection and maintenance of parks. The responsibility for the police and fire service can be at either the state or local level. Levels of government are connected by overlapping responsibilities and the transfer payments made among them.

The issue of fiscal federalism is not restricted to the design of government within countries. Indeed the recent impetus for the advancement of this theory has been issues involving the design of institutional structures for the European Union. The progress made toward economic and monetary integration has begun to raise questions about subsidiarity, which is the degree of independence that individual countries will maintain in the setting of taxes. Such arguments just involve the application of fiscal federalism, albeit at a larger scale.

These observations lead to a number of interesting economic questions. First, why should there be more than one level of government? Using the logic of economic reasoning, multi-level government can only be justified if it can achieve something that a single level cannot. Explanations of what multi-level government can achieve must revolve around access to information and how the information can be best utilized. If this argument is accepted, and it is explored in detail below, then a second question arises. How are the functions of government best allocated among the levels? A brief sketch of how this allocation works in practice has already been given: Is this outcome efficient, or does it reflect some other factors?
The next section of the chapter considers the rationale for multi-level government, focusing on the availability of information. An overview of arguments in favor of multi-level government are then given. This is followed by a more detailed analysis of some of the key issues, and the concept of an optimal structure is investigated. The issue of accountability and decentralization is analyzed next. The essential elements of interregional risk sharing are then presented, and the distinction between insurance and redistribution is discussed. Empirical evidence is provided on the extent of decentralization by countries and functions, and the main determinants of the observed decentralization.

19.2 Arguments for Multi-level Government

The economic arguments for having government are founded on the two principles outlined in chapter 5. If there is market failure, the government can intervene in the economy to increase efficiency. It can also intervene to improve equity, regardless of whether the economy is efficient. These arguments justify intervention; to justify multi-level government, the case must be made that objectives of efficiency and equity are better served by a combination of local and central government.

If the correct decisions are made about the level of public good provision and about taxes, then it does not matter at which level of government they are taken. Provided that there are no resources wasted in overlapping responsibilities, the number of levels of government is a matter of indifference. A case for multi-level government must therefore be sought in differences in information and political process that allow some structures to achieve better outcomes than others.

Decisions should be taken at a national level if they involve public goods that serve the entire economy. The obvious example here would be defense, whose benefits cannot be assigned to any particular community within the economy. It is common to argue that all citizens should have the same access to the law, have the same rights under the law, and be subject to the same restrictions. An application of this equity argument supports a legal system that is organized and administered from the center. Given the central provision of these services, it is natural to support them through centrally organized taxes.

Other public goods, namely the local public goods of chapter 7, benefit only those resident within a defined geographic area. The level of supply of these goods could be determined and financed at the national level, but there are three arguments to suggest a lower level decision is preferable.
First, determination at the local level can take account of more precise information available on local preferences. In this context, local ballots and knowledge of local circumstances may help in reaching a more efficient decision. Second, if a decision were to be made at the national level, political pressures may prevent there being any differentiation of provision among communities, whereas it might be efficient to have different levels of local public good provision in different areas. Finally the Tiebout hypothesis investigated in chapter 7 argued that if consumers have heterogeneous preferences, then efficiency requires numerous communities to form and offer different levels of public good provision. This will not be possible if decisions are taken at a national level.

If these arguments for determining and providing public goods at a local level are accepted, then it follows almost inevitably that financing should be determined at the same level. To do otherwise, and to set a tax policy that is uniform across the economy, would result in transfers among regions. Those regions choosing levels of public good provision that are high relative to their tax base would not generate sufficient tax revenue to finance their provision, while those with relatively low levels would raise excessive revenue. These deficits and surpluses result in implicit transfers among regions. Such transfers may not be efficient, equitable, or politically acceptable.

Similar arguments can be repeated for the other roles of government such as the control of externalities and some aspects of the reallocation of income. The provision of services, even if they have the nature of private goods (e.g., garbage disposal), can be subject to the same reasoning. This process suggests that different levels of government should be constructed to ensure that decisions are made at the most appropriate level. There is a limit, though, in that duplication of effort and wasted resources should be avoided. The precise design of the structure of government then emerges from the trade-off between increasing the number of levels to ensure that decisions are made at the correct point and reaping the resource benefits of having fewer levels.

The general observations made above can now be refined into more detailed arguments. We first explore the costs of imposing uniformity and then consider positive arguments for decentralization.

19.2.1 The Costs of Uniformity

Uniform provision of public goods and services by all jurisdictions will only ever exactly meet the needs of the entire population when preferences are homogeneous. When they are not, any form of uniform provision must be a compromise between
competing levels of demand. As such, it must involve some loss in welfare relative to differentiated provision.

This argument can be illustrated by considering an economy where there are two groups of consumers who have different tastes for the economy’s single public good. The public good is financed by a uniform income tax. Denote the two groups by $A$ and $B$, and assume that members of group $B$ have a relatively stronger preference for the public good than those of group $A$, taking into account the higher tax rate that this implies. The utility levels of the two groups can then be graphed against the quantity of public good provision as in figure 19.1. The preferred choices of public good provision are denoted as $G_A$ and $G_B$ (with $G_A < G_B$). Now consider the choice of a uniform level of provision, and let this level be $G_0$. Assume that this level lies between $G_A$ and $G_B$ (the argument easily extends to cases where it lies outside these limits). The loss of welfare to society is then given by $L = n_A[u_A(G_A) - u_A(G_0)] + n_B[u_B(G_B) - u_B(G_0)]$ compared to what would be achieved if each group could be supplied with its preferred quantity.

The value of the loss can be minimized by setting the location of $G_0$ so that the marginal benefit for group $B$ of having more public good, $n_Bu_B'(G_0) > 0$, just offsets the marginal loss of group $A$, $n_Au_A'(G_0) < 0$, but the essential point is that the loss remains positive. Furthermore the loss increases the more widely dispersed are preferences and the more members there are of each group.

**Figure 19.1**
Costs of uniformity
This analysis shows how uniformity can be costly in terms of forgone welfare. A policy of uniformity can then only be supported if the costs of differentiation exceed the benefit. Such costs could arise in the collection of information to determine the differentiation and in the administration costs of a differentiated system. These arguments will be explored further below. The next section considers, however, the limit of the benefits that can arise from differentiation.

19.2.2 The Tiebout Hypothesis

Although the costs of uniformity as illustrated above are indisputable, it is another step to show that decentralization is justified. The route to doing this is to exploit the Tiebout hypothesis that was analyzed in section 7.6 in connection with the theory of local public goods. The exact same arguments are applicable here. Each community can be treated as an independent provider of local public goods. If the consumers in the economy have heterogeneous tastes, then there will be clear advantages to jurisdictions having different levels of provision. Each can design what it offers (its tax rates, level of provision, and type of provision) to appeal to particular groups within society. By choosing the jurisdiction in which to live (i.e., by voting with their feet), the consumers reveal their tastes for public goods. In the absence of transactions costs, or other impediments to freedom of movement, an efficient equilibrium must ensue.

The limits to this argument explored in the context of local public goods are also applicable here. Transactions costs are relevant in practice, and the problem of optimally dividing a finite population into a limited number of jurisdictions will arise. The fact that the first-best allocation will not be achieved does not necessarily undermine the preference-matching argument for decentralization. There are clearly still benefits to decentralization, even when this cannot be taken to the level required by the Tiebout hypothesis. Starting from a uniform level of services that is too little for some consumers and too much for others, a move away from this uniform level by some jurisdictions must benefit some of the consumers. This way even restricted decentralization can be increasing in efficiency. This argument can be easily understood from figure 19.1.

The Tiebout hypothesis shows the benefits achievable by decentralization. Although these are not fully realizable, a limited version of the same argument suggests that even restricted decentralization will improve on uniform provision.
19.2.3 Distributive Arguments

The regions that constitute any economy are endowed with different stocks of resources. Some may be rich in natural resources, such as oil and coal, and others may have a well-educated workforce with high levels of human capital. Such differences in endowments will be reflected in disparities in living standards among regions.

The ability to differentiate public good provision among the regions then allows more accurate targeting of resources to where they are required. This is an equity argument for not having uniform provision. Decentralized decision-making allows each region to communicate its needs to the center and permits the center to make differential allocations to the regions.

This process will be designed to offset the differences in living standards caused by endowments. Typically there will be no compensation for differences in preference for public good provision such as giving more to a region wishing to spend more on public goods. This form of redistribution among regions is called an *equalization* formula and will be explored in chapter 20 when discussing intergovernmental grants.

19.3 Optimal Structure: Efficiency versus Stability

The previous arguments have explored a number of advantages of fiscal decentralization. These have involved both efficiency and equity aspects. The issue that remains is what is the optimal structure or the correct number of levels of administration. The difficulty that arises here is that the optimal division may differ from one public good to another. The examples in the introduction have discussed how fire services are organized at a very local level, education at a higher level, and defense at an even higher one. There are many other public goods provided by the federal government. If each were to be allocated at the correct level of decentralization, this would imply an equally large number of levels of government.

Each level of government brings with it additional costs. These involve all the factors that are necessary to provide administration. Buildings, staff, and equipment will all be required, as will elections to choose politicians. The politicians will also require compensation for the time devoted to political activity. These costs are replicated each time an additional level of government is introduced.

Consequently introducing further levels of government is not costless. The choice of the optimal degree of decentralization must take these costs into account and balance them against the benefits. From such a process will emerge the optimal structure. This
will depend on the relative sizes of costs and benefits but is most likely to result in a
level of decentralization such that some decisions are taken at a higher level than would
be best if decentralization were costless.

This argument is now illustrated in a simple spatial model that trades off scale econo-
 mies against diversity of preference. The point of departure is that centralized decision-
 making produces a “one-size-fits-all” outcome that does not reflect the heterogeneity of
tastes. The uniform provision follows from political economy considerations preventing
centralized majority voting from allocating different levels of public goods to different
districts. It is only by decentralizing the majority voting at the district level that it is
possible to differentiate public good provision, but at some cost of duplication.

Suppose that there is one public good that can be provided either at the federal level or
at the regional level. We model the federation as the line segment $[0, 1]$, with points on
the line representing different geographical locations. The public good can be located
anywhere along the line, and individuals are characterized by their ideal location for
the public good. With central provision the public good is located at $\frac{1}{2}$, the midpoint
of the line segment. The farther away from this point individuals are located, the less
they like the public good provided at the federal level. This is shown in figure 19.2.

Alternatively, provision can be decentralized. Each region is then represented by an
interval on the line segment. Region $L$ is the left-interval left $[0, \frac{1}{2}]$ and region $R$ the
right-interval left $[\frac{1}{2}, 1]$. Individuals are assumed to be uniformly distributed, so both
regions are of equal size. With decentralized provision the public good is located at the
midpoint of each interval, that is, $\frac{1}{4}$ in the left-region and $\frac{3}{4}$ in the right-region. There is
a fixed cost $C$ (per capita) of providing the public good at the central level, and due to
duplication the cost is $2C$ with decentralized provision (i.e., the number of individuals
across whom the cost of public good provision is spread is reduced by one-half).

\begin{figure}
\centering
\begin{tabular}{ll}
0 & Centralized 1 \\
\hline
\end{tabular}
\end{figure}

\begin{figure}
\centering
\begin{tabular}{ll}
Region $L$ & Region $R$ \\
0 & Decentralized 1 \\
\hline
\end{tabular}
\end{figure}

\textbf{Figure 19.2}
Centralization and decentralization
The utility function of each individual $i$ under centralization is

$$u^c_i = \left(1 - \alpha \left|\frac{1}{2} - i\right|\right) - C,$$  \hspace{1cm} (19.1)

and under decentralization

$$u^d_i = \begin{cases} 
\left(1 - \alpha \left|\frac{1}{4} - i\right|\right) - 2C & \text{for } i \in \left[0, \frac{1}{2}\right], \\
\left(1 - \alpha \left|\frac{3}{4} - i\right|\right) - 2C & \text{for } i \in \left[\frac{1}{2}, 1\right].
\end{cases}$$ \hspace{1cm} (19.2)

where $\left|\frac{1}{2} - i\right|$ and $\left|\frac{1}{4} - i\right|$ denote the distance between the public good's location and the ideal location of individual $i$, respectively, under centralization and decentralization for an individual $i$ located in the left-region ($\left|\frac{1}{2} - i\right|$ for one located in the right-region). The rate at which utility decreases with the distance is given by the parameter $\alpha$.

We can define when decentralization is socially optimal by considering the trade-off between duplicating the cost of public good provision against taking provision closer to individual preferences. The socially optimal solution maximizes the sum of all individual utilities. Since individual utilities differ only in the distance to the public good location (because of the equal cost sharing), the sum of utilities will depend on the average distance. Under the centralized provision the average distance from $\frac{1}{2}$ is, due to the uniform distribution, just equal to $\frac{1}{4}$. Note that this distance is actually minimized by locating the public good at the midpoint $\frac{1}{2}$. Decentralization brings the average distance in either region down to $\frac{1}{8}$. This positive effect of decentralization has to be balanced against the extra cost $C$ of providing the public good twice. Therefore decentralization is the optimal solution if and only if the extra cost $C$ is less than the advantage of reducing the average distance by $\frac{1}{4} - \frac{1}{8}$ evaluated at the rate $\alpha$, at which utility falls with distance, that is, $C \leq \alpha \left[\frac{1}{4} - \frac{1}{8}\right] = \frac{\alpha}{8}$. In summary, decentralization is optimal if and only if $C \leq \frac{\alpha}{8}$.

These arguments show that the optimal amount of decentralization will be achieved when the benefits from further decentralization, in terms of matching the diversity of tastes, outweigh the cost of differentiating the public good provision. In practice, the extent of decentralization is determined through the political process. Using this basic model, we can now illustrate the tendency for majority voting to lead to excessive decentralization.

In order to look at the incentive for decentralization under majority voting, we assume that decentralized provision prevails when a majority of voters are favorable in at least one region. This assumption is innocuous given the symmetry between regions: if there
is a majority in favor of decentralization in one region, there must also be an equivalent majority in the other region. We concentrate on the incentive of the left-region for decentralization.

The majority in the left-region is formed by those who are either to the left or to the right of the individual at the regional midpoint $\frac{1}{4}$. It is easily seen that if this central individual prefers decentralization, then all those to his left also have the same preferences because they are located further away from the centralized provision but share the cost equally. Therefore, there will be a majority in favor of decentralization in the left-region if the decisive individual $i = \frac{1}{4}$ prefers decentralization, that is, if

$$u_i^f = \left(1 - \alpha \left|\frac{1}{2} - \frac{1}{4}\right\right) - C \leq u_i^d = \left(1 - \alpha \left|\frac{1}{4} - \frac{1}{4}\right\right) - 2C. \quad (19.3)$$

It can therefore be concluded that decentralization is a majority voting equilibrium if and only if $C \leq \frac{\alpha}{4}$.

This result suggests that there will be excessive decentralization under majority voting because the critical cost level under majority voting is higher than the critical cost level for optimality. In particular, for any cost $C$ between $\frac{\alpha}{8}$ and $\frac{\alpha}{4}$ majority voting leads to decentralization ($C \leq \frac{\alpha}{4}$), although it is not socially optimal ($C > \frac{\alpha}{8}$). Therefore, under majority voting there is excessive decentralization: voters who are located at the extremes have an incentive to support decentralized provision to get a public good closer to what they want, but the democratic process does not internalize the negative externalities imposed on voters located in the center who suffer from the extra cost and little or worse preference matching.

### 19.4 Accountability

Politicians may pursue a range of different objectives. At times they may be public-spirited and dedicate themselves fully to furthering public interest. But they may also pursue their own interests, even if these differ from those of their constituents. Some may want to derive private gains while in office or actively seek perks of office. Some may extend clientilistic favors to their families and friends. But the most important way in which they can act against the best interests of their constituents is by choosing policies that advance their own interests or those of special groups to which they are beholden.

A government is accountable if voters can discern whether it is acting in their interest and sanction them appropriately if they are not, so that incumbents anticipate that they
Part VII: Multiple Jurisdictions

will have to render accounts for their past actions. The problem is then to confront politicians with a trade-off between diverting rents and losing office or doing what voters want and getting re-elected. In this view, elections can be seen as an accountability mechanism for controlling and sorting good from bad incumbents. By “good incumbent,” we mean someone who is honest, competent, and not easily bought off by special interests.

The standard view of how electoral accountability works is that voters set some standard of performance to evaluate governments, and they vote out the incumbent unless these criteria are fulfilled. However, elections do not work well in controlling and sorting politicians. There are severe problems in monitoring and evaluating the incumbent’s behavior in order to make informed decisions about whether or not to re-elect. Voters face a formidable agency problem because they are inevitably poorly informed about politicians’ behavior and type. Moreover, the electoral sanction (pass or fail) is such a crude instrument that it can hardly induce the politicians to do what the public wants.

From this perspective it might be reasonable to try to organize competition among politicians in order to control them. In this respect the Brennan and Buchanan (1980) view is that decentralization is an effective mechanism to control governments’ expansive tendencies. The basic argument is that competition among different decentralized governments can exercise a disciplinary force and break the monopoly power of a large central government. Comparing performance in office among different incumbents helps in sorting good types from bad types as well as controlling the quality of their decisions. Hence one votes against an incumbent if his performance is bad relative to others, in order to induce each incumbent to behave in the public interest.

To see the logic of the argument, consider a simple example. Suppose that the circumstances under which politicians make decisions can be good (state $a$) or bad (state $b$). Governments decide to adopt policy $A$, which is better for their constituents in the good state $a$, or policy $B$, which is better in the bad state $b$. Governments need not pursue the public interest and can rather advance their own interests by choosing policy $A$ in state $b$ and policy $B$ in state $a$ to get some private gains (e.g., a rent $r > 0$).

Suppose that politicians place a value $V$ on being re-elected and that this value satisfies $V > r$. The payoff matrix is shown in figure 19.3: the first number in each cell is the government payoff, and the second number is voters welfare. If the government is re-elected, it gets the extra value $V$. The government knows the prevailing conditions (i.e., whether $a$ or $b$ has occurred), but all that citizens observe is their current welfare.

To induce politicians to act as well as they can under this information structure, voters must set the correct re-election rule. If voters set the standard the incumbent must meet...
in order to be re-elected too high (e.g., committing to vote for the incumbent if the welfare level is at least 3), then the incumbent cannot be re-elected whatever he does if conditions turn out to be bad (state $b$). Consequently the incumbent has the incentive to obtain the rent $r$ and leave office. Alternatively, if the voters set the standard for re-election lower, say at 1, the incumbent will be able to divert rent when conditions happen to be good (state $a$) and be re-elected by giving voters less than what they could obtain. Then voters are in a quandary because whatever they decide to do, the politicians will sometimes escape from their control and divert rent.

Suppose now that the electorate can compare the outcome of its incumbent with other incumbents (in different constituencies) facing exactly the same circumstances. Then from the observation of outcomes elsewhere, voters can potentially infer whether the prevailing conditions are good or bad and thereby get the most they can under either condition. The information will be revealed if there is at least one government that chooses a different policy from that of the others. The voting rule then becomes:

When conditions are good, vote for the incumbent if the outcome is at least 3. When conditions are bad, vote for the incumbent if the outcome is at least 1. Otherwise, vote the incumbent out. Hence a government facing good conditions $a$ knows that by choosing the appropriate policy $A$, it will be re-elected for sure and get $V$, which is more than the rent $r$ it can get by choosing $B$ and being voted out. In turn a government facing bad conditions $b$ knows that by choosing $B$, it will be re-elected and get $V$, which is better than what it would get by adopting the wrong policy $A$ to get the rent $r$ but no chance of being re-elected. Therefore, comparing the performance of their incumbent with other incumbents facing similar circumstances, voters can gain increased control.
over their politicians and deduce what is attributable to circumstances as opposed to
government actions.

Another argument for why decentralization should lead to greater efficiency and
accountability is that a central decision maker does not need to please all jurisdictions to
get re-elected but simply a majority of them. However, this argument is usually balanced
against the fact that the value of holding office is larger in a centralized arrangement
and thus politicians are more eager to win election, which in a conventional political
agency model may increase accountability and efficiency.

19.5 Risk-Sharing

Interregional insurance is fundamentally about sharing risk among a group of regions so
that no region bears an undue amount of risk. Because of this risk-sharing, insurance can
arise even when all parties are risk averse. What is necessary for interregional insurance
to be possible is that the risks the parties bear are, to some degree, independent of each
other. That is, when one region suffers a loss, there are other regions (or group of regions)
that do not suffer a loss. While such independence is usually true of almost all individual
risks for which standard forms of insurance exist (fires, car accidents, sicknesses, etc.),
it is less obvious at the regional level.

There are some fundamental principles in mutual insurance. First, risk-sharing is
more effective, the broader the basis is on which risks are pooled. This is a consequence
of Borch’s theorem on mutual insurance. Second, it is more advantageous for any region
to engage in mutual insurance with other regions when risks are negatively correlated
across regions. Third, there must be a degree of symmetry across regions. The reason is
that with an asymmetric regional distribution of risks, some regions will systematically
and persistently subsidize others. The distributional considerations will then dominate
insurance aspects. Fourth, risk-sharing arrangements require reciprocal behavior: a
region with a favorable shock accepts to help out other regions if it can reasonably
expect that those regions will in turn help it out in bad circumstances. With voluntary
insurance, participants are free to opt out at any time, so there is also the possibility of
a risk-sharing agreement without commitment.

19.5.1 Voluntary Risk-Sharing

A model of voluntary insurance between two regions when aggregate income is constant
is as follows: In each period two regions, indexed \( i = \{a, b\} \), receive an income \( y_i \), and
one region is randomly selected to receive a monetary gain $\Delta > 0$. Each has the same probability $\frac{1}{2}$ of receiving this gain, and the total income is fixed at $Y = y_a + y_b + \Delta$. The regional income distribution is given in figure 19.4.

With constant aggregate income, risk-aversion requires the smoothing of regional income across states of nature. Optimal risk-sharing arrangements imply full insurance, which requires that the region receiving the gain $\Delta$ transfer one-half of this gain to the other region. Denoting such a transfer by $t^*$, then $t^* = \frac{\Delta}{2}$. Therefore the gain is equally shared among regions, and regional income is constant.

Let $u_i(x)$ denotes the utility of region $i$ from disposal income $x$. Then it is readily seen that both regions are better off with such an optimal risk-sharing arrangement, since

$$u_a \left( y_a + \frac{\Delta}{2} \right) \geq \frac{1}{2} u_a(y_a + \Delta) + \frac{1}{2} u_a(y_a), \quad (19.4)$$

$$u_b \left( y_b + \frac{\Delta}{2} \right) \geq \frac{1}{2} u_b(y_b + \Delta) + \frac{1}{2} u_b(y_b). \quad (19.5)$$

Without commitment, complete risk-sharing is not guaranteed. We must take into account the possibility that the region receiving the gain may refuse to transfer some of the gain to the other region. A risk-sharing agreement without commitment must be “self-enforcing” in the sense that no region has an incentive to defect unilaterally from the agreement. To be self-enforcing, the risk-sharing arrangement must be such that the expected net benefits from participating is at any time larger than the one-time gain from defection (by not making the transfer when called upon). If full insurance is not
possible, it is still possible to design partial insurance by limiting transfers when the participation constraint is binding.

Let $t_i$ be the transfer made by region $i$ to the other region when region $i$ receives the gain $\Delta$. On receiving the gain $\Delta$, region $i$ can trade off the immediate gain of defecting by refusing to make the transfer $t_i$, against the cost of being excluded from any future insurance arrangement, and to bear regional income variation alone. Taking region $a$, the gain from defection when receiving $\Delta$ is

$$u_a(y_a + \Delta) - u_a(y_a + \Delta - t_a).$$

(19.6)

The cost of losing insurance in the next period (which will be discounted at rate $\delta < 1$ when compared to the gain) is

$$\left[ \frac{1}{2} u_a(y_a + \Delta - t_a) + \frac{1}{2} u_a(y_a + t_b) \right] - \left[ \frac{1}{2} u_a(y_a + \Delta) + \frac{1}{2} u_a(y_a) \right].$$

(19.7)

The participation constraint holds if the gain from future insurance exceeds the cost of defecting. Comparing the two values and rearranging, we find that region $a$ has no incentive to defect if

$$\left[ 1 + \frac{\delta}{2} \right] u_a(y_a + \Delta - t_a) + \frac{\delta}{2} u_a(y_a + t_b) \geq \left[ 1 + \frac{\delta}{2} \right] u_a(y_a + \Delta) + \frac{\delta}{2} u_a(y_a).$$

(19.8)

and similarly region $b$ has no incentive to defect if

$$\left[ 1 + \frac{\delta}{2} \right] u_b(y_b + \Delta - t_b) + \frac{\delta}{2} u_b(y_b + t_a) \geq \left[ 1 + \frac{\delta}{2} \right] u_b(y_b + \Delta) + \frac{\delta}{2} u_b(y_b).$$

(19.9)

We can draw several implications from this simple model of risk-sharing without commitment. First, the time horizon will influence the amount of mutual insurance that is sustainable. Indeed the value attached to continued insurance depends on the discount rate (reflecting the time horizon). At one extreme when $\delta \to 0$ (extremely short horizon), the value of future insurance is zero and regions always defect. No insurance is possible. At the other extreme when $\delta \to 1$ (very long horizon), the value of future insurance is sufficiently high that full insurance is possible ($t_i = \Delta$).

And by a continuity argument, for intermediate discounting values $\delta \in (\underline{\delta}, \overline{\delta})$, with $0 < \underline{\delta} < \overline{\delta} < 1$, only limited insurance is possible ($t_i < \frac{\Delta}{2}$). Therefore the expected time horizon limits the amount of risk-sharing. For values $\delta \geq \overline{\delta}$, complete risk-sharing can be achieved. For intermediate values $\underline{\delta} < \delta < \overline{\delta}$, there is partial risk-sharing. And for values $\delta < \underline{\delta}$, no risk-sharing is possible.
The second implication is that the level of risk-sharing that regions can achieve increases with risk aversion. The reason is that regions put more weight on the gain from long-term insurance against the short-term gain from defecting. This is immediately seen from the participation constraints. Indeed the income distribution on the left-hand side of equations (19.8) and (19.9) is less uncertain than the income distribution on the right-hand side, which makes the participation constraints more likely to be satisfied under increased risk aversion.

A third implication concerns the effect of income inequality. Intuition would suggest that mutual insurance is more likely if regions are ex ante identical and that regional inequality limits the scope for insurance. But this is not true. The reason is that risk-sharing redistributes ex post from the region with a positive shock to the other region, but it does not redistribute ex ante from the rich to the poor regions. More surprisingly, the increased inequality, while maintaining constant the aggregate income, can improve insurance. To see this, start from income equality \( y_a = y_b \). Using the participation constraint, we can calculate the level of risk-sharing that is possible. Then we increase \( y_a \) and reduce \( y_b \) by the same amount. The participation constraints are then affected because income levels influence the demand for insurance. It is then possible to show that for some standard utility functions the amount of risk-sharing has increased with inequality.

### 19.5.2 Insurance versus Redistribution

In practice, interregional insurance is organized in a federation through federal taxes and transfers. The effect of such a federal tax system is to redistribute income from high- to low-income regions. By pooling income risk across the regions, the federal tax system provides insurance against region-specific shocks. However, to the extent that there is ex ante income inequality among regions, federal taxes also provide ex ante regional redistribution. We ignore the stabilizing effect of federal taxation, which refers to the possibility of smoothing shocks over time (between bad years and good years). The insurance motive for the federal tax system is explicitly recognized in many countries. For instance, in the United Kingdom part of the tax system is actually called National Insurance. To appreciate the amount of insurance federal taxes can provide, it is necessary to disentangle redistribution from insurance components. Redistribution acts on the initial income distribution, while insurance responds to income shocks (either permanent or temporary).

Assume that region \( i \)’s income at time \( t \) is subject to permanent shock \( \psi^i_t \) and temporary shock \( \eta^i_t \). Both shocks are assumed to have mean zero. Thus regional income at
time $t$ can be written as

$$y_t^i = y_0^i + \sum_{s=1}^{t} \psi_s^i + \eta_t^i.$$  \hfill (19.10)

Suppose that the federal tax system taxes all regions' incomes at the same rate $\tau$ and redistributes total tax revenue as a uniform transfer to all regions. It follows that region $i$ at time $t$ pays taxes $\tau y_t^i$ and receives transfers from the federation based on the average tax payment

$$\tau E_i \left[ y_0^i + \sum_{s=1}^{t} \psi_s^i + \eta_t^i \right] = \tau E_i \left[ y_0^i \right] = \tau y_0^i.$$  \hfill (19.11)

The regional income after tax and transfer is

$$x_t^i = y_0^i + \tau [y_0^i - y_t^i] + (1 - \tau) \left[ \sum \psi_s^i + \eta_t^i \right].$$  \hfill (19.12)

The income change can be decomposed into an insurance part and a redistribution part as follows:

$$x_t^i - y_t^i = \tau [y_0^i - y_t^i] - \tau \left[ \sum \psi_s^i + \eta_t^i \right].$$  \hfill (19.13)

Using this decomposition, it is interesting to measure the extent of insurance provided by federal taxation in practice. Empirical studies for the US federal tax system clearly suggest the presence of intranational insurance. Although there is disagreement about the exact magnitude of the insurance, all studies find that the redistribution effect largely dominates the insurance effect. They also find that insurance is rather modest, in the sense that it cannot smooth more than a ten cents on a dollar change in state income caused by asymmetric shocks.

### 19.6 Hard and Soft Budgets

In a federal system the central government can impose a budget constraint upon local government. This can be done by fixing the funding that is provided by the centre. The budget constraint of the local government may be soft or hard. A hard budget constraint is one that must be met. This requires the federal government to make a credible commitment about the level of funding. In contrast, a soft budget constraint arises when...
the amount of funding can be manipulated by the local government. Empirical research 
has provided evidence that local governments do face soft budget constraints. 

Soft budget constraints are a source of inefficiency since they encourage strategic 
behavior among local governments. The logic is that the funds transferred to local 
governments come from a common pool of resources controlled by federal government. 
This gives an incentive for local governments to make excessive claims on this common 
pool since the cost is shared. This is an application of the standard argument that a 
common-access resource will be overexploited. One way that this can occur is for 
local governments to borrow excessively in order to make current expenditures in the 
knowledge that they will receive a bailout from the federal government. 

The effect of the soft-budget constraint can be illustrated by considering the behavior 
of two local governments in a two-period setting. In period 1 the local governments 
determine how much public good to provide, how much to borrow in period 1, and what 
local tax rate to set. Once decided, the public good provides services in both periods, 
and there is no additional public good provision in the second period. After these 
choices are made, the federal government sets a uniform tax on each region in period 
1 and a differentiated transfer to each local government in period 2. The equilibrium 
for this two-stage game can be found by backward induction: the choices of the federal 
government are determined conditional on those of the local governments, then local 
government choices are determined. 

The local governments are denoted A and B. The residents of region j have aggregate 
income $y_j^t$ in each period. This income is taxed at the local rate $\tau_j^t$. Income left after tax 
is used to purchase quantity $z_j^t$ of a private good in period t, so $z_j^t = [1 - \tau_j^t] y_j^t$. The 
local government of region j collects tax revenue $\tau_j^t y_j^t$ in each period, pays a tax $T$ to 
the federal government in period 1, and receives a transfer $\Gamma_j^t$ in period 2. It borrows $b_j^t$ 
in period 1 and repays $[1 + r]b_j^t$ in period 2. The budgets of the local government are 

$$g_A^t = \tau_A^t y_A^t - T + b_A^t, \quad (19.14)$$ 

$$[1 + r]b_A^t = \tau_A^t y_A^t + \Gamma_A^t. \quad (19.15)$$ 

Equation (19.15) can be used to show that 

$$z_A^t = y_A^t + \Gamma_A^t - [1 + r]b_A^t. \quad (19.16)$$

The preferences of the residents in region j are represented by the quasi-linear utility function 

$$U^j = U(z_1^j) + U(z_2^j) + g^j. \quad (19.17)$$
Using the budget constraints, the local residents obtain utility

\[ U^j = U((1 - \tau^j)y^j) + \delta U(y^j + \Gamma^j - [1 + r]b^j) + \tau^j y^j - T + b^j, \quad j = A, B. \tag{19.18} \]

In the second stage of the game the federal government takes the policy choices \( \{\tau^A, b^A, \tau^B, b^B\} \) of the local governments as given when it optimizes. The budget constraint for the federal government is

\[ \Gamma^A + \Gamma^B = 2T, \tag{19.19} \]

and its objective function is

\[ W = U^A + U^B. \tag{19.20} \]

Using (19.18), we write the necessary conditions for the choice of the transfers as

\[ \delta U'(y^j + \Gamma^j - [1 + r]b^j) - 1 = 0, \quad j = A, B. \tag{19.21} \]

From the necessary conditions it can be calculated that

\[ \frac{d\Gamma^j}{db^j} = 1 + r, \tag{19.22} \]

so an increase in local government borrowing raises the transfer it receives in the second period to compensate.

At the first stage the local governments make their choices. Substituting the federal government budget into (19.18) gives the local government objective

\[
\max_{\{\tau^j, b^j\}} U((1 - \tau^j)y^j) + \delta U(y^j + \Gamma^j - [1 + r]b^j) + \tau^j y^j - \frac{\Gamma^A}{2} - \frac{\Gamma^B}{2} + b^j, \tag{19.23} \]

\[ j = A, B. \]

The key feature of (19.23) is that the apparent cost of the transfer, \( \Gamma^\prime \), to the region is halved since the cost of financing it is shared between two local governments. The necessary condition for choice of \( b^j \) is

\[ \delta U'(y^j + \Gamma^j - [1 + r]b^j) \left[ \frac{d\Gamma^j}{db^j} - [1 + r] \right] - \frac{1}{2} \frac{d\Gamma^j}{db^j} + 1 = 1 - \frac{1 + r}{2}. \tag{19.24} \]

The first-order condition is positive provided that \( r < 1 \), so for these quasi-linear preferences the local government has an incentive to continue borrowing without limit.
central message is that there is a strategic motive to borrow excessively since half the cost of borrowing is financed by the other local government.

This ordering of moves implies a soft-budget constraint. The federal government has an incentive to equalize private good provision across regions in the second period. It does this by providing transfers in period 2, but since the cost of the transfer is shared between regions, this gives each region an incentive to borrow excessively and over-provide the public good in period 1. The regions do this with the knowledge that they will be bailed out by the federal government in period 2. The softening of the budget constraint imposes a cost on all regions and distorts the level of public good provision. Such soft-budget constraints will always be a problem when the federal government has economic and political reasons for providing bailouts.

The same model can be used to analyze a hard-budget constraint. This can be done by reversing the order of moves in the game so that federal government sets the first-period tax and the second-period transfer before the local governments make their choices. Under this ordering of moves the choice problem of the local governments is still described by (19.23), but $\Gamma^j$ is taken as given. The necessary condition for $b^j$ is

$$- [1 + r] \delta U'(y^j + \Gamma^j - [1 + r] b^j + 1 = 0,$$

so the level of borrowing is efficient given the level of the transfer.

It is therefore socially preferable for the federal government to impose a hard-budget constraint. The difficulty in doing this in one of credibility. What will the federal government do if a local government tests the hard-budget constraint by excessive borrowing? It may be politically difficult not to offer a bailout. The position of the federal government is made worse if there is uncertainty, since it may be hard to distinguish between borrowing that is excessive because of a poor realization of second-period income for a region and borrowing that is deliberately excessive in expectation of a bailout. With hidden information there is a clear moral hazard problem.

19.7 Evidence on Decentralization

19.7.1 Decentralization around the World

The degree of decentralization of government activity can be measured in several different ways. Oates (1972) distinguishes three measures of fiscal decentralization: (1) share of total public revenue collected by the central government, (2) share of the central government in all public expenditures (including income redistribution payments), and (3) share of the central government in current government consumption expenditures.
Table 19.1
Share of central government expenditure in total expenditures

<table>
<thead>
<tr>
<th>Countries</th>
<th>1975</th>
<th>1985</th>
<th>1995</th>
</tr>
</thead>
<tbody>
<tr>
<td>Developed</td>
<td>0.57</td>
<td>0.49</td>
<td>0.46</td>
</tr>
<tr>
<td>Russia</td>
<td>n.a.</td>
<td>0.61</td>
<td>0.63</td>
</tr>
<tr>
<td>Latin America</td>
<td>0.76</td>
<td>0.71</td>
<td>0.70</td>
</tr>
<tr>
<td>Asia</td>
<td>0.79</td>
<td>0.74</td>
<td>0.72</td>
</tr>
<tr>
<td>Africa</td>
<td>0.88</td>
<td>0.86</td>
<td>0.82</td>
</tr>
<tr>
<td>World</td>
<td>0.76</td>
<td>0.68</td>
<td>0.64</td>
</tr>
</tbody>
</table>


The first measure based on revenue collection raises the problem that the center may collect revenue for regions. It underestimates the degree of decentralization to the extent that regions get back substantial portions of the revenue collected at the central level. The second measure, including income redistribution payments, also underestimates the degree of decentralization because the redistribution of income is mostly the role of central governments regardless of how decentralized a country is. The same argument applies for excluding defense spending, which is the other public good that is uniformly provided by central governments. So the more appropriate measure is the concentration of total government current consumption. Such information is readily available in the rich database at Brown University, in which total government expenditures are the consolidated sum of all expenditures at different government levels. Consolidation matters to prevent double-counting of intergovernmental grants and transfers.

Table 19.1 shows the patterns of decentralization around the world and suggests some clear trends. Developed countries are generally more decentralized. Latin America countries decentralized mostly during the period 1980 to 1995. However, government consumption in Latin America remains substantially more centralized, with spending at the central level close to 70 percent against central spending less than 50 percent in developed countries. African countries are the most centralized and display little decrease in centralization (with almost all government spending occurring at the central level). Developed countries exhibit the most substantial decreases among all regions in centralization. The world level average (involving up to 48 countries) also reveals a general trend toward greater decentralization, with the central spending share declining from 75 percent in 1975 to 64 percent in 1995.
19.7.2 Decentralization by Functions

Previously the degree of decentralization was shown to differ quite substantially among countries. It is instructive to measure the decentralization of public expenditures by function to see whether this is consistent with normative advice. From a normative point of view decentralization is desirable when the need to tailor spending to local preferences dominates the possible economies of scale and interregional spillovers.

The Government Finance Statistics of the IMF contain the data for breaking down government activities by functions and levels. All local expenditures refer to expenditures of the state, regional and provincial governments. Table 19.2 indicates the functional decentralization of government activity country by country. Housing and community amenities are the most decentralized, with an average of 71 percent, followed closely by education and health with an average of 64 percent each. The least decentralized are the expenditures for social security and welfare with an average of 18 percent. This is consistent with the normative view that income redistribution is better achieved at the central level.

Table 19.2
Local expenditures as a percentage of total government expenditure by function, 1995 to 1999

<table>
<thead>
<tr>
<th>Country</th>
<th>Education</th>
<th>Health</th>
<th>Social welfare</th>
<th>Housing</th>
<th>Transport</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>72</td>
<td>48</td>
<td>10</td>
<td>77</td>
<td>85</td>
<td>50</td>
</tr>
<tr>
<td>Canada</td>
<td>94</td>
<td>96</td>
<td>31</td>
<td>74</td>
<td>90</td>
<td>60</td>
</tr>
<tr>
<td>Denmark</td>
<td>45</td>
<td>95</td>
<td>55</td>
<td>29</td>
<td>51</td>
<td>56</td>
</tr>
<tr>
<td>France</td>
<td>37</td>
<td>2</td>
<td>9</td>
<td>82</td>
<td>42</td>
<td>19</td>
</tr>
<tr>
<td>Germany</td>
<td>96</td>
<td>28</td>
<td>21</td>
<td>93</td>
<td>57</td>
<td>38</td>
</tr>
<tr>
<td>Ireland</td>
<td>22</td>
<td>48</td>
<td>6</td>
<td>70</td>
<td>43</td>
<td>25</td>
</tr>
<tr>
<td>Netherlands</td>
<td>33</td>
<td>5</td>
<td>14</td>
<td>79</td>
<td>35</td>
<td>26</td>
</tr>
<tr>
<td>Norway</td>
<td>63</td>
<td>78</td>
<td>19</td>
<td>87</td>
<td>31</td>
<td>38</td>
</tr>
<tr>
<td>Russia</td>
<td>83</td>
<td>90</td>
<td>10</td>
<td>96</td>
<td>68</td>
<td>39</td>
</tr>
<tr>
<td>Spain</td>
<td>71</td>
<td>63</td>
<td>6</td>
<td>93</td>
<td>62</td>
<td>36</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>68</td>
<td>0</td>
<td>20</td>
<td>40</td>
<td>61</td>
<td>26</td>
</tr>
<tr>
<td>United States</td>
<td>95</td>
<td>43</td>
<td>31</td>
<td>32</td>
<td>75</td>
<td>49</td>
</tr>
<tr>
<td>Average</td>
<td>64</td>
<td>64</td>
<td>18</td>
<td>71</td>
<td>56</td>
<td>38</td>
</tr>
</tbody>
</table>

19.7.3 Determinants of Decentralization

Decentralization is a complex process, and we have provided but a snapshot of the enormous normative literature on how best to allocate different responsibilities between central and local governments and the possible efficiency gains of decentralization. However, the positive issue of why and when decentralization occurs deserves also some attention.

The positive literature on decentralization suggests certain empirical regularities concerning the forces that promote decentralization. Oates (1972) finds in a cross-sectional analysis that both country size and income per capita play a crucial role in explaining decentralization. The empirical evidence suggests that for different measures of decentralization, larger and richer countries are more decentralized. To better control for interregional geographic and cultural differences, Oates and Wallis (1988) use a panel analysis of 48 US states. They find that diversity as measured by urbanization increases decentralization.

In a very interesting study, Panizza (1999) estimates a theoretical model of decentralization that allows a test for the preference sorting effect. He uses a linear country model like the one presented previously in the optimal structure section. The level of public good provision is determined at the central level by majority voting (i.e., by the median voter’s preference). The central government provides a uniform public good level whose value decreases with the spatial distance between citizens and the central government. Local government provision is closer to the preferences of citizens and so more valuable. Since voters benefit more from local provision, increasing central provision reduces overall demand for the public good. Central government decides its share in provision of the public good, anticipating how that share influences overall demand for the public good. Using cross-sectional analysis and standard measures of decentralization, Panizza finds that decentralization increases with country size, income per capita, the level of democracy, and ethnolinguistic fractionalization.

An important limit of the existing empirical testing of decentralization is that it ignores a central force in the process of decentralization, namely the threat of separation. The possibility of secession has been a powerful force in limiting the ability of the central government to exploit peripheral minorities of voters for the sake of the majority of the population. The idea is that a unitary government is more willing to devolve more power and responsibility when the threat of secession is more credible. It has been a recurrent feature in Europe that the decision to decentralize is not necessarily guided by efficiency considerations but is in fact driven by distributional and political forces. When rich regions, which today transfer large amounts of income to poorer regions, demand more
decentralization, it is to limit their net contributions. They often do that because they no longer believe in the mutual insurance effect that such transfers might change directions in the near future. Also the size of the regional redistribution has become so visible that it creates an insurmountable political problem. The perception is that rich regions become better off by seceding, and to prevent such countries from breaking apart, concessions in the form of larger devolution of responsibilities and resources to regions have been taking place. Italy and Belgium are two good illustrations of the sort of decentralization forced by the pressing demand of the rich regions Lombardy and Flanders, respectively. It is clear that in those cases the efficiency argument that decentralization allows policy choices that better reflect local preference was not the key force. Richer regions demand more autonomy because the regional income inequality is such that mutual insurance becomes pure redistribution. Moreover the demand for more autonomy is exacerbated, rightly or wrongly, by the perceptions in the rich regions that the regional transfers are largely influenced by opportunistic behavior of the receiving regions (i.e., some form of moral hazard problem at the regional level).

19.8 Conclusions

There is a considerable controversy as to what public activities should be decentralized or centralized. There is also empirical evidence of increasing decentralization around the world. In this chapter we considered the costs and benefits of decentralization. We saw that one important advantage of a decentralized system is the tailoring of the provision of public goods and services to local preferences. The idea is that local government is closer to the people and so more responsive to their preferences than central governments. Another advantage of decentralization is to foster intergovernmental competition, making government more efficient and more accountable to the electorate.

There are also disadvantages of a decentralized system. Some, like fiscal competition, are covered in the next chapter. The main disadvantage of decentralization is probably its failure to exploit all the economies of scale in the provision of public goods. Another disadvantage is to limit the scope for interregional risk-sharing through the federal fiscal system.

Optimal federalism results from trade-offs among the various costs and benefits due to decentralization. It provides normative conclusions about the allocation of responsibilities between central and local levels. However, from a more positive perspective, political and distributional considerations can lead to different conclusions. The best illustration is that too much decentralization will result from a democratic choice.
Further Reading

The optimal government structure with overlapping jurisdictions is analyzed in:


The excessive decentralization with majority rule is in:


The empirical analysis of the determinants of decentralization is in:


The study of secession and breakup of nations is in:


Government accountability is in:


Risk-sharing is analyzed in:
Chapter 19: Fiscal Federalism


The relation of the federal fiscal system and insurance is in:


An analysis of hard and soft budgets can be found in:


Exercises

19.1 Two jurisdictions have preferences described by $U^A = - |\theta^A - G^A|^2$ and $U^B = - |\theta^B - G^B|^2$, where $G^j$ is the quantity of the local public good in jurisdiction $j$ and $\theta^j > 0$ is a parameter.

a. What is the optimal quantity of public good for the two jurisdictions?

b. If the public good is centrally provided so that $G^A = G^B$, find the quantity that maximizes $U^A + U^B$.

c. Calculate the loss from enforcing uniformity of provision.

19.2 The Tiebout hypothesis has been likened to consumers “voting with their feet.” In many voting situations the electors can gain by voting strategically. Why will a consumer never make a strategic choice of jurisdiction?

19.3 Is the allocation of the population between jurisdictions likely to be the efficient division in an economy where property rental is the norm or one where property ownership is the norm?

19.4 (Preference matching 1) Consider a society with many persons who can choose freely to live in either region 1 or region 2. It is more expensive to live in region 2: it costs $c_1$ to live in region 1 and $c_2 = c_1 + \Delta$ to live in region 2 (with $\Delta > 0$). Individuals differ in their incomes, denoted by $y$. Income takes on values between 0 and 1 and is uniformly distributed. Individuals care about the income of those living in their region. The mean income of a region $j = 1, 2$ is a function of the average value of $y$ in that region, denoted by $\bar{y}_j$. An individual with income $y$ choosing to live in region $j$ with mean income $\bar{y}_j$ derives utility net of cost of $U = [1 + y] [1 + \bar{y}_j] - c_j$. Hence richer individuals place greater value on living together with other rich residents.
a. Suppose that all individuals simultaneously make their location choices. Show that in any equilibrium (where no one wishes to move given the location choice of everyone else) both regions must be occupied if $\frac{1}{2} < \Delta < 1$; that is, the cost differential is neither too high nor too low. What would happen if $\Delta > 1$ or $\Delta < \frac{1}{2}$?

b. For $\frac{1}{2} < \Delta < 1$ show that in any equilibrium where the mean income differs across regions, every resident of region 1 must have a lower income than every resident of region 2.

c. Show that there exists a critical level of income $y^*$ such that all individuals with higher income choose to live in region 2 and all individuals with lower income choose to live in region 1. Provide an expression for this critical income level.

d. Show that in the equilibrium with income sorting, it is possible to make everyone better off by changing slightly the residential choices. *(Hint: Consider a small change in the critical income $y^*$.)*

19.5 *(Preference matching 2)* Consider two districts $A$ and $B$ with two types of residents, rich ($R$) and poor ($P$). Rich residents have an income of $Y_R = 2,000$ and poor residents have an income of $Y_P = 1,000$. Both districts provide a local public good for their residents. The rich residents value the local public good more than the poor residents. That is, the value of the local public good to each resident is $V_i = \frac{Y_iG}{10} - \frac{G}{2}$ for $i = R, P$, where $G$ is the level of local public good provision. The cost of the local public good per resident is $C = 5G$.

a. What are the marginal value and the marginal cost of the local public good for each type of resident?

b. What is the willingness to pay of the rich residents for the local public good? What is the willingness to pay of the poor residents?

c. In district $A$ there are 400 rich residents and 200 poor residents, whereas in district $B$ the numbers are reversed. What would be the public good provision in each district if it was decided by majority voting? What type of residents would not be happy with this voting outcome?

19.6 Consider the previous exercise and now suppose that both types of residents can migrate to the other district.

a. Which residents will move?

b. What will be the equilibrium distribution of residents?

c. Are there still residents unhappy with the amount of local public good?

d. Is the provision of public good efficient (according to the Samuelson rule)? Explain why or why not.

19.7 Consider the previous exercise with the situation before migration, and suppose that the government requires the rich residents to contribute $\frac{3}{4}$ of the cost of local public good provision where they live and the poor residents to contribute $\frac{1}{4}$ of the cost.

a. How much public good will be provided under majority voting in each district?

b. Is there any resident who wants to move (and, if so, who and where)? Why or why not?

19.8 Provide arguments for an against the decentralization of the following policy areas:

a. Law and order
b. Education  
c. Public transportation  
d. Welfare  
e. Traffic law  
f. Employment policy  
g. Regional development  

19.9 There are two regions. Both can undertake some investment in a local public good that improves the welfare of their residents. However, there are some spillovers in these investment decisions. If region 1 provides the public good, region 2 obtains some spillover benefits, and vice versa. More precisely, let $g_1$ and $g_2$ be the local public good levels in region 1 and 2 and their respective welfare levels be $W^i(g_i, g_j) = 2\left[a\sqrt{g_i} + b\sqrt{g_i g_j}\right] - c g_i$ for $i \neq j$ (with $i, j = 1, 2$), where $a > 0$ and $0 < b < c$.

a. Find the Nash equilibrium levels of $g_1$ and $g_2$ when public investment decisions are taken simultaneously. What is the equilibrium welfare level of each region?

b. Suppose that public investment decisions are centralized. What levels of $g_1$ and $g_2$ maximize the total welfare $W^1(g_1, g_2) + W^2(g_2, g_1)$? Are these levels higher or lower than in part a? Explain briefly. Is the welfare of each region higher than in part a? Why or why not?

c. Discuss a possible interregional transfer scheme based on public investment in the other region that could induce each region to choose noncooperatively in a Nash equilibrium the same public investment levels as in part b. Explain.

19.10 (Oates’s decentralization theorem). There are two regions $j = \{1, 2\}$ with the same population size. The population in each region is represented by a continuum of residents with a mass of unity. There are three goods: a private good $x$ and two regional public goods $g_1$ and $g_2$. One unit of either of the public goods requires $c$ units of private good to produce. In each region residents differ in their preference for public good. The resident with preference of type $\theta$ in region $j$ has utility function $u_{\theta}(x, g_j, g_{-j}) = x + \theta \left[(1 - \gamma) \log(g_j) + \gamma \log(g_{-j})\right]$, where $\gamma \in \left[0, \frac{1}{2}\right]$ measures the degree of inter-regional spillovers. With $\gamma = 0$, there are no spillovers, and residents only care about local public good provision. With $\gamma = \frac{1}{2}$, the residents care equally about public goods in both regions. In each region preference types $\theta$ are distributed according to the CDF $F_j(\theta)$ on the same interval $[0, \theta]$ with mean $\mu_j < \frac{\theta}{2}$. The mean is assumed to be equal to the median in both regions. Region 1 is assumed to display stronger average preference for public good, $\mu_1 > \mu_2$.

a. Calculate the (efficient) public good levels that maximize the aggregate surplus from public good provision.

b. In a decentralized regime the quantity of public good provision is chosen independently by each region so as to maximize the regional surplus from public goods and public spending is financed by a uniform tax on local residents. Thus, if region $j$ produces $g_j$, then each resident in region $j$ must pay a tax $t_j = c g_j$. What are the public good levels in a decentralized regime? Compare with efficient levels in exercise 19.11.
c. In a centralized regime, public good levels are chosen by a central government with the uniformity requirement that \( g_1 = g_2 \) and the equal sharing of the cost so that each resident in either region must pay \( t_j = \frac{c(g_1 + g_2)}{2} \). What is the (uniform) level of public good \( g \) maximizing aggregate public goods surplus? Compare with efficient levels in part a.

d. Show that if regional preferences are the same, \( \mu_1 = \mu_2 \), then centralization produces a higher level of surplus than decentralization if and only if there are spillovers \( \gamma > 0 \).

e. Show that if regional preferences are different, \( \mu_1 > \mu_2 \), then there exists a critical level of spillovers \( 0 < \gamma^* < \frac{1}{2} \) such that centralization produces higher level of surplus if and only if \( \gamma \geq \gamma^* \). Discuss the result and the tradeoff between the two regimes.

19.11 There are two regions \( j = 1, 2 \) and two goods, a private good \( x \) and a pure public good \( g \). There are \( n_j \) identical residents in region \( j \). Each resident is endowed with a fixed amount of private good that can be turned into a public good at a unit cost equal to one. The utility of each resident in region \( j \) is \( U_j = x_j + \theta_j \log(g) \).

a. Suppose that \( \theta_1 < \theta_2 \) and \( n_1 < n_2 \). Suppose also that each region chooses its public good provision independently so that the common level of public good is the sum of regional provisions \( g = \sum_j g_j \). How much public good will each region provide in the (Nash) equilibrium? (Hint: Beware of corner solutions.)

b. How does the equilibrium outcome change if we assume that \( \theta_1 < \theta_2 \) and \( n_1 > n_2 \)?

c. Now assume that public good provision is decided jointly for the two regions by a central authority (centralized setting). Majority voting implies that public good provision is then based on the preference of the median (larger) region; that is region 2 in part a and region 1 in part b. The advantage is to spread the cost more widely, and the disadvantage is to impose the preference of one region on the other. Find the optimal public good provision and compare with spending levels in part a and part b. Is it possible for spending to fall with centralization?

d. What is the welfare effect of centralization for each region when region 1 is larger? Does your conclusion change if region 2 is larger? Why or why not?

19.12 (Caplin and Nalebuff 1992) There are two regions \( j = 1, 2 \) competing for residents through their policy choice. The policy choice of each region \( j \) is represented by a point \( x_j \) on the interval \([0, 1]\). Prospective residents differ in their policy preferences. A resident of type \( \theta \) who lives in region \( j \) with policy \( x_j \) has utility \( U(x_j; \theta) = -[\theta - x_j]^2 \). Thus type \( \theta \)'s preferred policy is \( x = \theta \). Assume a triangular population distribution on the interval \([0, 1]\).

Given the policy choice of each region, each resident chooses the region with the policy closer to his most preferred policy. In equilibrium no region wishes to change its policy given the policy of the other region and no individual wishes to move given the policy choices.

a. If both regions are utilitarian (i.e., maximize the sum of their residents’ utilities), show that in equilibrium the two regions locate at \( \frac{1}{4} \) and \( \frac{3}{4} \) respectively with the population equally divided between them.

b. Show that if both regions are Rawlsian (i.e., they maximize the minimum utility of their residents), they will locate at \( \frac{1}{4} \) and \( \frac{3}{4} \) respectively, again with the population evenly distributed. Who are the worst-off residents? Compare with your answer in part a.

c. Now assume region 1 is Rawlsian and region 2 is utilitarian. Show that the equilibrium involves region 1 locating at 0.2 and region 2 locating at 0.6 with those to the left of 0.4 living...
in region 1 and the larger group to the right of 0.4 living in the utilitarian region 2. Who are the worst-off residents?

d. Comparing parts a and c, show that if one region switches to a more egalitarian objective (while the other is utilitarian), then society as a whole becomes less egalitarian. (Hint: Compare the welfare of the worst-off residents in parts a and c.)

e. Similarly, comparing parts b and c, show that if one region switches to a more egalitarian objective (when the other is Rawlsian), then society as a whole becomes more egalitarian.

19.13 Re-do the previous exercise with an inverted triangular distribution of population. Show that if one region switches to a more egalitarian objective, the welfare of the worst-off resident always improves no matter what the objective of the other region is. (Hint: The population distribution is bimodal.)

19.14 There are two time periods. In each period, politicians in office choose the level of public good. Between periods, there is an election. All politicians are benevolent (there is no rent diversion). They differ in their competence (i.e., ability to transform taxes into public output. There is uncertainty about the quality of politicians: politicians are “good” with probability $\pi$ and “bad” with probability $1 - \pi$. The good type is always low cost and can produce the public good at a unit cost $\theta_L$. The bad type is always high cost and can produce the public good at a unit cost $\theta_H$ (where $\theta_H > \theta_L$). All voters have an identical per-period utility function

$$W(G, T) = B(G) - \mu T,$$

where $B(G)$ is the increasing concave benefit from the public good and $\mu \geq 1$ is the marginal cost of public funds. The intensification of tax competition is represented by an increase in $\mu$ (i.e., a basic implication of the tax competition theory). There is no discounting between periods. There is a private benefit from holding office of $R > 0$.

a. Derive the equilibrium outcome and per-period voter welfare.

b. Show that more competition reduces the equilibrium level of public good provision and voter welfare.

19.15 Consider the above exercise and assume that the bad incumbent can “delay” the revelation of the high cost by borrowing $b$ (borrowing is only observable after the election). This creates an incentive for the bad type to appear competent by acting as if cost is low and creating a deficit during election years (the political budget cycle). In the first period, the incumbent observes the unit cost $\theta \in \{\theta_L, \theta_H\}$ and then chooses the level of provision $G$ and the amount of borrowing $b$, giving a total tax bill of $T = \theta G - b$. Voters observe the choices of $T$ and $G$ prior to the election. Voters make an inference about the incumbent’s type and re-elect the incumbent if he is at least as likely to be “good” as the challenger who is good with prior probability, $\pi$. In the second period, the politician then in office again chooses $G$ given $\theta$ and pays back $b$, making a tax bill of $T = \theta G + b$. There are no further elections.

a. Calculate the equilibrium outcome and show that there exists a critical value $\mu^{*}$ such that the bad incumbent will separate and lose the election if and only if $\mu > \mu^{*}$, and he will pool and be re-elected otherwise.

b. Show that increasing competition can raise voter welfare when politicians are likely to be bad.
Part VII: Multiple Jurisdictions

19.16 What principles should govern the allocation of power across different levels of government?

19.17 Is it possible that a federal social insurance scheme oversupplies risk-sharing? Explain briefly.

19.18 Discuss the following statement: “A community of a higher order should not interfere in the internal life of a community of a lower order, . . . but rather should support it in case of need and help coordinate its activity with the activities of the rest of society always with a view to the common good.” (John Paul II)

19.19 Describe the argument of yardstick competition to control policy makers. What are the main difficulties if yardstick competition is to be applied in practice to control performance of policy makers?

19.20 Discuss the relative costs and benefits of decentralization.
20 Fiscal Competition

20.1 Introduction

What is the role of competition among governments? If competition is the fundamental force for efficient economic performance in the private sector, why should it be different for the public sector? Why cannot the same disciplining effect of competition be applied to the public sector as well? In the private sector, competition will promote efficiency because firms that best satisfy consumers’ preferences will survive and prosper while others will lose customers and fail. Extending this argument to the public sector, competition among governments and jurisdictions should induce them to best serve the will of their residents. If they fail to do so, residents will vote with their feet and leave for other jurisdictions that offer a better deal.

The purpose of this chapter is to show that if the private competition analogy has some merit, it also needs to be seriously qualified. The chapter is organized as follows: First, the efficiency aspects of fiscal competition are presented. Second, the distributional aspects of residential mobility are evaluated. The key issue is how mobility limits the possibility of redistributing income. Third, the role of intergovernmental transfers is discussed both in terms of efficiency and redistribution. Fourth, some evidence on fiscal competition and intergovernmental interactions is given. Last, the main results from fiscal competition theory are summarized and evaluated in the concluding section.

20.2 Tax Competition

Tax competition refers to the interaction among governments due to interjurisdictional mobility of the tax base. It does not include the fiscal interaction among governments resulting from public good spillovers, where residents of one jurisdiction consume the public goods provided by neighboring jurisdictions. The cause of tax competition is that independent jurisdictions finance public expenditure by placing a tax on a mobile tax base. For example, a tax on capital will cause capital to seek a better return in alternative jurisdictions, and an income tax will cause workers to move to another jurisdiction if they are mobile. The loss of tax base by one jurisdiction represents a gain to the others, so mobility causes an externality among jurisdictions.

In the model we consider tax competition that arises because jurisdictions finance the provision of a public good with a tax on locally employed capital. Capital moves across
jurisdictions in response to tax differentials while residents are typically immobile (or at least less mobile). In the competitive version of tax competition, jurisdictions are “too small” (relative to the economy) to affect the net return to capital that is determined worldwide. As a result each jurisdiction sets its tax on locally employed capital, taking as given the net-of-tax price of capital. Tax rates in other jurisdictions do not matter, and there is no strategic interaction among jurisdictions when setting their taxes. We say that jurisdictions behave competitively. When jurisdictions are “large” relative to the economy, each jurisdiction can affect the net return to capital by varying its own tax rate. In this case the tax rate chosen in one jurisdiction varies with the taxes in other jurisdictions. Jurisdictions behave strategically: they set their tax in response to the tax rates in other jurisdictions.

Both the competitive and strategic versions of the tax competition model produce the same important conclusion, namely that public goods are underprovided relative to the efficient Samuelson rule level. The reason is that each jurisdiction perceives the mobility of capital and keeps its tax low to preserve its tax base. To understand the inefficiency arising from intergovernmental competition, it is useful to consider a simple model in which we assume, in turn, that jurisdictions behave competitively and then strategically.

20.2.1 Competitive Behavior

The assumption of competition means that the mobile factor of production is available to the “small” jurisdiction at a fixed price. Suppose that capital is the mobile factor of production and that the jurisdiction seeks to impose a tax on capital and to use the revenue to provide public goods and services to its residents, or to directly transfer cash to them. If capital were perfectly immobile, a local source-based tax on capital would reduce the net rate of return to capital by the exact amount of the tax. This capital tax would make the residents better off at the expense of capital owners (who are not necessarily residents).

In contrast, when capital is costlessly mobile, local capital taxation cannot affect the net return to capital. The reason is that the imposition of the local tax drives capital out of the jurisdiction until the increase in the gross rate of return is sufficient to compensate capital owners for local taxes. However, the outflow of capital from the jurisdiction reduces the remuneration of labor. The resulting loss of income to the residents will exceed the value of the tax revenue collected from capital taxation. Except for the case where public expenditures have greater value to local residents than the tax revenue used to finance them, the net effect of capital taxation is to harm immobile residents.
Therefore, with competition and costless mobility, the taxation of capital is impossible, whereas it may be desirable without mobility. Capital taxes that help immobile residents when capital is immobile harm them when this factor is perfectly mobile.

Assume that the local production process uses mobile capital and immobile labor. The production function is \( F(K_i, L) \), where \( K_i \) is the aggregate capital and \( L \) is aggregate labor employed in jurisdiction \( i \). Each worker is endowed with one unit of labor.

Under constant returns to scale, \( F(K_i, L) = LF\left( \frac{K_i}{L}, 1 \right) = Lf(k_i) \), where \( k_i \) is the capital–labor ratio. The production function \( f(k_i) \) gives the per capita output, which is increasing and concave \( (f''(k_i) < 0 < f'(k_i)) \). The concavity of the production function reflects diminishing returns to capital as it is combined with the immobile stock of labor.

Let \( \rho \) denote the net return to capital outside the jurisdiction, and let \( t_i \) denote the per unit tax on the capital employed in the jurisdiction \( i \). With costless mobility of capital, the local supply of capital equates its net return in jurisdiction \( i \) with its net return elsewhere:

\[
f'(k_i) - t_i = \rho. \tag{20.1}
\]

With exogenously fixed \( \rho \), the fact that \( f'(k_i) \) is decreasing in \( k_i \) implies that a higher tax drives capital away, so \( \frac{dk_i}{dt_i} < 0 \). Assuming that the net revenue collected from local capital taxation accrues to workers in the form of cash transfers or public goods of the same value, the net income of workers will be

\[
y_i = f(k_i) - f'(k_i)k_i + t_i k_i = f(k_i) - \rho k_i, \tag{20.2}
\]

where the second equality follows from the arbitrage condition \( f'(k_i) = \rho + t_i \). Because taxation reduces the amount of capital in the jurisdiction, it is then easily seen that the welfare of the workers, as measured by their net income \( y_i \), is maximized by setting \( t_i = 0 \).

### 20.2.2 Strategic Behavior

It is now assumed that jurisdictions behave strategically. The strategic interaction makes the equilibrium analysis more delicate, and it is useful to describe the equilibrium outcome rigorously. Consequently this section will use more calculus than usual and can be skipped with little loss of continuity by those who wish to.

Consider two countries \((i = 1, 2)\) that levy a tax on the return to capital. Capital is mobile and is used together with some fixed amount of labor to produce output. The
production function is \( F(K_i, L) \), where \( K_i \) is the aggregate capital and \( L \) is aggregate labor employed in country \( i \). The quantity of labor available and the production technology are the same for the two countries, and each worker is endowed with one unit of labor. Under constant returns to scale, \( F(K_i, L) = LF\left(\frac{K_i}{L}, 1\right) = Lf(k_i) \), where \( k_i \) is the capital–labor ratio. The production function \( f(k_i) \) gives the per capita output, which is increasing and concave \( (f''(k_i) < 0 < f'(k_i)) \). There is a fixed stock of capital \( \bar{k} \) that allocates itself between the two countries, so \( k_1 + k_2 = \bar{k} \). Each country levies a per unit tax \( t_i \) on the capital that is employed within its boundaries. The revenue raised is used to supply a level of public services of \( G_i = t_i k_i \). Due to capital mobility the tax choice of one jurisdiction affects the size of the tax base available to the other country.

Given the pair of tax rates, costless mobility implies the equality of the after-tax return to capital across countries:

\[
f'(k_1) - t_1 = f'(k_2) - t_2 = f'(%20.3)k - k_1) - t_2.
\]

This arbitrage condition produces an allocation of capital across countries that depends on the tax rates, as illustrated in figure 20.1.

The partition of the capital stock between the two countries is represented on the horizontal axis with the capital levels measured from the two corners (from left to
right for country 1). The corresponding marginal product of capital in each country is measured on the vertical axis (the left axis for country 1). Note that if the tax rates do differ (e.g., $t_1 > t_2$), then capital is inefficiently allocated because the marginal product of capital differs across countries ($f'(k_1) > f'(k_2)$). It can also be seen that an increase in the tax rate in country 1 reduces the net return to capital in that country and causes some capital to move away to country 2. The converse holds when the tax in country 2 is increased.

These observations can be demonstrated formally by taking the total differential of the arbitrage condition (20.3) with respect to $t_1$ and $k_1$ to give

$$f''(k_1)dk_1 - dt_1 = -f''(k_2)dk_1.$$  

(20.4)

Then the variation in $k_1$ in response to the tax change $dt_1$ is

$$\frac{dk_1}{dt_1} = \frac{1}{f''(k_1) + f''(k_2)} < 0.$$  

(20.5)

The sign of this expression follows from the assumption of a decreasing marginal product of capital, $f'' < 0$. Note that this assumption implies some regulating forces in the allocation of capital, since, when capital moves from 1 to 2, its marginal product decreases in country 2 at rate $f''(k_2)$ and rises in country 1 at rate $-f''(k_1)$. When setting its tax rate, country 1 will take into account how capital responds. That is, it will incorporate the movement of capital described above into its decision problem.

Assuming that the net revenue collected from local capital taxation accrues to workers in the form of cash transfers or public goods of the same value, we have the net income of workers (or residents) in country 1 as

$$y_1 = f(k_1) - f'(k_1)k_1 + t_1k_1.$$  

(20.6)

Each country maximizes the net income of its residents while taking into account capital flows resulting from tax changes. Because the amount of capital employed in each country also depends on the other country’s tax rate, there is strategic fiscal interaction among countries: neither can set its own tax rate without taking into account what the other is doing.

The optimal choice of each country is found by applying the usual Nash assumption: each takes the tax rate of the other as given when maximizing. By this reasoning, the best response of country 1 to the other country’s tax $t_2$ is described by the following first-order condition:
Part VII: Multiple Jurisdictions

\[
\frac{dy_1(t_1, t_2)}{dt_1} = -k_1 f_1' \frac{dk_1}{dt_1} + k_1 + t_1 \frac{dk_1}{dt_1} = [k_1 f_2'' + t_1] \frac{dk_1}{dt_1} = 0, \quad (20.7)
\]

where the second equality is obtained by using (20.5) to substitute for \( k_1 = k_1 \left[ f_1'' + f_2'' \right] \frac{dk_1}{dt_1} \) and the subscript on \( f \) denotes the country. Therefore the best-response function for country 1 can be written as

\[
t_1 = -k_1 (t_1, t_2) f_2'' \left( \bar{k} - k_1 (t_1, t_2) \right), \quad (20.8)
\]

which can be solved for \( t_1 \) to give

\[
t_1 = r_1(t_2). \quad (20.9)
\]

Similarly, for country 2,

\[
t_2 = -k_2 (t_1, t_2) f_1'' \left( k_1 (t_1, t_2) \right), \quad (20.10)
\]

which implies

\[
t_2 = r_2(t_1). \quad (20.11)
\]

A Nash equilibrium is a pair \( (t_1^*, t_2^*) \) such that the tax choice of each country is a best response to the other country’s tax choice, \( t_1^* = r_1(t_2^*) \) and \( t_2^* = r_2(t_1^*) \). The symmetry of the model implies that both countries choose the same taxes in equilibrium, so \( t_1^* = t_2^* \), and consequently capital is evenly distributed between jurisdictions with \( k_1 = k_2 = \frac{\bar{k}}{2} \). The Nash equilibrium in taxes is thus

\[
t_1^* = t_2^* = -\frac{\bar{k}}{2} f'' \left( \frac{\bar{k}}{2} \right). \quad (20.12)
\]

We can also find the slope of the best-response function \( r_1(t_2) \) to evaluate the nature of strategic interdependency between the two countries. The first-order condition \( \psi_1(t_1, t_2) = k_1 f_2'' + t_1 = 0 \) implicitly defines \( t_1 \) as a function of \( t_2 \), and we need to know how the optimum choice of \( t_1 \) will respond to changes in \( t_2 \). Differentiating the first-order condition totally, we have

\[
\frac{\partial \psi_1}{\partial t_1} dt_1 + \frac{\partial \psi_1}{\partial t_2} dt_2 = 0. \quad (20.13)
\]
Recall that with a fixed stock of capital, the loss of tax base by one jurisdiction represents a gain to the others, so \( \frac{dk_1}{dt_1} = -\frac{dk_2}{dt_1} = -\frac{dk_1}{dt_2} \). This gives the slope of the best response as

\[
\frac{dr_1}{dt_2} = \frac{\left[ f''_2 - k_1 f'''_2 \right] dk_1/dt_1}{1 + \left[ f''_2 - k_1 f'''_2 \right] dk_1/dt_1},
\]

(20.14)

where \( \frac{dk_1}{dt_1} < 0 \), and \( f'' < 0 \). It follows that for \( f''' \geq 0 \) the slope of the best-response function satisfies \( 0 < \frac{dr_1}{dt_2} < 1 \). It can be seen in figure 20.2 that the tax rates are strategic complements: a higher tax in country 2 drives capital to country 1, which in response raises its own tax rate.

It is now easily seen that such a Nash equilibrium with \( t_1^* = t_2^* = t^* \) involves inefficiently low taxes and that jointly increasing taxes to \( t > t^* \) is beneficial to both countries. First, observe that from the perspective of the two countries together, the stock of capital is fixed at \( k \). Hence it is simply a fixed factor. Provided that both countries levy the same tax rate \( t = t_1 = t_2 \), half of the capital, \( \frac{k}{2} \), will be located in each country regardless of the level of the taxes. The welfare of the workers in each country, as measured by their net income \( y = f(k) - f'(k)k + tk \), is then improved, since \( t \frac{k}{2} > t^* \frac{k}{2} \). In fact, with the cooperative tax setting, the countries can maximize the net income of their residents by fully taxing the mobile factor, whereas the noncooperative equilibrium leads to a lower tax on capital. Welfare is higher with cooperation. This loss in potential welfare is the efficiency cost of fiscal competition.
Although the model just considered leads to the extreme conclusion that the tax is pushed to its maximum with cooperation, this was not the major point of the analysis. What the model does illustrate are the forces that are at work when an attempt is made to tax a mobile factor of production. The movement of the factor generates an externality among countries that is not internalized when their governments conduct individual optimization of taxes. This externality is due to the fact that a higher tax in one country pushes some of the factor to the other country. This has the beneficial effect of increasing the other country’s tax base and its tax revenue at any given tax rate. Cooperation among countries in the choice of taxes internalizes this externality and allows them to choose a mutually preferable set of tax rates.

Consequently competition for mobile factors of production results in tax rates that are lower than is optimal for the countries involved. Implicitly each country can be understood to be trying to undercut the other to attract the mobile factor of production. This undercutting puts downward pressure on tax rates to the detriment of all countries. The policy principle that emerges from this scenario is that international cooperation on the setting of tax rates is beneficial.

As already noted, although the argument has been phrased in terms of countries, the same results would apply within a federal structure in which the separate jurisdictions at any level set their own tax rates. It is possibly more relevant in such a context because the factors of production may be more mobile than they are between countries. Furthermore, for tax competition to arise, the tax base need not be a factor of production but simply needs to be mobile between jurisdictions. For example, the argument applies equally well to the taxation of commodities provided that purchases can be made mobile through cross-border shopping. The resulting equilibrium with cross-border shopping will have inefficiently low commodity taxes. The taxes will be lower, the higher the perceived elasticity of cross-border shopping. This is so as long as goods are taxed according to the origin principle (i.e., the goods are taxed where they are produced). If the destination principle is applied, the goods are taxed in the country where they are consumed and the incentive for cross-border shopping would disappear. However, the destination principle of taxation is costly to operate because it requires that all taxes levied in the country of production be rebated when the good is exported. This is only possible if border controls are maintained. The same commodity taxes must be levied on all imports into the country of final consumption, with tax revenue also accruing to this country.

As such, the tax competition argument provides some important reasons for being cautious about the benefits of fiscal federalism that were described in the previous chapter. Giving jurisdictions too much freedom in tax setting may lead to mutually damaging reductions in taxes—the so-called race to the bottom.
20.2.3 Size Matters

Differences in country size, production technologies, factor endowments, or residents’ preferences can be expected to cause the countries to choose different tax rates. An interesting aspect of the ensuing asymmetric tax competition is the so-called benefit of smallness. The idea is that although fiscal competition is inefficient, it can actually benefit small countries. Of course, such a gain comes at the expense of larger countries.

This can be seen in our simple two-country model by assuming that they differ in their number of residents only. Suppose that country 1 is “large” with a share \( s > \frac{1}{2} \) of the total population and country 2 is “small” with population share \( 1 - s < \frac{1}{2} \). The capital market-clearing condition is then

\[
s k_1(t_1, t_2) + [1 - s] k_2(t_1, t_2) = \tilde{k},
\]

where \( \tilde{k} \) is the (worldwide) average capital–labor ratio. The arbitrage condition implies equality of the after-tax return on capital across countries:

\[
f'(k_1) - t_1 = f'(k_2) - t_2
\]

\[
= f'\left( \frac{\tilde{k}}{1 - s} - \frac{s k_1}{1 - s} \right) - t_2.
\]

Differentiating the arbitrage condition gives the capital outflow in response to a domestic tax increase,

\[
\frac{d k_1}{d t_1} = \frac{1 - s}{[1 - s] f''(k_1) + sf''(k_2)} < 0,
\]

and by analogy, for the small country,

\[
\frac{d k_2}{d t_2} = \frac{s}{[1 - s] f''(k_1) + sf''(k_2)} < 0.
\]

From (20.17) and (20.18) it follows that both countries face a capital outflow after an increase in their own tax rate, but this outflow is less severe in the large country. Indeed, when \( t_1 = t_2 \), we have \( k_1 = k_2, f''(k_1) = f''(k_2) \), and thus \( \frac{d k_2}{d t_2} < \frac{d k_1}{d t_1} < 0 \) for \( s > 1 - s \). The larger country faces a less elastic tax base and thereby chooses a higher tax rate than the smaller country, so in equilibrium \( t_1 > t_2 \). Because the small country charges a lower tax on capital, it will employ more capital per unit of labor, \( (k_2 > k_1) \), increasing per capita income and making its residents better off than the residents of the large country. It is even possible that for a sufficiently large difference in size, the small country will be better off than it would be without tax competition.
This benefit of smallness is illustrated in figure 20.3, where for country $i = 1, 2$ per capita income is denoted $c_i = f(k_i) - f'(k_i)k_i$ and tax revenue is denoted $g_i = t_i k_i$. The net return to capital, denoted by $R$, is the same for both countries by arbitrage and is adjusted to tax choices in order to clear the market. It is then readily seen that the residents of the small country are better off taxing less, since $c_2 + g_2 > c_1 + G_1$.

We can obviously extend this reasoning to show that if the number of countries competing for capital increases, each country having a lower population share will perceive a greater elasticity of its tax base and choose lower taxes: the larger the number of countries, the more intense the competition and the lower the equilibrium taxes. Again, commodity tax competition displays a close similarity to capital tax competition when countries differ in population size, with the smaller country setting the lower tax rate as it perceives a higher elasticity of its domestic tax base.

### 20.2.4 Public Input Provision

In the standard model of tax competition there is no specific need for revenue. The capital tax is used solely to extract the rent from capital and the revenue raised in each jurisdiction is simply added to the net incomes of the workers. The addition to income can be interpreted as public good provision, but only if a constant marginal utility of public good is assumed. Different conclusions about tax competition can emerge if the use of tax revenue is considered in more detail. If it is used to finance a public input...
that is used with private capital in production, then there is a motive for levying the tax that is missing in the standard model. This can have an effect on the nature of the tax externality between jurisdictions.

Assume that there are two jurisdictions. Each jurisdiction, \( i \), provides a public input, \( g_i \), which combines with private capital, \( k_i \), in the production function

\[
y_i = f (k_i, g_i) .
\]  
(20.19)

The public input is financed by the tax on capital, so the quantity in \( i \) is determined by

\[
g_i = t_i k_i , \quad i = 1, 2 .
\]

If an increase in the tax rate, \( t_i \), raises additional revenue, it will increase \( g_i \). If the increase in public input raises the marginal productivity of capital, this will offset the effect of the tax increase on the net return to capital. When this effect is strong enough, it is possible for an increase in the tax rate to attract mobile capital. This can give an incentive to set taxes above the efficient level.

This reasoning is made clear when the externality underlying the tax competition is explored. The standard model without a public input has a positive tax externality

\[
\frac{\partial y_2}{\partial t_1} = \left[ -f k_2 k_2 + t_2 \right] \frac{\partial k_2}{\partial t_1} > 0 ,
\]  
(20.20)

so that an increase in the tax rate in jurisdiction 1 increases income in jurisdiction 2. This is because the tax increase reduces the net return to capital in jurisdiction 1 and mobile capital flows out. Now consider the situation where the public input is included in the model. The equilibrium level of the public input in jurisdiction 2 must depend on the tax rates in the two countries

\[
g_2 = g_2 (t_1, t_2) .
\]  
(20.21)

This dependence adds an additional component to the externality effect of taxation. The tax externality then becomes

\[
\frac{\partial y_2}{\partial t_1} = \left[ -f k_2 k_2 + t_2 \right] \frac{\partial k_2}{\partial t_1} + \left[ f g_2 - f k_2 k_2 \right] \frac{\partial g_2}{\partial t_1} .
\]  
(20.22)

The budget constraint implies that

\[
\frac{\partial g_2}{\partial t_1} = t_2 \frac{\partial k_2}{\partial t_1} ,
\]  
(20.23)

which gives

\[
\frac{\partial y_2}{\partial t_1} = \left[ 1 + f g_2 \right] t_2 - \left[ f k_2 + f k_2 g_2 t_2 \right] k_2 \frac{\partial k_2}{\partial t_1} .
\]  
(20.24)
It can be observed from (20.24) that it is possible that a sufficiently strong complementarity in production \((f_{k_2 g_2} >> 0)\) can create a negative externality. This occurs when

\[
f_{k_2 g_2} > \frac{1 + f_{g_2}}{k_2} - \frac{f_{k_2 k_2}}{t_2}.
\]  

(20.25)

If (20.25) is satisfied, the equilibrium tax rates will be higher than the efficient level at the Nash equilibrium.

20.2.5 Tax Overlap

A common feature of fiscal federalism is that higher and lower levels of government share the same tax base. This tax base overlap gives rise to vertical fiscal externalities. With tax competition among jurisdictions, the horizontal fiscal externality on other regions is positive—an increase in tax rate by one region raises the tax base of others. In contrast, if different levels of government share the same tax base, then the tax levied by one government will reduce the tax base available to other levels of governments. This introduces a negative vertical externality. Not surprisingly, such vertical externalities lead to overtaxation in equilibrium because each level of government neglects the negative effect of its taxation on the other levels of government.

The joint taxation of cigarettes by Canadian federal and provincial governments is a good example. Figure 20.4 illustrates this tax overlap problem. The supply curve, \(S\), is assumed to be perfectly elastic, and the demand curve, \(D\), is downward-sloping. Suppose that the initial federal excise tax rate is \(T_0\) and the provincial tax rate is \(t_0\). The corresponding price is \(p_0\) and the quantity of cigarettes consumed is \(q_0\). Tax revenue is \(T_0 q_0\) for the federal government, and \(t_0 q_0\) for the provincial government. If the provincial government raises its tax rate to \(t_1 = t_0 + \Delta\), the consumer price increases by the amount of the tax increase \(p_1 = p_0 + \Delta\), and the quantity consumed decreases to \(q_1\). The tax revenue of the provincial government increases by \([t_1 - t_0] q_1 - t_0 [q_0 - q_1]\), but due to the reduction in the consumption of cigarettes, revenue for the federal government decreases by \(T_0 [q_0 - q_1]\), as represented by the shaded area in figure 20.4. A similar vertical externality (but in the opposite direction) would arise if the federal government were to raise its tax rate. If both levels of government neglect the revenue losses incurred by the other government when making their tax choices, then both governments are underestimating the cost of raising tax revenue from the common tax base and will tend to choose tax rates that are inefficiently high.
When vertical and horizontal externalities are combined, the noncooperative equilibrium outcome is ambiguous: it involves excessively low taxes if the horizontal externalities dominate the vertical externalities. Canada again provides an important example because most provincial governments levy their personal income tax as a fraction of the federal income tax. On top of this, each province levies a surtax on high-income residents. The bias in the perceived marginal cost of taxation caused by tax base overlap may explain why Canadian provinces have introduced high-income surtaxes when tax competition for mobile high-income taxpayers predicts the reverse.

### 20.2.6 Tax Exporting

In any country some of the commodities that are sold within its borders will be purchased by nonresidents (especially cross-border shoppers). This will be particularly true if the country is important in the context of international tourism. It will also be encouraged under fiscal federalism with a single market covering all jurisdictions. Similarly some of the productive activity carried out in a country will be undertaken by firms that repatriate their profits to another country. Whenever there is such economic activity by nonresidents, the possibility for tax exporting arises.

Tax exporting is the levying of taxes that discriminate against nonresidents. A simple example would be the imposition of a higher level of VAT on restaurants located in centers of tourism. The motive for such tax exporting is to shift some of the burden of
revenue collection onto nonresidents and to lower it on residents. All else held constant, this is clearly of benefit to residents. However, all else is not constant in practice, and the same argument will apply to all countries. As for tax overlap, tax exporting provides an argument for why tax rates may be set too high when countries compete.

Another form of tax exporting is the taxation of capital employed in the country but owned by nonresidents. The simplest version of this form of tax exporting can be described with the previous model of tax competition by assuming that country 1’s residents have a capital endowment $k_1 > 0$ that differs from that of country 2, $k_1 \neq k_2$. Capital owners in each country are free to invest their capital in their home country or abroad. The level of social welfare in country 1 is measured by the net income of its workers, $y_1 = f(k_1) - f'(k_1)k_1 + t_1k_1$, where $k_1$ is the amount of capital employed in country 1, plus the net income of its capital owners, $\rho k_1$. The capital market-clearing condition requires that

$$k_1 - k_1 = -[k_2 - k_2].$$  \hspace{1cm} (20.26)

That is, if $k_1 < k_1$, country 1 is employing less capital than its endowment, and its net export of capital has to be equal to the net import of capital from country 2. For country 1 the problem is to set its tax $t_1$ on capital given the tax of country 2 so as to maximize $f(k_1) - f'(k_1)k_1 + t_1k_1 + \rho k_1$, where $k_1 = k_1(t_1, t_2)$ is the amount of capital employed in country 1 given tax rates $t_1, t_2$ and $\rho$ is the net return to capital. Using $\rho = f'(k_1) - t_1$, the objective function can be written as follows:

$$W_1 = f(k_1) + \rho [k_1 - k_1];$$  \hspace{1cm} (20.28)

that is, country welfare is equal to total production $f(k_1)$ plus the net return to capital exports $\rho [k_1 - k_1]$. When deriving the first-order condition, we must differentiate $W_1$ with respect to $t_1$, taking into account the change in capital supply $\frac{dk_1}{dt_1}$ and the change in the net return to capital $\frac{d\rho}{dt_1}$. This gives

$$\frac{dW}{dt_1} = [f'(k_1) - \rho]\frac{dk_1}{dt_1} + [k_1 - k_1]\frac{d\rho}{dt_1}$$

$$= t_1\frac{dk_1}{dt_1} + [k_1 - k_1]\left[f''(k_1)\frac{dk_1}{dt_1} - 1\right] = 0.$$  \hspace{1cm} (20.29)

To solve this first-order condition, we can use (20.5) to get $f''(k_1)\frac{dk_1}{dt_1} - 1 = -f''(k_1)\frac{dk_1}{dt_1}$, which gives
Chapter 20: Fiscal Competition

\[ t_1 = f''_1 [\bar{k}_1 - k_1], \quad (20.30) \]

and by analogy for country 2 using (20.26),

\[ t_2 = -f''_1 [\bar{k}_1 - k_1], \quad (20.31) \]

with \( f''_i < 0 \). Therefore in any equilibrium, if country 1 has the larger endowment of capital so that \( \bar{k}_1 > \bar{k}_2 \), it will export capital \( (k_1 < \bar{k}_1) \) and prefer to subsidize capital, \( t_1 < 0 \). The reason is the terms-of-trade effect. By subsidizing capital, the country with large endowment of capital can raise the net return to capital. Because the other country will import capital from country 1, it will tax capital, \( t_2 > 0 \), as a means of taxing nonresidents. This is the tax-exporting effect. Next, note that the initial asymmetry in capital endowments leads countries to set different tax rates. This nonuniform tax equilibrium has important implications in terms of productive efficiency. Indeed the efficient allocation of capital requires its marginal product to be equalized across countries. But because country 1 subsidizes capital and country 2 taxes it, the marginal product of capital is higher in country 2 than in country 1, \( f'_2 > f'_1 \). Therefore, in equilibrium, country 1 attracts too much capital and country 2 too little, relative to what efficiency recommends.

20.2.7 Efficient Tax Competition

Tax competition has been seen as producing wasteful competition. There are circumstances, however, where tax competition may be welfare enhancing. We consider two examples.

The first example is the case where countries seek to give a competitive advantage to their own firms by offering wasteful subsidies. In equilibrium all countries will do this, so the effect of each country’s subsidy cancels out with that of the others. Since the subsidies cancel, no country gains an advantage, and all countries will be better off giving no subsidy. This is the prisoners’ dilemma once again. Tax competition may help overcome this inefficient outcome by allowing firms to locate wherever they choose and preventing governments from discriminating between domestic and foreign firms operating within a country. The mobility of the firms will force governments to recognize that their subsidy will not only give a competitive advantage to their domestic firms but that it will also attract firms from other countries. Because the government cannot discriminate between the domestic and foreign firms operating within its borders, it will have to pay the subsidy to both types of firms, thereby eliminating the
competitive advantage. Therefore mobility eliminates the potential gains from the subsidy and raises its cost by extending its payment to foreign firms.

Tax competition can therefore improve welfare by reducing the incentive for countries to resort to wasteful subsidies to protect their own industries. Notice that the nondiscrimination requirement plays a crucial role in making tax competition welfare improving. If discrimination were possible, then governments could continue to give wasteful subsidies to their domestic firms.

The second example is the use of tax competition as a commitment device. In the tax competition model, governments independently announce tax rates and then the owners of capital choose where to invest. A commitment problem arises here because the governments are able to revise their tax rates after investment decisions are made. If there were a single government and investment decision were irreversible, then this government would have an incentive to tax away all profits. The capital owner would anticipate this incentive when making its initial investment decision and choose not to invest capital in such a country.

Tax competition may help solve this commitment problem. The reason is that intergovernmental competition for capital would deter each government from taxing away profits within its borders because it would induce reallocation of capital between countries in response to differences in tax rates. Tax competition is a useful commitment device as it induces governments to forgo their incentive to tax investment in an effort to attract further investment or to maintain the existing investment level.

The original insight that tax competition leads to inefficiently low taxes and public good provision was obtained in models with benevolent decision makers. An alternative approach is to consider public officials that seek in their decision-making to maximize their own welfare and not necessarily that of their constituencies. From this perspective, tax competition may help discipline nonbenevolent governments. For instance, if we view governments as “leviathans” mainly concerned with maximizing the size of the public sector, then tax competition may improve welfare by limiting taxation possibilities and thereby cutting down the size of government that would be otherwise excessive. This argument suggests that the public sector should be smaller, the greater the extent to which taxes and expenditures are decentralized. The evidence on this is, however, mixed. In fact there is not much evidence on the relationship between fiscal decentralization and the overall size of the public sector.

An analogous argument applies to governments with some degree of benevolence, possibly due to electoral concerns. When political agency problems are introduced, this inefficiency of competition among governments is no longer so clear. Intergovernmental competition makes the costs of public programs more visible, as well as
their benefits, in ways that make public officials accountable for their decisions. Stated briefly, competition may induce government officials to reduce waste and thus reduce the effective price of public goods.

20.3 Income Distribution

When the powers for tax setting are devolved to individual jurisdictions, the Tiebout hypothesis asserts that the outcome will be efficient. The basis for this argument is that there are enough jurisdictions for individuals to sort themselves into optimal locations. For practical purposes it is not possible to appeal to this large-numbers assumption, and questions need to be asked about the outcome that will emerge when only a small, predetermined number of jurisdictions exist. The Tiebout hypothesis is also silent about how the policy of a jurisdiction emerges. It is possible that equilibrium results in all the residents of any jurisdiction being identical, so that there is no need to resolve different points of view. More generally, though, it is necessary to explore the consequences of political decision-making, expressed through elections, on the choice of policy.

An important set of issues revolves around income distribution and the role that this has in determining the composition of the population in jurisdictions. For instance, will it always be the case that the rich wish to detach themselves from the poor so that they can avoid being subject to redistributive taxation? Also, if they have the option, would the poor wish to live with the rich? These questions are now explored under perfect and imperfect mobility.

20.3.1 Perfect Mobility

The difficulty that mobility poses for redistribution is seen most strongly in the following example: Consider individuals who differ only in income level \( y \) and who can choose to reside in one of the two available jurisdictions. The jurisdictions independently set a constant tax rate \( t \) between 0 and 1 and pay a lump-sum transfer \( g \) subject to a budget balance constraint. Individuals care only about their income after taxes and transfers, so their preferences are given by

\[
u(t, g; y) = g + (1 - t) y. \tag{20.32}\]

The tax-transfer pair \((t, g)\) in each jurisdiction is chosen by some unspecified collective decision rule (e.g., majority voting). We are looking for an equilibrium in which the
two jurisdictions differ and offer different tax and transfer schemes, thereby inducing the sorting of types across jurisdictions (as the Tiebout hypothesis would predict).

With no loss of generality, suppose that it is jurisdiction 1 that sets the higher tax rate. Then, to attract any individuals, it must also provide a higher level of transfer, that is, \((t_1, g_1) > (t_2, g_2)\). Individuals with different income levels differ in their preferences for redistribution. From (20.32), if a type \(y\) prefers the high-tax jurisdiction 1, then all those with lower income levels will also prefer this jurisdiction. And, if a type \(y\) prefers the low-tax jurisdiction 2, then all those with higher income levels will also prefer this jurisdiction. Therefore, if both jurisdictions are occupied in equilibrium, there must exist a separating type \(y^*\) who is just indifferent between the two tax schemes and all those who are poorer, with \(y \leq y^*\), join jurisdiction 1 and all those who are richer, with \(y > y^*\), join jurisdiction 2. That is, the jurisdiction undertaking more redistribution attracts the poorest individuals.

However, this cannot be an equilibrium because the richest individual in the poor jurisdiction loses out from intrajurisdictional redistribution and will prefer to move to the rich jurisdiction and become a net beneficiary of redistribution as its poorest resident. Therefore there cannot be an equilibrium with different tax-transfer schemes.

There remains the possibility of a symmetric equilibrium with individuals evenly divided between the two jurisdictions. In other words, perfect mobility leads to harmonization of tax-transfer schemes, even though agents differ in their preferences. Mobility does not lead to the sorting of types across jurisdictions as Tiebout predicts. The possibility for the rich to detach themselves from the poor to escape redistributive taxation induces jurisdictions either to abandon any taxation or to choose the same tax rate.

### 20.3.2 Imperfect Mobility

Suppose now that consumers have one of two income levels. Those with the higher income level are termed the “rich,” and those with the lower income are the “poor.” The two groups are imperfectly mobile but to different degrees. The rich (group 1) have income 1 and poor (group 0) have income 0. For simplicity, there is an equal number of poor and rich in the total population. The focus is placed on one of the jurisdictions (e.g., region 1) and the proportions from each group residing there are denoted \(x_1\) and \(x_0\), where the subscript denotes income group. The remainder, \([1 - x_1]\) and \([1 - x_0]\), are located in the other jurisdiction. Redistribution implies that the rich are subject to taxation and the poor are recipients of a transfer. Accordingly, each jurisdiction levies a head tax \(t\) on its rich residents to pay a transfer \(b\) to each of its poor residents. The feasible choices are restricted by the budget constraint \(tx_1 = bx_0\).
In addition to income differences each individual is characterized by a preference for location \( x \), with \( 0 \leq x \leq 1 \), where a low \( x \) implies a preference for region 1 and a high \( x \) implies a preference for region 2. It is assumed that \( x \) is uniformly distributed within income groups. Individuals care only about their net income and their location. Given the pair of transfers \((b, b^*)\) in the two regions, the payoff of a poor individual with preference \( x \) is
\[
\begin{align*}
&b - d_0 x \quad \text{in region 1,} \\
&b^* - d_0 [1 - x] \quad \text{in region 2,}
\end{align*}
\] (20.33)
where \( d_0 \) measures the degree of attachment to location of the poor, with higher attachment equivalent to lower mobility. Given the tax pair \((t, t^*)\), the payoff of a rich individual with preference \( x \) is
\[
\begin{align*}
&[1 - t] - d_1 x \quad \text{in region 1,} \\
&[1 - t^*] - d_1 [1 - x] \quad \text{in region 2.}
\end{align*}
\] (20.34)

Given the tax policies, the population is divided between the two regions. The proportion of poor joining region 1, \( x = x_0 \), is defined by the type that is indifferent between the two regions, so
\[
b - d_0 x_0 = b^* - d_0 [1 - x_0],
\] (20.35)
and thus
\[
x_0 = \frac{1}{2} + \mu_0 \left[ \frac{b - b^*}{2} \right],
\] (20.36)
with \( \mu_0 = \frac{1}{d_0} \) denoting the mobility of the poor. Higher mobility of the poor increases their migration in response to transfer differential. The poor are evenly distributed across regions in case of uniform transfers \( b = b^* \).

Similarly the proportion of rich, \( x = x_1 \), joining region 1 is given by the indifference condition
\[
[1 - t] - d_1 x_1 = [1 - t^*] - d_1 [1 - x_1].
\] (20.37)
Defining the mobility of the rich by \( \mu_1 = \frac{1}{d_1} \) yields
\[
x_1 = \frac{1}{2} + \mu_1 \left[ \frac{t^* - t}{2} \right],
\] (20.38)
with equal taxes inducing equal division of the rich between the two regions, \( x_1 = \frac{1}{2} \). Taxing more drives out some of the rich (i.e., \( x_1 \) decreases with \( t \)).

Budget balance implies that the transfer a region can afford to pay depends on who it attracts. The transfer paid to each poor resident is

\[
\begin{align*}
\text{in region 1,} \\
\text{in region 2.}
\end{align*}
\]

(20.39)

Assume that both governments follow a policy of maximal redistribution. Then region 1 sets its tax rate \( t \), taking as given the tax rate of the other, \( t^* \), so as to maximize the transfer given to its poor residents, \( b \), correctly anticipating the induced migration. The migration response of the rich to a small tax change is proportional to their mobility

\[
\frac{dx_1}{dt} = -\frac{\mu_1}{2} < 0.
\]

(20.40)

How the poor respond to a small tax change depends on the migration response of the rich and is given by total differentiation

\[
dx_0 = \frac{\mu_0}{2} \left[ \frac{d(b - b^*)}{dt} dt + \frac{d(b - b^*)}{dx_0} dx_0 \right].
\]

(20.41)

Evaluating this expression around \( t = t^* \) (i.e., with \( x_i = \frac{1}{2} \)) for separate changes in \( t \) and \( x_0 \) gives

\[
\frac{d(b - b^*)}{dt} = \frac{x_1 + \frac{dx_1}{dt}}{x_0} - \frac{t^* d(1 - x_1)}{1 - x_0} = 1 - 2t \mu_1,
\]

(20.42)

\[
\frac{d(b - b^*)}{dx_0} = -\frac{dx_1}{[x_0]^2} - \frac{t^* [1 - x_1]}{[1 - x_0]^2} = -4t.
\]

(20.43)

Therefore the migration response of the poor to a domestic tax change can go either way,

\[
\frac{dx_0}{dt} = \frac{\frac{1}{2} - t \mu_1}{\mu_0 + 2t} \gg 0.
\]

(20.44)

It is worth noting that more taxation can drive out the poor: \( \frac{dx_0}{dt} < 0 \) if \( t > \frac{1}{2\mu_1} \). The reason is that if the rich are sufficiently mobile (high \( \mu_1 \)), a tax increase induces so many rich to leave that the poor will find it better to follow them. In such circumstances the poor will chase after the rich.
Putting these points together the (symmetric) equilibrium tax choice can be determined. Region 1 chooses its tax rate $t$ so as to maximize the transfer to its poor residents $b$ taking as given the tax choice of the other region and correctly anticipating the migration responses of the rich and the poor. The necessary first-order condition is

$$\frac{db}{dt} = \frac{x_1}{x_0} + \left[ \frac{t}{x_0} \right] \frac{dx_1}{dt} + \frac{db}{dx_0} \left[ \frac{dx_0}{dt} \right] = 0.$$  \hspace{1cm} (20.45)

Using the migration changes as given by (20.40) and (20.44) and evaluating the condition at the symmetric outcome in which both regions pick the same tax-transfer scheme and each group divides evenly between the two regions, we obtain the first-order condition

$$\frac{db}{dt} = 1 - t \mu_1 - 2t \left[ \frac{1 - t \mu_1}{\mu_0 + 2t} \right] = 0.$$  \hspace{1cm} (20.46)

This gives the following symmetric equilibrium

$$t = t^* = \frac{1}{\mu_1 - \mu_0}.$$  \hspace{1cm} (20.47)

Consequently the equilibrium level of redistributive taxation is inversely proportional to the difference in the mobility of the rich and the poor. Higher mobility of the rich reduces taxation, but this is partially offset by the mobility of the poor. The reasoning behind this is that in equilibrium the poor chase after the rich, so it is not possible for the rich to detach themselves from the poor.

The same logic applies in modelling capital and labor mobility. By extension, the possibility for taxing capital increases with the mobility of labor, so the problem of tax competition is more connected to the relative mobility of capital with respect to the mobility of labor.

### 20.3.3 Race to the Bottom

In a context where there are no legal barriers to migration, so that the forces of fiscal competition are at work, any attempt at redistribution or the provision of social insurance in a country would be impossible because it would induce emigration of those who were supposed to give (the rich) and immigration of those who were supposed to receive (the poor). The most extreme predictions of this form imply a “race to the bottom” but
receive little theoretical or empirical support. This is probably due to the presence of significant costs and barriers to migration.

For example, welfare shopping is discouraged in Europe by portability limits between member states and by the eligibility requirement of previous employment in the country. However, underprovision of social insurance in an integrated market is an issue that cannot be ignored in the European Union. Even if it has not been a pressing issue to date, fiscal competition for capital and labor factors has already arrived. The Irish success with reduced corporation taxes suggests that with the EU enlargement this issue will become ever more evident.

20.4 Intergovernmental Transfers

The reasons for organizing intergovernmental transfers are twofold: efficiency and redistribution. We consider the two in turn.

20.4.1 Efficiency

A critical insight from the analysis of tax competition is that increasing the tax rate in one region benefits other regions by increasing their tax bases. We now consider how transfers between regions can be employed to secure efficiency.

If we take the tax base to be the capital stock and the aggregate supply of capital to be fixed, then the tax-induced outflow of capital from the region taxing more represents an inflow of capital to the other regions. In particular, another region $j$ benefits from increased revenue by the amount $t_j \Delta k_j$, where $t_j$ is its tax rate and $\Delta k_j$ the fiscally induced capital inflow. The problem facing each region is to choose the tax rate on capital to finance the public good level that maximizes the welfare of its residents subject to the budget constraint $G = tk(t)$. The optimal regional level of a public good is given by the fact that the marginal benefit of public good $MB$ must be high enough to not only cover its marginal cost, $MC$, but also to offset the negative impact of capital outflow on tax revenue, denoted by $t \Delta k < 0$. Then, following the modified Samuelson rule, we obtain

$$MB = MC - t \Delta k.$$  \hspace{1cm} (20.48)

The context of identical regions provides a useful reference for isolating the fiscal-externality inefficiency from other equity and efficiency aspects that would arise when regions differ and choose different tax rates and public good levels. With identical tax
rates, $t$, the cost of a capital outflow from one region is exactly offset by the benefits from the resulting capital inflows to other regions. It follows that if regions were to take into account such external benefits, they would no longer perceive capital outflows as a cost. The efficient provision of public good (as given by the usual Samuelson rule) would then obtain with

$MB = MC$. \hspace{1cm} (20.49)

The central authority can achieve this efficient outcome by means of revenue-matching grants. The idea is to correct the externality by providing a subsidy to the revenue raised by each region. The matching rate to a region is the additional revenue that accrues to other regions when this region raises its tax rate. Then regions are correctly compensated for the positive externalities generated when they raise their taxes.

Differences among regions bring about a second inefficiency from tax competition, namely that different tax rates induce a misallocation of capital across regions, such that the marginal product of capital is relatively high in high-tax regions (see figure 20.3). It follows that matching rates should be differentiated to induce all regions to choose the same tax rate and at the same time to internalize the fiscal externality. Tax harmonization requires the payment of a higher subsidy to regions with a low preference for taxation and public goods. In practice, however, the central government may not have the political authority nor the information required to impose differentiated matching grants. The information problem is rather severe because all regions can claim to be of the low-tax type in order to obtain a higher subsidy. With tax overlap, the matching rate will be negative to represent the reduction in tax revenues for other levels of government when the region raises additional revenue from the common tax base (see figure 20.4).

Expenditure externalities can also be corrected with expenditure-matching grants. For example, spending by a local government on education or public infrastructure improves the potential earnings of its residents by making them more productive, and this will increase the federal government’s revenue from income, payroll, and sales taxes. To induce the local government to internalize this vertical expenditure externality, the federal authority can use expenditure-matching grants. Matching grant programs specify that the federal government matches on a dollar-for-dollar basis local expenditure up to some maximum. The effect is to lower the price of local public goods and thereby offset the tendency for local governments to under provide public expenditures generating positive externalities.

An important example of vertical expenditure externalities in Canada is the substitutability between expenditures on unemployment benefits at the federal level and
welfare benefits at the provincial level. If the federal government reduces unemployment benefits or their duration, more people will apply for welfare benefits, increasing spending at the provincial level. Conversely, employment programs by provincial government that allow welfare recipients to regain eligibility to unemployment benefits will lead to higher spending at the federal level.

In the United States, until 1996, the federal government could bear 50 to 80 percent of the cost of some welfare expenditures undertaken by states (e.g., Aid to Families with Dependent Children, Food Stamp, and Medicaid programs). In 1996 the AFDC matching system was replaced by a lump-sum grant. Interestingly, as illustrated in figure 20.5, this matching system has proved to be more effective in stimulating local public expenditures than a lump-sum subsidy of the same amount. The reason is simply that the lump-sum grant can be used in any way the recipient wishes, in contrast to the matching grant that is increasing with the amount of public spending. In this perspective, matching grants are also called “conditional” grants because they place some restrictions on their use by the recipient, and the lump-sum grants are called “unconditional” grants. Figure 20.5 also indicates the distortionary effect of matching grants: higher welfare can be attained at the same cost with lump-sum grants. This is the advantage of the freedom of choice that “unconditional” grants provide.

The attraction of matching grants is to internalize expenditure externalities. However, lump-sum grants also have their own attraction, which is to maintain fiscal discipline.
Chapter 20: Fiscal Competition

(at the heart of the EMU). The idea here is that by creating a hard budget constraint, they impose very useful discipline on decentralized expenditure decisions. More generally, a hard-budget constraint implies that decentralized governments must place a basic reliance on their own sources of revenue and must not be overly dependent on transfers from the federal government. Self-financing is a powerful incentive device, and it is essential that local governments do not turn to the federal authority to bail them out of fiscal difficulties by resorting to expansible matching grants.

20.4.2 Redistribution

Intergovernmental grants are also used to channel resources from wealthy jurisdictions to poorer ones. Such transfers are based on equalization formulas that measure the fiscal need and fiscal capacity of each jurisdiction, locality, or province. Fiscal equalization then involves higher grants to those jurisdictions with the greatest fiscal need and the least fiscal capacity. If the objective is to equalize taxable capacity, the central government can supplement the revenue base of poorer jurisdictions by matching any revenues they collect by the addition of a further percentage. This form of equalization is sometimes called “power equalization.”

In practice, equalization grants play a major role in countries like Australia, Denmark, Canada, Germany, Sweden, and Switzerland and involve substantial transfers from wealthy to poor jurisdictions. In the United States such equalizing grants have never played an important role in allowing poorer states to compete effectively with fiscally stronger ones, but the equalization formula has been the basis of local school district finance in many states.

Typically an equalization system sets the transfer to each region equal to the difference between its observed tax base and the average tax base of all regions, multiplied by some standard tax rate, usually equal to the average tax among all the regions. Accordingly, if $b_i$ is the tax base of region $i$ with $\bar{b}$ the average tax base among regions and $\bar{t}$ the average tax rate, then the equalization transfer to region $i$ is given by

$$T_i = \bar{t}(\bar{b} - b_i) \geq 0 \quad \text{for} \quad b_i \leq \bar{b}.$$  \hspace{1cm} (20.50)

The use of the average tax rate as the standard tax rate is to accommodate a diversity of regional spending behavior. The intention is that equalization compensates for a difference in fiscal capacities but not for the difference in preferences for public spending. Indeed, when all regions choose the same tax rate, the formula guarantees equal revenues. The equalization formula can also correct for fiscal externalities. A tax cut by one region increases not only its tax base at the expense of other regions but
also relative to the average tax base, thereby reducing the entitlement of this region to equalization grants.

Fiscal equalization is a contentious issue. In some cases, as in Canada, it may provide the cement that holds the bricks of the federation together. In other cases, like Italy or Belgium, it may become the source of division, where rich regions weary of large and durable transfers to poor regions, actually seek the break up of the federation.

### 20.4.3 Flypaper Effect

When considering the budgetary decisions of the recipients of intergovernmental grants, models of rational choice suggest that the response to a lump-sum grant should be roughly the same as the response to an equal increase in income resulting from a federal tax cut. But empirical studies of the response to grants have rejected this equivalence. There is instead strong evidence that local government spending is more sensitive to grants than it is to increasing income through tax cuts.

Among the best estimates of this for the United States: the marginal propensity for state and local governments to spend out of personal income in the state is about 10 percent, but the marginal propensity for state and local governments to spend out of grants from the federal level is around 80 to 90 percent. This has been known as the “flypaper effect” to say that money sticks where it hits. This is intriguing because it suggests that the same budget could give rise to different choices depending on what form the increment to the budget takes. It has been suggested that this may reflect the behavioral regularity that money on hand (from grants) has a different effect on spending than when the money must be raised (by taxation).

This can be understood with the following thought experiment. Say you have lost your ticket for the cinema and you must decide whether to buy a new one. Now suppose instead that you lose the same amount of money, would you be as willing to buy the ticket in the first place. Although in both cases you face exactly the same budget constraint, it is less likely that you will buy the new ticket after losing the original one than if you lost the equivalent amount of money.

The fly paper effect also casts serious doubt as to the notion that local governments are more responsive to local demand. Indeed, from the estimates above, one might think that if the local government were strictly responding to local demand, $100 per capita of federal grants would lead to about $90 per capita tax reduction and $10 additional spending. It is, of course, entirely the other way around, with about $90 additional spending and $10 of reduced local taxes.
20.5 Evidence

20.5.1 Race to the Bottom

The central result of the tax competition model is that increasing mobility of capital will drive down the equilibrium tax on capital. This canonical model is at the heart of concerns about capital tax competition within the European Union. In response to this growing concern, the OECD published a report (OECD 1998) comprising about twenty recommendations to counter what was perceived as “harmful” tax competition of capital income. This issue was also taken seriously by the European Union in December 1997. The EU commission agreed to a “Code of Conduct” in business taxation, as part of a “package to tackle harmful tax competition.” The Code is aimed at identifying tax measures that reduce the level of tax paid below the “usual” level. In particular, a measure is considered “harmful” if the tax advantage is restricted only to nonresidents, or if it is “ring-fenced” from the domestic market, or if the tax break is granted without any real economic activity taking place.

The central motivation for these reforms is the race to the bottom in capital taxation. To appreciate the relevance of this, we should evaluate the existence and magnitude of this race to the bottom. Table 20.1 shows the statutory corporate income tax rates in 1982 and 2001 for a group of EU and G7 countries. The statutory tax rate includes local tax rates and any supplementary charges made. Except for Italy and Ireland, all countries have significantly reduced their statutory tax rate. In 1982, Ireland had the lowest rate at 10 percent and Germany the highest rate at 64 percent, while both the United States and the United Kingdom had a rate around 50 percent. In 2001, Ireland had still the lowest rate at 10 percent but both Germany and the United States had reduced their rate just below 40 percent, and the rate in the United Kingdom was down to 30 percent. Over the same period Austria, Finland, and Sweden had cut their statutory rate by more than one-half.

Table 20.2 shows the fall in the median statutory corporate income tax rates over the last two decades for the same group of countries. Between 1982 and 2001, the median statutory tax rate for this group fell from 50 percent to 35 percent. The statutory tax rate is likely to be important in determining the incentive for firms to shift investment between countries. However, the tax base is also likely to be relevant. A higher tax rate does not necessarily imply higher tax payments, since effective tax payments also depend on the definition of the tax base. Governments with different tax rates can also adjust their rates of depreciation allowances for capital expenditure. The rate allowed
for firms to spread the cost of capital against tax varies considerably across countries. Adjusting the statutory tax rate to take account of this effect and other difference in tax base, we obtain the “effective” average tax rate. It measures the proportion of total profit taken in tax. The evolution of the “effective” rate does not replicate the statutory rate. There is a decline from 43 percent in 1982 to 32 percent in 2001, but the fall is less pronounced than for the statutory rate. The lower fall in the effective rates indicates that the reduction in the statutory rates has been partially offset by less generous allowances for capital expenditures (broader tax base).

Table 20.1
Statutory corporate income tax

<table>
<thead>
<tr>
<th></th>
<th>1982</th>
<th>2001</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austria</td>
<td>61</td>
<td>34</td>
</tr>
<tr>
<td>Belgium</td>
<td>45</td>
<td>40</td>
</tr>
<tr>
<td>Canada</td>
<td>45</td>
<td>35</td>
</tr>
<tr>
<td>Finland</td>
<td>60</td>
<td>28</td>
</tr>
<tr>
<td>France</td>
<td>50</td>
<td>35</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>53</td>
<td>30</td>
</tr>
<tr>
<td>Germany</td>
<td>62</td>
<td>38</td>
</tr>
<tr>
<td>Greece</td>
<td>42</td>
<td>38</td>
</tr>
<tr>
<td>Ireland</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Italy</td>
<td>38</td>
<td>40</td>
</tr>
<tr>
<td>Japan</td>
<td>52</td>
<td>41</td>
</tr>
<tr>
<td>Netherlands</td>
<td>48</td>
<td>35</td>
</tr>
<tr>
<td>Portugal</td>
<td>55</td>
<td>36</td>
</tr>
<tr>
<td>Sweden</td>
<td>61</td>
<td>28</td>
</tr>
<tr>
<td>United States</td>
<td>50</td>
<td>49</td>
</tr>
</tbody>
</table>

Source: Devereux et al. (2002).

Table 20.2
Statutory and effective corporate income tax rates

<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>Median statutory</td>
<td>50</td>
<td>48</td>
<td>46</td>
<td>43</td>
<td>39</td>
<td>38</td>
<td>37</td>
<td>36</td>
<td>37</td>
<td>35</td>
</tr>
<tr>
<td>Average effective</td>
<td>43</td>
<td>42</td>
<td>41</td>
<td>38</td>
<td>36</td>
<td>37</td>
<td>36</td>
<td>36</td>
<td>34</td>
<td>32</td>
</tr>
</tbody>
</table>

Source: Devereux et al. (2002).
20.5.2 Race to the Top

It is natural for economists to think that competition among jurisdictions can stimulate public decision makers to act more efficiently and limit their discretion to pursue objectives that are not congruent with the interest of their constituency. Tests of this hypothesis have led to substantial empirical research investigating whether intergovernmental competition through fiscal decentralization affects public expenditures.

The evidence as reviewed in Oates (1999) supports strongly the conclusion that increased competition tends to restrict government spending. But the fact that spending falls with more competition does not mean that resources are more efficiently allocated as competition increases. The problem is that it is hard to come up with measures of the quality of locally provided public services. However, there is one notable exception, that is education where standardized test scores and postgraduation earnings provide performance measures that are easily comparable across districts. Following this strategy, Hoxby (2000) finds that greater competition among school districts has a significant effect both in improving educational performances and reducing expenditures per student.

Besley and Case (1995) develop and test a political model of yardstick competition in which voters are poorly informed about the true cost of a public good provision. They use data on state taxes and gubernatorial election outcomes in the United States. The theoretical idea is to see how much of a tax increase is due to the economic environment and how much to the quality of their local government, so voters can use the performance in other jurisdictions as a “yardstick” to obtain an assessment of the relative performance of their own government. The empirical evidence supports the prediction that yardstick competition does indeed influence local tax setting. From that perspective intergovernmental competition is good for disciplining politicians and limiting wasteful public spending.

20.5.3 Tax Mimicking

A substantial body of empirical studies has emerged testing for interdependence among jurisdictions in tax and expenditure choices. An early and very influential piece of work is by Case et al. (1993) who test a model in which a state’s expenditure may generate spillovers to nearby states. The novelty of this work is that it allows for spatially correlated shocks as well as spillovers. Data from a group of states showed strong evidence of fiscal interdependence and the effects arising from interdependence to be
large. A dollar increase in spending in one state induced neighboring states to increase their own spending by seventy cents.

Brueckner and Saavedra (2001) test for the presence of strategic competition among local governments using data for seventy cities in the Boston metropolitan area. Where capital is the mobile factor and population fixed, local jurisdictions were found to choose property tax rates that take into account the mobility of capital in response to tax differentials. Property taxes are the only important source of local revenue. The authors use spatial econometric methods to relate the property tax rate in one community to its own characteristics and to the tax rates in competing communities. They found that the tax rate in one locality is positively and significantly related to tax rates in contiguous localities. This means that the tax interdependence generates upward-sloping reaction functions. The same conclusion was obtained with similar methodology by Heyndels and Vuchelen (1998) in their study of property-tax mimicking among Belgian municipalities.

Turning to welfare migration, Saavedra (1998) uses spatial econometric estimates of cross-sectional welfare benefits (AFDC) for the years 1985, 1990, and 1995 for all states in the continental United States. She finds strong evidence that a given state’s welfare benefit choice is affected by the benefit levels in nearby states for each year. Moreover her findings show significant and positive spatial interdependence, suggesting that a given state will increase its benefit level as benefits in nearby states rise.

20.6 Conclusions

Competition may be thought of as a means to secure better fiscal performance, or at least exposes fiscal inefficiency. If market competition by private firms provides households with what they want at least cost, why cannot intergovernmental competition lead to better governmental activities? Poorly performing governments will lose out, and better performing ones will be rewarded. Although appealing, the analogy can be misleading, and the competitive model is not directly transferable to fiscal competition among governments. Once there is more than one jurisdiction, the possibility is opened for a range of fiscal externalities to emerge. Such externalities can be positive, as with tax competition, and lead to tax rates that are too low. Competition among governments to render high-quality services may give way to competition for undercutting tax rates to attract mobile factors from neighboring jurisdictions. Given capital mobility, any attempt by local government to impose a net tax on capital will drive out capital until its net return is raised to that available elsewhere. The revenue gain from the higher
A tax rate will be more than offset by an income loss to workers due to the reduction in the locally employed capital stock. Fiscal harmonization across jurisdictions would be unanimously preferred. Alternatively, the externalities can be negative, as with tax overlapping, and put upward pressure on tax rates. This is the common pool problem leading to overspending and overtaxation.

Interestingly, when there is no clear division of power, there is competition between central and local governments not only for the same tax base but also for the same local voter base. The federal government is competing with the local government for the provision of what might otherwise be local government services (child care, education, police, etc.). The implication of that is the overexpansion of public spending. To make things worse, tax bases as well as consumers of public services are mobile and need not move together. By voting with their feet to sanction inefficient government, consumers of public services can move to escape fiscal obligations rather than to obtain efficient public services. At the extreme, given household mobility, public services may be used and benefits may be received in one region, while income is derived and taxes are paid in another.

Such externalities can be corrected, and intergovernmental transfers are one means of doing so. Grants may be either conditional (matching grants and categorical grants) or unconditional (lump-sum block grants). Each type of grant involves different incentives and induces different behaviors of local governments. The final mix of increased expenditure versus lower local taxes depends on the preferences guiding local choices. Empirical studies are essential to compare the costs and benefits of intergovernmental competition. The presence of fiscal interaction between jurisdictions is not compelling evidence of harmful tax competition. Tax interactions can also be due to a political effect where electoral concern induces local governments to mimic tax-setting in neighboring jurisdictions. In such a case competition can be an effective instrument to discipline and control officials.

Recall the question raised at the beginning of this chapter on the analogy between market competition and government competition. The main lesson from the fiscal competition theory is that intergovernmental competition limits the set of actions and policies available to each government. There is no doubt that the constraints that are imposed on the authority of governments do constrain or limit actions, and this way both “good” and “bad” actions can be forestalled. Whether or not we view intergovernmental competition as harmful depends on our perception of the quality of governments. Unconstrained actions of a benevolent governments are good, but they can be very costly when governments abuse power.
Further Reading

The earliest studies on fiscal interaction due to interjurisdictional spillovers were:


The competitive version of tax competition was first analyzed by:


Further contributions include:


The pioneering studies of strategic tax competition are:


Asymmetric tax competition is analyzed in:


The public input model is analyzed and empirically tested in:


Mobility and community formation is in:


Chapter 20: Fiscal Competition


Mobility and redistribution are discussed in:

On the design of intergovernmental transfers, see:

Tests of the Tiebout hypothesis are described in:

Evidence on fiscal competition is given in:

For evidence on efficiency-enhancing competition, see:

For evidence on strategic fiscal interaction, see:


**Exercises**

20.1 Why is source-based taxation of capital income likely to induce undertaxation of capital?

20.2 How does source-based taxation compare with residence-based taxation for the efficiency of cross-border investment?

20.3 What are the implications of source-based taxation for the efficiency of cross-border investment?

20.4 Consider a world economy consisting of $N$ identical countries, each endowed with one unit of labor. Labor is immobile. The world economy also contains one unit of capital that is freely mobile across countries. All countries have identical production functions given by $F(L, K) = L^{3/4}K^{1/4}$, where $L$ denotes labor and $K$ denotes capital. The price of output is fixed at $1$.

   a. Suppose that none of the countries tax either capital or labor. Find the equilibrium interest rate and allocation of capital across countries. What is the total income received by capitalists (the owners of the fixed factor of production) and workers? Evaluate the interest rate and income levels for $N = 2$ and $N = 20$.

   b. Consider the impact of a tax at rate $\tau$ on capital income in country 1 if other countries do not tax capital income. Assume that tax revenues are used to buy output at the fixed price of $1$. What is now the after-tax return on capital invested in country 1? What is the equilibrium interest rate and allocation of capital across countries? Find the total income received by capitalists and workers and the tax revenues in country 1 as a function of $N$ and $\tau$.

20.5 Consider the exercise above and set the capital income tax rate in country 1 at $\tau = 0.20$.

   a. Find the change in total capital income, total labor income, and the revenue raised in country 1 for $N = 2$ and $N = 100$.

   b. What happens to the before-tax marginal product of capital in the countries without taxes for $N = 2$ and $N = 100$?

   c. Are the workers in country 1 better off as a result of the tax? What about the impact of the tax on the welfare of the workers in nontaxing countries?

   d. Discuss the tax-shifting between workers and capitalists as the number of countries increases from $N = 2$ to $N = 100$.

20.6 Consider a world economy composed of two countries, $A$ and $B$. There is a fixed stock of capital, $K$, that allocates between the countries on the basis of the after-tax return. Let the after-tax return in country $i$, $i = A, B$, be equal to $r - mK^i - t^i$, where $K^i$ is the quantity of capital that locates in country $i$ and $t^i$ is the tax rate in country $i$.

   a. Provide an interpretation of the parameter $m$. 

---

**Notes:**

- **Source-Based Taxation:**
  - Source-based taxation generally refers to the taxation of income earned in a jurisdiction where the source of income is located, as opposed to residence-based taxation, which bases taxation on the residence of the individual.

- **Cross-Border Investment:**
  - Cross-border investment involves investments made by individuals or entities in one country into another country.

- **Equilibrium Interest Rate:**
  - The equilibrium interest rate is the rate at which the supply of capital equals the demand for capital, typically associated with the market clearing condition.

- **Tax Revenues:**
  - Tax revenues are the amounts raised by taxation and are used to fund public goods and services or to reduce government spending in the form of transfers or subsidies.

- **Before-Tax Marginal Product of Capital:**
  - The before-tax marginal product of capital is the additional output that can be obtained by increasing the quantity of capital by a small amount, measured at the existing price.

- **Tax-Shifting:**
  - Tax-shifting involves the redistribution of the burden of taxation between workers and capitalists, often as a result of different tax regimes or tax policies across jurisdictions.
Chapter 20: Fiscal Competition

b. Assuming that each country chooses its tax rate to maximize tax revenue, calculate the Nash equilibrium choice of tax rates.

c. What is the effect on the equilibrium tax rates of reducing $m$? Explain this result.

d. What are the efficient tax rates?

e. What are the implications of these findings for tax policy?

20.7 Consider an economy a fixed stock of capital $K$. The capital stock is mobile between regions. The residents of region $i$ is denoted $K_i$ and the (fixed) labor force in $i$ by $L_i$. The residents of region 2 are considering the imposition of a tax $t$ on each unit of capital. Wage rates ($w_1$, $w_2$), outputs ($y_1$, $y_2$) and national returns to capital ($r$) are determined by market clearing. The two regions produce the same good, which acts as numéraire. The production function in country $i$ is $y_i = K_i^{1-\alpha} L_i^\alpha$.

a. Show that $w_i = \frac{\alpha K_i^{1-\alpha} L_i^\alpha}{L_i}$, $i = 1, 2$.

b. Show that $r + t_i = \frac{[1-\alpha]K_i^{1-\alpha} L_i^\alpha}{K_i}$.

c. Use part b and the market-clearing condition for capital, $K = K_1 + K_2$, to show that $K = L_1 \left[ \frac{1-\alpha}{r} \right]^{1/\alpha} + L_2 \left[ \frac{1-\alpha}{r + t} \right]^{1/\alpha}$.

d. Now assume that regional capital ownership is proportional to regional workforce. If $\alpha = 0.5$, $L_1 = 100$, and $L_2 = 200$, which policy gives region 2 a higher level of national income?
   i. Region 2 imposes no tax ($t = 0$).
   ii. Region 2 imposes a tax of $t = 0.15$.
   iii. Region 2 imposes a subsidy of $t = -0.15$.

20.8 Consider the economy of the exercise above but now assume that labor is also mobile and that region $i$ levies a tax on capital $t_i \geq 0$.

a. Show that $w = \frac{\alpha K_i^{1-\alpha} L_i^\alpha}{L_i}$, $i = 1, 2$.

b. Show that $r + t_1 = \frac{[1-\alpha]K_i^{1-\alpha} L_i^\alpha}{K_i}$, $r + t_2 = \frac{[1-\alpha]K_i^{1-\alpha} L_i^\alpha}{K_i}$.

c. Demonstrate that the result in part a implies $K_i L_i = K_i L_i$. How can this be consistent with the result in part b?

d. What will be the optimal taxes for the two regions if they wish to maximize national income?

20.9 A car manufacturer can choose to locate a new plant in country $A$ or country $B$. Your job is to determine where to locate this new plant. The only inputs used in car production are labor and capital, and the production function is Cobb–Douglas: $F(L, K) = L^{1/2} K^{1/2}$, where $L$ is the labor input and $K$ the capital input. In country $A$, labor costs $7 per unit and capital costs $7 per unit, while in country $B$, labor costs more ($8) but capital costs less ($6).

a. In which country should you locate the new plant so as to minimize cost per unit of output (i.e., average cost)?
b. Now assume that country A subsidizes labor so that labor costs $6 per unit in country A. Does it change the location decision of the firm?

c. Instead of subsidizing labor, suppose that country A subsidizes capital so that capital costs $6 per unit. Does it change the location decision of the firm?

d. What happens if both countries act identically in either taxing or subsidizing capital and labor? What would be the location decision of the firm? Has any country an incentive to alter its tax-subsidy choice, and if so, how?

20.10 The federal government would like to give grants to local states to promote free medical vaccinations for children. However, because of a limited budget the federal government wishes to target its spending as efficiently as possible.

a. Define what would be a matching grant, and explain why it could be preferable to a block grant.

b. Now consider a closed-end matching grant that matches state spending dollar for dollar up to a specified amount, at which point the subsidy is phased out. Do you think a closed-end matching grant is preferable to an open-end matching grant?

20.11 Consider two countries, 1 and 2, with population shares $\lambda_1$ and $\lambda_2$ such that $\lambda_1 + \lambda_2 = 1$. There is a world stock of capital, equal to 1, to be allocated between the two countries. The production function is $f(k_i) = k_i - \frac{k_i^2}{2}$, $k_i$ standing for per capita capital invested in country $i = 1, 2$. Production factors are paid their marginal productivity. Capital moves freely between the two countries. Denote its net price by $\rho$. The representative consumer in country $i = 1, 2$ owns $\bar{k}_i$ units of capital. Country $i$ chooses $t_i$ so as to maximize the net income of its representative consumer:

$$y_i = f(k_i) - f'(k_i)\bar{k}_i + \rho\bar{k}_i + t_i\bar{k}_i.$$

a. What are the first-best tax levels $t_1$ and $t_2$ in this setting?

b. Express the demand for capital in country $i$ as a function of $\rho$ and $t_i$. Using the equilibrium condition of the international capital market, $\lambda_1k_1 + \lambda_2k_2 = 1$, solve for $\rho$ as a function of tax rates and population shares. If $t_1$ increases, how will $\rho$ react? How do you interpret this result?

c. Show that each country’s objective function can also be written as $f(k_i) + \rho [\bar{k}_i - k_i]$.

d. Find the first-order condition for country $i$’s choice of the tax rate, and show how the sign of the tax rate depends on the net exporting position of the country.

e. Suppose now that $\lambda_1 = \lambda_2 = \frac{1}{2}$, $\bar{k}_1 = \frac{1}{2}$ and $\bar{k}_2 = \frac{3}{2}$. Draw the tax response functions, and find the equilibrium taxes and the corresponding values of $\rho$, $k_1$, and $k_2$. Is this equilibrium outcome efficient? How can you measure the inefficiency of tax competition?

f. Express equilibrium taxes as functions of the parameters $\lambda_1$, $\lambda_2$, $\bar{k}_1$, and $\bar{k}_2$, and show that the equilibrium taxes must be of opposite signs.

20.12 Which factor of production is more mobile: capital or labor?

20.13 Consider a linear city composed of two different jurisdictions. The two extremes of the city are $-\frac{1}{2}$ and $\frac{1}{2}$, and the city center is located at 0. Jurisdiction 1 is located to the left of the city center and jurisdiction 2 to the right. The city is inhabited by a continuum of individuals that are uniformly distributed with a total mass of 1. Each jurisdiction has a population of $\frac{1}{2}$. The
median voter of jurisdiction 1 is located at $-\frac{1}{4}$, while the median voter of jurisdiction 2 is located at $\frac{1}{2}$. Each jurisdiction has a job center, located respectively at $-\frac{1}{2}$ and $\frac{1}{2}$. Wages are exogenous and equal to $\gamma > 1$ in jurisdiction 1 and 1 in jurisdiction 2. There are ad valorem taxes on wages $t_1$ and $t_2$ paid to the jurisdiction where work is undertaken. Individuals incur a commuting cost equal to $\frac{1}{2}$ per unit of distance to their workplace. Jurisdictions are run by majority-elected governments that maximize net income. Net income is the sum of the median voter’s net wage plus fiscal revenue per capita, $[1 - t_i] w_j + t_j N_j w_j$ when the median voter in $j$ works in jurisdiction $i$, with $N_j$ denoting the number of people working in $j$.

a. Suppose that each individual works in the jurisdiction where she gets the highest after-tax wage net of commuting costs. Compute the number of workers $N_1, N_2$ in each jurisdiction as a function of the wage taxes $t_1$ and $t_2$.

b. Suppose that in equilibrium jurisdiction 2’s median voter works in jurisdiction 1. Find the Nash equilibrium in taxes. Do the jurisdictions tax or subsidize wages? Explain briefly. Give a condition on $\gamma$ under which this equilibrium exists. (Hint: In such an equilibrium the median voter in 2 chooses to work in jurisdiction 1.)

c. Suppose now that in equilibrium jurisdiction 2’s median voter works in jurisdiction 2. Find the Nash equilibrium in taxes. Does either jurisdiction subsidize the wage? Why or why not? Give a condition on $\gamma$ such that this equilibrium exists. Compare your answer to that of part b. Interpret the difference.

20.14 The 1996 welfare reform in the United States devolved responsibility for welfare programs to the states. Does the theory suggest it was the right the policy?

20.15 What is meant by the “race to the bottom”? Does evidence suggest that such a race is occurring?

20.16 In many countries the capital owned by nonresidents is taxed. What are the effects of such tax exporting? Is it profitable to subsidize capital owned by nonresidents?
21.1 Introduction

The recent explosion of globalization has followed a reduction of institutional barriers to the free movement of goods, services, and factors of production. The effect that this has had on trade between nations is dramatic: world GDP has increased by a multiple of five since 1960 but the volume of world trade has increased by a multiple of eleven. It has also become easier and cheaper to move production facilities and corporate headquarters between countries. This in turn allows taxable profits to be shifted between tax jurisdictions. The political landscape has changed as well, with national governments having to accept constraints upon their ability to unilaterally implement policy. These constraints can derive from membership of institutions such as the European Union, from agreement to tax treaties, and from the mobility of the tax base.

The series of issues we consider in this chapter derive from the incentives of firms and consumers to exploit international tax differentials to their advantage. The process of globalization has enhanced these incentives with consequences for the tax choices of individual countries and the willingness to participate in international agreements. Some of the issues have been brought into sharp focus by the continuing evolution of the European Union. The ever-closer economic union between member states and, in particular, the completion of the single market in 1993 increased the mobility of firms and made it legal for consumers to purchase anywhere in the European Union at the local tax rate. This has raised questions about the design of an appropriate tax structure for a customs union and has lead to several new policy initiatives. The issues facing the European Union have been the motivation for much recent research but the process of globalization makes many of the issues of significant worldwide relevance.

The first section considers the characterization of international efficiency and emphasizes that efficiency has to be defined relative to mobility. One consequence of globalization has been the increased international mobility of capital. We look at the implications of mobility for the taxation of capital in general, and the corporation in particular. Increased internationalization of firms with production in many different countries allows a firm to choose where to produce and, to some extent, where to earn profit. Firms can relocate profit through the transfer prices used to account for
transactions between divisions. We analyze the effect of taxation on transfer pricing and assess policies used to mitigate the practice. The focus is then placed more directly on the issue of taxation and location choice. Some empirical evidence is reviewed. The chapter then moves onto the design of indirect taxes in an international setting. The benefits of harmonization are considered and alternative tax principles are assessed. The chapter is completed by a study of tariff policy and the role of trade agreements in liberalizing trade.

21.2 International Efficiency

In a closed economy production efficiency is achieved when the ratio of the marginal products of any two inputs is the same in the production of all goods. When this condition is attained, it is not possible to reallocate inputs to increase the output of one good without reducing the output of another. The discussion of commodity taxation in chapter 15 demonstrated that Pareto-efficiency is achieved when final consumption goods are taxed but intermediate inputs are not. The key property of a value-added tax (VAT) is that it is designed to achieve precisely this outcome. With a VAT producers can reclaim the tax paid on the inputs they use, so only the value added at each stage of the production process is taxed. At the end of the production process the sum of value added is equal to the final consumption value. This appealing property of VAT is one of the reasons for the increasing number of countries using the system. The United States remains a notable exception, which brings to mind the famous quip about when the United States will finally move to a VAT. As the New York Times reported way back in 1988, Larry Summers predicted that the United States will have a VAT when liberals figure out that it is a money machine and conservatives see that it is a tax on the poor.

In an international setting there are additional considerations that need to be taken into account in order to describe an efficient allocation. The criteria that must be satisfied for efficiency have to include the allocation of production and consumption within and across countries. The important additional point is that the characterization of efficiency across countries depends on the degree of mobility of factors and goods. To see this, consider an extreme case where factors of production are absolutely immobile. In this case efficiency has to be defined with the initial factor allocation taken as given even if a reallocation of factors across countries could achieve an increase in world output. In contrast, if factors are freely mobile, then their allocation across countries must be determined as part of the description of an efficient location.
Chapter 21: Issues in International Taxation

21.2.1 Efficient Allocations

These points are now developed by considering a world with two countries, \( A \) and \( B \), and two inputs, \( K \) and \( L \). The inputs are in fixed supply but may be mobile between the countries. The inputs are used to produce a single output in each country. The efficient allocation of inputs is found by considering the decision problem of a world social planner. The constraints on the optimization reflect the mobility of factors.

Suppose that initially the inputs and the produced good are perfectly mobile between the countries. This implies that it does not matter in which country production takes place. The efficient outcome is then described by the allocation of capital and labor to the two countries that maximizes the sum of outputs

\[
\max_{\{K^A, K^B, L^A, L^B\}} Y^A + Y^B.
\]  

(21.1)

This optimization is subject to the production constraints

\[
Y^A = F(K^A, L^A), \quad Y^B = F(K^B, L^B),
\]  

(21.2)

and the constraints on input use

\[
K = K^A + K^B, \quad L = L^A + L^B.
\]  

(21.3)

Note that the output level can be written \( Y^A + Y^B = F(K^A, L^A) + F(K - K^A, L - L^A) \). So the necessary condition for the allocation of capital is

\[
F_K(K^A, L^A) = F_K(K - K^A, L - L^A),
\]  

(21.4)

and the necessary condition for labor is

\[
F_L(K^A, L^A) = F_L(K - K^A, L - L^A).
\]  

(21.5)

In this case production efficiency requires the marginal product of each input to be the same in both countries. These are arbitrage conditions that require identical returns on each input in the two countries and are a consequence of the mobility of the two factors. These conditions are more restrictive than the more familiar efficiency condition for the closed economy involving the ratio of marginal products. We can divide (21.4) by (21.5) to show that the arbitrage conditions imply the standard efficiency condition

\[
\frac{F_K(K^A, L^A)}{F_L(K^A, L^A)} = \frac{F_K(K^B, L^B)}{F_L(K^B, L^B)}.
\]  

(21.6)
The consequences of different assumptions on mobility are now considered. If one of the factors was immobile, then the efficiency condition for the other factor must be defined with respect to the fixed allocation. For example, if labor was immobile, then there is only the efficiency condition for capital
\[ F_K(K^A, \bar{L}^A) - F_K(K - K^A, \bar{L}^B) = 0. \] (21.7)
In this case the ratio condition (21.6) does not apply because the immobility of labor prevents the ratio of marginal products from being equalized.

Alternatively, if it is assumed that the factors of production are mobile but the final product is not, then the allocation of production between countries matters. The efficient allocation solves
\[ \max_{\{K^A, K^B, L^A, L^B\}} F(K^A, L^A), \] (21.8)
subject to the constraints that
\[ F(K^B, L^B) = \bar{Y}^B \] (21.9)
and
\[ K = K^A + K^B, \quad L = L^A + L^B, \] (21.10)
where \( \bar{Y}^B \) is some fixed level of output in country B. The solution to this optimization is characterized by the standard efficiency condition for input use (21.6) and the constraint (21.9).

These examples illustrate the general principle that the characterization of an efficient allocation depends on the mobility of factors of production and the final product. This observation is important for judging the efficiency, or otherwise, of tax systems and policies.

21.2.2 Markets and Efficiency

In a closed economy the operation of a competitive market ensures that an efficient outcome is achieved. The same claim is valid for an open economy provided that the market has the same degree of mobility of the factors and product as the planned outcome. This need not be the case. For example, capital may be mobile from the perspective of a world social planner, but arbitrary regulations may prevent it being mobile in the market setting. The market will then not achieve efficiency. In the absence
of any differences in mobility, it can be asserted that the competitive market will achieve a Pareto-efficient equilibrium.

The extent to which efficiency is achieved when there is monopoly (or any form of imperfect competition) needs a more detailed analysis. The mobility of the factors and the final product restrict the choices of the firm in the same way that they restricted the choices of the social planner.

Assume first that the product and the factors are both mobile. The total output of the monopolist is divided between the supplies, \( x^A \) and \( x^B \), to the two countries. Hence

\[
x^A + x^B = F(K^A, L^A) + F(K^B, L^B).
\]  

(21.11)

The firm chooses the allocation of inputs to maximize profit

\[
\pi = x^A p^A(x^A) + x^B p^B(x^B) - rK^A - rK^B - wL^A - wL^B,
\]  

(21.12)

subject to (21.11). The assumption of a common rental rate, \( r \), for capital and wage rate, \( w \), for labor is a consequence of the mobility assumption: if the factors are mobile, then the prices must be the same in the two countries. The optimization has necessary conditions

\[
p^A + x^A p'^A = p^B + x^B p'^B,
\]  

(21.13)

\[
F_{K^A} = F_{K^B},
\]  

(21.14)

\[
F_{L^A} = F_{L^B},
\]  

(21.15)

where \( p'^A \) and \( p'^B \) are the derivatives of the inverse demand functions. The first condition determines the allocation of output across markets to equalize marginal revenues. The second and third are the efficiency conditions that the social planner achieves with full mobility. The firm therefore allocates the inputs it uses between the two countries using the same criterion as the social planner.

Consider now the case where the product cannot be transported. Then the output must be used to supply the market in which it is produced. The profit of the monopoly firm is

\[
\pi = \sum_{j=A,B} \left[ F(K^j, L^j) p^j(F(K^j, L^j)) - rK^j - wL^j \right].
\]  

(21.16)

The necessary conditions are

\[
F_{K^A}(p^A + Fp^A) = F_{K^B}(p^B + Fp^B),
\]  

(21.17)

\[
F_{L^A}(p^A + Fp^A) = F_{L^B}(p^B + Fp^B).
\]  

(21.18)
Dividing the conditions for $K^i$, $i = A, B$, by those for $L^i$ gives
\[
\frac{F_{K^A}}{F_{L^A}} = \frac{F_{K^B}}{F_{L^B}} \quad (21.19)
\]

This condition requires that the ratio of marginal products be equal. This is the standard between-firm efficiency condition for the allocation of inputs, and again, it is the same as the efficiency condition for the social planner under these mobility assumptions.

The allocation of inputs by a profit-maximizing monopolist and the allocation by a social planner are characterized by the same efficiency conditions. Therefore the inputs used by the monopolist are allocated efficiently between countries. This does not mean that monopoly will achieve an identical outcome to the social planner because the monopolist produces less output than is efficient in order to increase price and extract a rent. More precisely, what the results show is that given the chosen level of output, the monopolist will efficiently allocate inputs.

### 21.2.3 Taxation and Efficiency

It is now possible to assess the desirability of production efficiency when governments use commodity taxes to raise revenue. In a closed economy the Diamond–Mirrlees lemma shows that production efficiency should be maintained. This conclusion was established by the argument that a move from an inefficient point inside the production set would be beneficial provided that there was a good that all consumers enjoyed (or one that all disliked).

This argument does not extend directly to an open economy. The reason for this is that each country has a separate government budget constraint. The effect of these budget constraints is to make it more difficult to find a direction of movement from a point inside the world production possibility set that benefits all consumers in all countries simultaneously. Since a potentially beneficial move may adversely affect the budget constraint of one or more governments, it will only be actually beneficial if the governments that lose can be compensated by those that gain. This problem does not arise in the single-country case.

The production efficiency argument can only be established if compensatory transfers between countries are permitted. If there are no such transfers, then production inefficiency may be desirable. It is interesting to observe in this context that the operation of the European Union has the features of a transfer scheme between member states. The EU budget is funded by contributions from member states and is used to finance payments to member states. The net effect is equivalent to transfers between
Chapter 21: Issues in International Taxation

member states. There is no claim that the transfers made by the European Union are the ones required to support production efficiency and justify VAT, but it is interesting to observe the existence of a system that could deliver the transfers.

21.3 Capital and Corporations

The growth of globalization has particularly strong implications for the taxation of capital. Globalization has increased the mobility of capital, which has increased the potential for tax competition to occur. Mobility has other implications. It can result in capital being invested in one tax jurisdiction but the return paid to an owner located in another. It can also result in a return being earned by the joint operation of capital located in several different jurisdictions. Such outcomes raise questions about which jurisdiction should have the right to tax the income flows from the capital to the owners. This section addresses these issues in the context of capital taxation generally and, in more detail, for the taxation of corporate profit in an international environment.

21.3.1 Capital Taxes

Capital taxes include taxes on corporate profit, on interest income, on dividends from the ownership of stock, and on the capital gains from asset ownership. It is important for capital taxation in a globalized world that income can arise from an asset located in one country and accrue to an individual located in another country. For example, a firm that is headquartered in the United Kingdom may earn profit by producing in China and paying a dividend to a shareholder who is resident in the United States. This multiplicity of international locations creates complexity in allocating capital income to tax jurisdictions.

Capital income can be taxed on either a source basis or a residence basis. When a source-based tax is in operation, the income from capital is taxed in the country in which the income is generated. For example, a resident of the United States with funds deposited in a German bank account will pay tax in Germany on the interest added to the account. With a residence-based tax, income from capital is taxed in the country of residence of the owner of the capital. For example, a UK resident who owns shares in a US company will pay tax on the dividends in the United Kingdom. The alternative systems have different incidence effects. These effects are now demonstrated by considering a small open economy that takes the world return to capital as given.
The introduction of a source-based system implies that the return offered on investment in the country will have to rise by the amount of tax to ensure that the net-of-tax return is equal to the fixed world return. This can only happen if the capital stock in the country falls (so that the marginal product of capital, and hence the return, rise), which means that capital must flow out of the country. The owners of capital will not be affected by the tax, since they always earn the world return. Instead, the incidence will fall upon the residents of the small open economy who earn income from non-mobile factors of production.

In contrast, the introduction of a residence-based tax will not affect the return earned by investors resident outside the country. This implies that the pre-tax return in the country has to equal the world rate. If it did not, there would be an opportunity for international investors to increase returns by investing in the country. Consequently the post-tax return for domestic investors will be reduced by the level of the tax. The cost of investment for the firms located in the country is not affected by the tax, so investment decisions will not be affected. This places the incidence of the tax on the owners of capital resident in the country.

The different effects of source and residence systems imply that they have different neutrality properties. A tax is defined as being neutral with respect to an economic choice when it does not change the relative values of marginal benefits and marginal costs involved with that choice. The choice involved here concerns the flows of capital into and out of a country, and how the imposition of a tax system affects the location of capital across countries. A source-based tax applies equally to all capital within a country regardless of its country of origin. Therefore the source-based tax achieves capital import neutrality provided that all countries exempt capital income earned abroad from taxation (which they will if they all operate a source-based system). Conversely, the residence-based system treats all capital income the same way regardless of where the income arises. It therefore achieves capital export neutrality provided that all countries practice the system or the domestic country provides a full tax credit for foreign taxes.

The system of taxation in operation can also affect the international pattern of ownership of capital by altering relative return on assets across countries. In many cases the ownership pattern should not matter for the real choices of the corporation, but it can do in some circumstances, such as when intangible assets are involved. When ownership patterns are not distorted, the tax system is defined as satisfying capital ownership neutrality. This can be achieved if countries levy a residence-based capital tax.

The choice of system for capital taxation has become increasingly important with increased globalization and the enhanced mobility of capital. The ownership of corporations is more internationally diversified than ever before, and an increasing number
Chapter 21: Issues in International Taxation

of firms are multinational. These changes have increased the number of issues that have to be taken into account in the design of the tax system. In a globalized world no country is isolated, and tax changes in one country can cause a reallocation of financial and productive capital that affects all countries.

21.3.2 Corporate Taxation

The taxation of corporate income involves the transformation of profit into tax revenue. Globalization has made corporations increasingly mobile across jurisdictions. If the tax system attempts to tax profit where it is generated, then a multinational firm will have an incentives to change the location of profit to where it is the most advantageous from a tax perspective. This can be achieved either through changing the physical location of the firm (or parts of the firm) to earn profit in a different location or by restructuring financial flows to make it appear that profit is earned in a different location.

The increasing exploitation of low-tax jurisdictions by corporations is illustrated in figure 21.1. The figure reports the profits of the subsidiaries of US corporations located in four low-tax jurisdictions (often called tax havens) and the effective tax rates in those jurisdictions. The profits reported in each of the jurisdictions rose substantially over the four-year period considered. The one exception is the significant fall in the year that the effective tax rate in Switzerland increased. The sharp fall in profit resulting from this tax increase further illustrates the international mobility of these profits.

In an international setting the first step in allocating the tax base between jurisdictions is to determine where profit is generated. If this can be done, the next step is to decide whether profit should be taxed on the basis of where it is generated, or on some other basis such as where it accrues to owners. Both choices are possible when taxes are being designed. Tax design must also take into account the ability of the headquarters of a firm to move from one country to another. Production and office facilities can also move. The important feature of this mobility is that it often involves a discrete choice: the headquarters may be in the United States or in Germany, the plant in Vietnam or Thailand (but not a bit in both). Usually in tax analysis it is the effect of the marginal rate of tax on marginal decisions that is important. The situation is different when a discrete choice has to be made. In such a case it is the average rate of tax that determines the chosen outcome.

These points can be related to the decisions that a company must make. The choice of where to produce is made by calculating the level of profit net of tax liability in each of the alternative locations. The location delivering the highest profit net of tax will be chosen. It is for this choice that the average tax rate is relevant. Conditional on the
Figure 21.1
Profits and effective corporate tax rates
Choice of location the company must then select the level of output. This is a standard marginal decision for which the marginal tax rate is relevant. The final decision for the firm is to choose the location of profit. This is not the same as the real decision about where to produce, but refers to the nominal allocation of profit through accounting choices. The following discussion of corporate taxation will treat each of these issues. They will be discussed separately to isolate the key points, but it is clear that they are interrelated. For example, a firm will choose to locate where the average tax rate is low, and will produce more at the location if the marginal rate is low. It will also choose to locate profit in that location to take advantage of the low tax rate.

Consider a corporation that operates divisions in two or more tax jurisdictions. Four of the alternative tax bases are now introduced and some of the issues relating to the choice of base are discussed. The four tax base alternatives are:

- **Source-based taxation** Corporate income earned in the country where productive activity takes place.
- **Residence-based taxation (corporate)** Income accruing to the residence country of the corporate headquarters.
- **Residence-based taxation (personal)** Income accruing to the residence country of the corporate headquarters or personal shareholders (residence-based taxation)
- **Destination-based taxation** The sales (net of costs) in the destination country where the goods or services are finally consumed.

The first point that needs to be made concerning the choice between these tax bases is that it is difficult to define the source of profit. Difficulties can arise even when a corporation has a simple structure but undoubtedly becomes more pronounced as corporate structure increases in complexity. A corporation may have many different functional units, such as management, production, marketing, finance, sales, and research and development (R&D), and each of these functions can be divided between different divisions of the firm in different jurisdictions. The problem confronting source-based taxation of the corporation is the allocation of profit across these functions and across the divisions. Each one of the functions is necessary for the operation of the firm and so ultimately has a claim to be a source of profit. Neither the internal accounting of the firm nor the information observed by tax authorities will generally be adequate to allocate profit to sources.

One resolution to this difficulty is to define prices at which the products of one of the divisions are sold to another. If such prices exist then the profit can be distributed across sources. Some products may be traded within the firm, and some may be sold
outside as well as used inside. In these cases observable prices exist. For other products there may be no observable prices. For example, the supply of management services may not be priced within the firm. However, the distinction between these cases may not be that significant. What the tax authorities require are arm’s-length prices that represent the price that the product or service would be sold at between unrelated parties. Since the prices used within the firm may not represent arm’s-length prices (we explore why when we look at transfer pricing later), the observation of a price within the firm may not actually be much more helpful than having no price. A more fundamental problem is the possibility that no price can be defined for the transaction. To see this point, assume that the discoveries of two distinct R&D divisions are critical for the production process of the corporation, but neither discovery is of any value alone. The arm’s-length price is zero individually or positive when the discoveries are combined. However, pricing the combined discoveries does not assist with the identification of the source of profit.

These comments demonstrate the difficulties, both practical and conceptual, involved in allocating profit on a source basis. These problems would have to be faced by any implementation of a source-based tax. In practice, the problems are dealt with through rules that attempt to approximate what is economically correct and through occasional litigation.

At first sight, it might seem easier to identify a residence country than to identify a source country. Close inspection shows that this is not necessarily the case. Defining the jurisdiction of residence of the headquarters of a multinational can be problematic. In practice, it is defined through the concept of the location of control and management. The ruling on this point in United Kingdom law summarizes the position as “the question where control and management abide must be treated as one of fact or ‘actuality’.” In brief, this means there is no absolute principle but a judgment on the facts in each case.

Profits are usually taxed when they are repatriated to the parent company. Taxing on repatriation provides an incentive not to repatriate, which might mean that excessive investments are undertaken abroad. So an alternative is to tax profits when they are accrued. The benefit of a residence-based tax is that all that must be observed is the total profit of the company. It does not matter where it is earned. If all countries were to operate on the residence basis, then companies would not benefit from shifting profit around and using tactics such as transfer pricing. However, there are two problems with this. First, the holding company of a multinational is mobile, and it may well move to take advantage of tax benefits. Consider the case of a holding company located in one country but with all productive activity and sales taking place elsewhere, and the shareholders resident elsewhere. There is a serious question about the moral legitimacy of residence-based tax in this case, and possibly also considerable mobility of the
holding company since it has no real connections to location. Second, a multinational may have many divisions across numerous countries. This is not a problem in theory, but in practice it may be very difficult for the residence country to monitor the taxable activity across the range of divisions. Even though it is only the sum of profit that needs to be known, this can only be monitored by using information on the individual components going into that sum.

An alternative is to tax the shareholders of a firm on a residence basis when they receive dividend payments. There are two sources of practical problems with this choice of tax base. The first problem confronting a tax authority is the difficulty of observing the dividend income that arises from foreign shareholdings. There may be regulations requiring domestic corporations to report payments to the domestic tax authority, but these need not extend internationally. Observation of dividend income is made particularly problematic by the fact that countries have an incentive to offer lax regulations in order to attract corporate headquarters. Such tax havens do exist, and can play host to an exceptionally high numbers of firms (Liechtenstein has a population of 35,000 people but approximately 73,000 registered corporations). The second problem is that the country of residence of individuals is not always clear. People can be internationally mobile with their country of residence ill-defined and open to dispute.

In summary, both source- and residence-based systems of corporate taxation have to confront a range of practical and conceptual difficulties. These difficulties prevent the implementation of either system in a manner that conforms with the theoretical idea. This has led to proposals for the implementation of systems that approximate the ideal but can be applied in practice. One such proposal is now considered.

21.3.3 Formula Apportionment

A practical solution to the problems of implementing the source- or residence-based systems is to employ formula apportionment to allocate the tax base across jurisdictions. Formula apportionment aggregates the activities of a multinational regardless of location and then allocates the tax base across countries according to a pre-agreed formula. The formula can be chosen to approximate either system of taxation. The European Commission has proposed the use of formula apportionment using a common coordinated EU tax base. The apportionment proposed by the European Commissions is based on the proportion of total sales that take place in each country. If the profit margin was constant and identical across countries, then this would be a source-based system. Otherwise, it will provide an approximation to a source-based system. The consequences of using this version of formula apportionment are now investigated with
the main focus placed on whether it is consistent with production efficiency and the
effect it has on the choices of a corporation.

To explore the implications of this system, consider a monopolist producing in two
countries using two internationally mobile inputs. The output produced by the firm can
also be transported costlessly between countries. There is a supranational government
(representing the European Union) that implements formula apportionment.

With the proposed form of formula apportionment a proportion $q$ of profit is taxed
in country $A$ and $1 - q$ in country $B$. With tax rates $t^A$ and $t^B$ the level of profit is

$$
\pi = \left[ x^A p(x^A) + x^B p(x^B) - r \left( K^A + K^B \right) \right. \\
\left. - w \left( L^A + L^B \right) \right] \left[ 1 - qt^A - [1 - q]t^B \right].
$$

(21.20)

This is maximized subject to the constraint that

$$
x^A + x^B = F(K^A, L^A) + F(K^B, L^B)
$$

(21.21)

and the formula apportionment rule

$$
q = \frac{x^A p(x^A)}{x^A p(x^A) + x^B p(x^B)}.
$$

(21.22)

The necessary conditions for the capital levels in the two countries are

$$
-r \left[ 1 - qt^A - [1 - q]t^B \right] + \lambda F_{K^j} = 0, \quad j = A, B,
$$

(21.23)

where $\lambda$ is the Lagrange multiplier on the production constraint. The necessary
conditions imply that

$$
F_{K^A} = F_{K^B},
$$

(21.24)

which is the efficiency condition for the mobile factor. The necessary conditions for $L^A$
and $L^B$ provide the same conclusion. This shows that formula apportionment based on
sales does not affect the international efficiency of factor allocation. This is a point in
its favor.

It should not be concluded that formula apportionment is entirely distortion-free.
Formula apportionment provides a motive to move profit to the low-tax country. This
is achieved by directing more output to that country. If the demand function is the same
in both countries, then this implies the price must be lower in the low-tax country to
raise the proportion of sales in that country. Hence a system of formula apportionment
operates as a form of implicit tax in the high-tax country.
21.3.4 Summary

The process of globalization has made capital and corporations more mobile. This makes it harder to identify the source of profit and the place of residence of the recipients of profit. In principle, there are clearly defined systems for taxing profit, or the return to capital generally, on a consistent basis. These systems differ in the effects that they have and in their neutrality properties. The drawback is that all systems are faced with practical difficulties that hinder implementation. Formula apportionment is a possible solution, and this can have desirable efficiency properties.

21.4 Transfer Pricing

Consider a multinational firm with divisions in two countries. Assume that one of the firm’s divisions produces a good while the other uses it, either as an input or for final sale. The price at which the good is transferred between the two divisions of the firm is known as the transfer price. For the division of the multinational firm that produces the good, the transfer price determines revenue. Conversely, for the division that uses the good the transfer price determines costs. The transfer prices on the transactions between the divisions of the firm should guide resource allocation within the firm. From this perspective, the firm has an incentive to choose transfer prices that mirror the true costs of production.

The choice of transfer price also has another implication for the firm: it affects the international allocation of profit between the countries in which the different units of the firm are located. A high transfer price will raise the profit of the unit producing the good, whereas a low transfer price will raise the profit of the user. If the countries in which the units of the firm are located have different corporate tax rates, the firm can set the transfer price to ensure the largest possible profit is made in the low-tax country. This will increase net-of-tax profit by exploiting the tax differential. It also affects the tax revenues of the countries and provides an incentive for countries to reduce corporate tax rates. The OECD Transfer Pricing Guidelines describe this in the following way: “Transfer prices are significant for both taxpayers and tax administrations because they determine in large part the income and expenses, and therefore taxable profits, of associated enterprises in different tax jurisdictions.”

To prevent firms from using transfer pricing to reduce tax liability many governments have adopted rules on transfer pricing. The central feature of the rules is the concept of an arm’s-length price. This is defined as the price that would be charged if the
intermediate good was sold to an independent buyer. This is easy to apply if there is a market for the intermediate good and the transfer price can be compared to the price actually charged. If there is no market, which will be the case for many specialized intermediate inputs, the rules specify a process for determining what constitutes the correct arm’s-length prices. The rules typically allow the transfer price to be freely set, but permit the tax authorities to adjust that price when it falls outside what is judged to be a reasonable range for the arm’s-length price.

There are practical difficulties in operating the arm’s-length standard. It has already been noted that the good in question may not be openly marketed. It may also be the case that there is no sufficiently similar item on open sale to permit a price comparison, or that the terms and conditions of sale may vary among transactions. Transfer-pricing rules usually allow the use of several methods for testing whether the transfer price is appropriate. The common methods include the use of comparable uncontrolled prices, cost-plus pricing, resale price or markup, and profitability-based methods.

We begin by exploring the incentives the firm faces when choosing a transfer price in the absence of restrictions on behavior. Suppose that the firm produces in country A. Some of the output is sold in country A, whereas the remainder is supplied to a division in B for sale in that market where there is no other production. The price in country \( i, i = A, B, \) is denoted by \( p_i \). The cost of production in country \( A \) is \( C = C (X_A + X_B) \), where \( C' > 0 \). Country \( i, i = A, B, \) levies a profit tax at rate \( t_i \).

With a transfer price \( \phi \) the aggregate profit of the firm is

\[
\pi = \left[ 1 - t_A \right] \left[ p_A X_A + \phi X_B - C(X_A + X_B) \right] + \left[ 1 - t_B \right] \left[ p_B X_B - \phi X_B \right]. \tag{21.25}
\]

This can be written as

\[
\pi = \left[ 1 - t_A \right] \left[ p_A X_A - C(X_A + X_B) \right] + \left[ 1 - t_B \right] \left[ t_B X_B + \phi \left[ t_B - t_A \right] X_B \right]. \tag{21.26}
\]

The choice of transfer price is clear from (21.26). If

\[
t_B - t_A < 0, \tag{21.27}
\]

then the firm will want to set the lowest possible transfer price to shift profit to the low-tax country \( B \). In this case the incentive is to maximize the profit accruing in country \( B \). Conversely, if

\[
t_B - t_A > 0, \tag{21.28}
\]

then the firm will want to set the highest possible transfer price to shift profit to the low-tax country \( A \). These conditions make it clear that the incentives for transfer pricing arises from international tax differences.
The analysis can be extended to explore the effect of arm’s-length regulation of the transfer price. It has been argued that it may be hard to exactly determine the correct arm’s-length price. This suggests that there may be a range of acceptable prices that can be used. Such a range can be represented by placing upper and lower bounds on the transfer price. The two natural bounds to place on the transfer price are that it cannot be below marginal cost and that it cannot be above the market price in country \( A \). Hence \( \phi \) must satisfy
\[
C' \leq \phi \leq p_A. \tag{21.29}
\]
Consider the choice of a monopoly firm when (21.27) is satisfied so county \( B \) has the lower tax rate. In this case \( \phi = C' \), so the profit level for the firm is
\[
\pi = [1 - t_A]\{p_A(X_A)X_A - C(X_A + X_B)\}
+ [1 - t_B]\{p_B(X_B)X_B + C'[t_B - t_A]X_B\}. \tag{21.30}
\]
The necessary conditions for the choice of \( X_A \) and \( X_B \) are respectively
\[
[1 - t_A]\{p_A + p_A'X_A - C'\} + C''[t_B - t_A]X_B = 0, \tag{21.31}
\]
\[
[1 - t_B]\{p_B + p_B'X_B - C'\} + C''[t_B - t_A]X_B = 0. \tag{21.32}
\]
The effect of the bound on the transfer price can be seen to be dependent on the sign of \( C' \). If \( C'' = 0 \), the lower bound is unaffected by changes in output. In this case the second terms vanish from the two necessary conditions. The supply to country \( A \) equates marginal revenue to marginal cost. The same condition holds true in country \( B \), so supply also equates marginal revenue and marginal cost. When \( C'' > 0 \), the cost-increase effect has to be taken into account. An increase in output raises marginal cost and tightens the bound on the transfer price. This implies that the supply to country \( A \) is reduced below the level that equates marginal revenue and marginal cost. The same distortion arises in country \( B \). Conversely, when \( C'' < 0 \), an increase in output reduces the lower bound and encourages greater supply to both countries.
Now consider the choices of the firm when \( t_B - t_A > 0 \) so that is at the upper bound of the transfer price rule. The level of profit is
\[
\pi = [1 - t_A]\{p_A(X_A)X_A - C(X_A + X_B)\}
+ [1 - t_B]\{p_B(X_B)X_B + p_A(X_A)[t_B - t_A]X_B\}, \tag{21.33}
\]
with necessary conditions for the choice of output levels
\[
[1 - t_A]\{p_A + p_A'X_A - C'\} + p_A'[t_B - t_A]X_B = 0, \tag{21.34}
\]
\[
[1 - t_B]\{p_B + p_B'X_B - C'\} + [t_B - t_A][p_A - C'] = 0. \tag{21.35}
\]
The effect of the bound here is to reduce supply to country $A$ because this increases the upper bound on the transfer price. Since (21.34) implies that $p_A C' > 0$, the supply to country $B$ beyond the level that equates marginal revenue to marginal cost despite country $B$ levying the higher tax.

In effect, the consequence of a transfer price rule with upper and lower boundaries is that it gives an incentive for the firm to change its supply to the two markets in an attempt to manipulate the rule. Hence introducing a rule causes additional distortions. Of course, one of the countries will gain from the rule by ensuring that its corporate tax revenues are increased. The other must lose.

Firms can use transfer pricing to obtain a favorable tax treatment of profit. If there is no limitation on the transfer price then the firm will set an extreme value in order to locate as much profit as possible in the low-tax jurisdiction. This process undermines the ability of jurisdictions to levy corporate taxes and encourages tax competition. To prevent this from happening transfer price rules are needed. The rules typically make use of the concept of an arm’s-length price. Placing upper and lower limits on the transfer price can cause a distortion if firm sees these limits as endogenous.

21.5 Location

It has already been mentioned in chapter 20 that the European Union operates an agreement to prevent harmful tax competition. One aspect of harmful tax competition is that some EU member states may try to attract firms from other member states. Even with the agreement the differences in tax treatment of corporations across member states still give rise to relocations of firm, with the media focusing on the role of taxation in these moves. This raises the question of the sensitivity of location choice to tax rates. The previous section considered the location decisions as given. This section goes one step further and looks at the issue of how taxation affects the choice of location.

For most economic choices at intensive margins, such as the decision to work a little bit more or less, it is the marginal tax rate that matters. The location decision of a firm is instead a choice at the extensive margin: it will either locate in one country or another. For such discrete choices it is the average tax rate that is important. This matters from the policy design perspective. It also matters from an empirical perspective since, to analyze location, it is necessary to construct effective average tax rates. This is a point to which we return after some theory.
21.5.1 Locational Choice

Firms are able to extract rents from competing jurisdictions. The extent of the rent depends on the benefits to the jurisdiction and location-specific benefits for the firm. Controlling competition between jurisdictions will reduce the rent extracted.

Consider two countries, $A$, $B$, that will derive benefits $Y_A$ and $Y_B$ respectively if the firm decides to locate within their borders. The firm has a preference for country $A$ since it earns an additional rent if it locates there. This location-specific rent could arise because of a skilled workforce in $A$ or local availability of necessary inputs. The location-specific rent for the firm in $A$ is $\theta > 0$. The two countries compete by choosing the taxes that they will charge the firm.

Denote the tax levied by country $i$ by $T_i^i$ (so it is a subsidy if $T_i^i < 0$). The firm obtains $\pi + \theta - T_A$ in country $A$ and $\pi - T_B$ in country $B$. The firm will locate in country $A$ provided that

$$T_A \leq \theta + T_B. \tag{21.36}$$

The maximum subsidy in $B$ must leave the country with a nonnegative payoff so that

$$Y_B + T_B \geq 0. \tag{21.37}$$

Given that country $A$ wishes to minimize its subsidy (or maximize its tax), it will choose the level of tax $T_A^A = \theta - Y_B$. The tax can be positive or negative depending on the values of $\theta$ and $Y_B$.

With this value of $T_A^A$ country $A$ receives

$$Y_A + \theta - Y_B, \tag{21.38}$$

and the firm receives the payoff

$$\theta - T_A = Y_B. \tag{21.39}$$

If $T_A^A > 0$, then the location-specific rent is shared between the firm and country $A$ (with $\theta > Y_B$). If $T_A^A < 0$, the firm extracts a payoff in excess of the location-specific rent (with $\theta < Y_B$).

This analysis illustrates how mobile firms are able to extract rents when countries compete to host them. The countries will compete to provide tax concessions, and the firm will choose the location that provides the highest payoff after taking these concessions into account. The other message of the analysis is that when location is a discrete choice (either $A$ or $B$), it is the total tax payment or, equally, the average
Part VII: Multiple Jurisdictions

tax rate that matters for choice. This is in contrast to the usual focus on the marginal
tax rate. This observation is important for empirical analysis of location choice since
the average tax rate is often very difficult to determine—it is often the consequence of
a complex set of taxes and allowances—and has to be constructed. This is discussed
further shortly.

21.5.2 Agglomeration Rents

In addition to taxation, location choices are determined by other agglomeration and
dispersion forces. The New Economic Geography, developed in the early nineties by
Krugman and others, states that even when high corporate tax rates deter firms from
locating in a country, there can still be agglomeration rents that attract them.

The agglomeration force is the access to the market for inputs and outputs. The
closer one is to the center of activity, then (1) the better the access to the market to sell
output, and (2) the better the access to inputs and intermediate goods. Because of these
factors, firms want to locate close to their workers and workers want to locate close to
firms. These agglomeration forces can result in *external economies of scale*: as more
firms locate in a given region and aggregate industry output rises the average cost of
production falls. External economies can be due to knowledge spillovers and access
to intermediate goods and trained workers. This raises the incentive for more firms to
locate in the same region. The location of computer chip manufacturers in California’s
Silicon Valley is an example of such agglomeration. There is no reason why the first
chip manufacturer located there other than an accident of history, but once the process
began, it gathered its own momentum.

The attraction force of agglomeration is offset by the dispersion force that arises
from the competition effect. The closer a firm is to center of activity, the tougher is the
competition from other firms located in the same area. The proximity of location will
result in products being close substitutes, which enhances competition.

How the two forces are balanced depends on the level of transportation costs (or,
equivalently, of trading costs). A key result of New Economic Geography is that in-
creasing returns to scale (at the firm level) and imperfect competition combined with
transportation costs may result in agglomeration. From this perspective, corporate taxes
deter firms from a location while agglomeration rents attract firms. So it is possible to
tax the agglomeration rent to some extent without driving firms away. Tax differences
across locations are explained by a countervailing gap in agglomeration rents. For both
low and high transportation costs, the agglomeration rents are low. For intermediate
transportation costs, the agglomeration rents are higher, which implies that taxes can
be set at a higher level in the center compared to the periphery of the economic activity. In this approach transportation is a proxy for trade costs, so the agglomeration rent is a bell-shaped function of the trade costs.

21.5.3 Evidence

The discussion of the international tax treatment of capital assumed that firms will relocate profit or capital to take advantage of international tax differentials. The practical significance of the effects we have identified depends on the extent to which firms do respond to taxation. There is a large literature that has analyzed the data to measure this responsiveness.

The movement of profit between locations is much easier than the physical movement of firms. It is therefore not surprising that there is ample evidence that international profit-shifting does indeed take place. This is despite the efforts of governments to contain it via transfer-pricing regulations. Evidence of profit-shifting has been found between the United States and various other countries and, more generally, within the OECD area. The typical evidence is that reported income within a country falls in response to a unilateral tax increase by that country.

The second issue is the response of the international location of real investment to differences in tax rates. This is a more difficult question to address. The chosen location is observed, but the rejected alternatives are not, so the factors that entered the decision process of the firm cannot be observed directly. The choice of location depends on the average tax rate that a firm faces (controlling for agglomeration rents). This tax rate cannot simply be read from legislation for several reasons. First, the tax system may involve several marginal rates of tax that depend on the profit level. Second, there may be reductions in the marginal rate in special circumstances such as location in an enterprise zone or as a start-up incentive. Third, there can be allowances for R&D expenditures and special depreciation provisions. These factors make the average rate of tax potentially specific to the circumstances of each firm and have required that it be computed as part of an empirical analysis.

The analysis in section 21.5.1 considered a firm making a single-period location decision. In practice, a firm must assess how the choice of location affects its value as determined by the future flows of dividends to the owners. This flow of dividends is determined by both the investment policy and the financial policy of the firm over the future lifetime of the firm. The analysis of the lifetime decision problem is complex, but for the purpose of analyzing how taxation affects location choice, it can be considerably simplified. The method for doing this is to use a perturbation argument:
investment is increased by one unit at time $t$ and reduced by $1 - \delta$ units at time $t + 1$, where $\delta$ is the depreciation rate of capital. This perturbation leaves the capital stock unchanged from period $t + 1$ onward, so the flow of profit is only changed (or perturbed) in periods $t$ (where extra investment is financed) and $t + 1$ (where extra output is produced and investment reduced). The total effect of this perturbation can be summarized by calculating the net present value (NPV) of the additional costs and revenues.

For each potential location a $NPV$ can be defined before tax ($NPV_B$) and after tax ($NPV_A$). The choice of location is determined by the comparison of the after-tax $NPV$s. This comparison can be related to taxation by using the $NPV_A$ and $NPV_B$ to define an effective average tax rate ($EATR$). The standard way to do this is to determine the net present value of the increase in revenues from the perturbation, $NPV^R$. The perturbation increases investment by one unit at $t$ so raises capital by one unit at $t + 1$. Assume that the extra unit of capital produces additional output that earns revenue $p$ at time $t + 1$. The value of this additional revenue at $t$ is $NPV^R = \frac{p}{1 + r}$, where $r$ is the interest rate. The $EATR$ is then defined by

$$EATR = \frac{NPV_B - NPV_A}{NPV^R}. \quad (21.40)$$

The reason for measuring the $EATR$ relative to $NPV^R$ is that this will almost certainly be positive (because it is an increase in revenue without subtracting costs), whereas either of $NPV_B$ or $NPV_A$ could be zero (or negative) because they include the additional cost of investment. Computing the $EATR$ for each location provides a comparison of the level of taxes.

One line of research has considered the effects of US tax policy. It has been shown that companies that can claim tax credits against their home-country tax bill for state income taxes paid in the United States are less likely to avoid high-tax states. This is consistent with the theoretical prediction. In addition the evidence demonstrates that US firms shifted away from international joint ventures in response to the higher tax costs created by certain provisions of the US Tax Reform Act of 1986.

An alternative line of research estimated the sensitivity of foreign direct investment (FDI) to changes in tax regimes. Many studies have been undertaken that suggest an elasticity somewhere in the range $-0.6$ and $-1.5$, with a value that varies across countries. Some studies have reported even higher figures.

The theoretical analysis shows how the mobility of firms can enable rent extraction. The empirical studies provide evidence that taxation does have a significant effect on the location of profit and of activity. The design of the international tax structure is
therefore important. This is why there is so much concern about competition for the location of firms that will distort location choice.

21.6 Harmonization of Taxes

In many areas of policy the European Union finds the idea of approximation—the closer alignment of the policies of member states—to be appealing. When applied to international tax policy, the process of approximation implies a reduction in the dispersion of tax rates across member states. The European Union announced an acceleration of the process of approximation of VAT rates in 1985. At the outset of this process, the aim was that by 1993 the tax rate on any commodity in any member state would not deviate more than 2.5 percentage points from a community norm. The process did not reach its intended outcome but did make some progress. The process of approximation remains an EU policy objective.

In the economic literature the EU terminology of approximation is replaced by that of harmonization, which means moving toward, though not necessarily achieving, a common value for tax rates. Complete harmonization would replace a range of different tax rates across countries with a single tax rate. This could be, for example, the average of the initial tax rates. A partial harmonization would move the individual tax rates a little closer to a common target.

It is not immediately obvious that harmonization can raise economic welfare. Applying the analysis of commodity taxation in chapter 15 provides a reason to believe that harmonization may well reduce welfare. Recall that the inverse elasticity rule demonstrates that tax rates should be differentiated across goods to reflect the demand conditions for each. Extending this argument to countries suggests that VAT rates should be based on local demand elasticities. If these elasticities vary across countries, then implementing a process of harmonization will move the system away from the efficient outcome. There are two reasons why this anti-harmonization argument may not apply. First, the existing tax rates need not be efficient. Second, there may be inefficiencies that are not resolved (or may even be enhanced) by the imposition of differentiated taxes. These arguments are now briefly explored.

It is always possible that an existing set of tax rates was selected for reasons other than efficiency. The tax rates might have been the outcome of lobbying or of historical precedent. Whatever the case, if the tax rates are not efficient, then harmonization may be able to raise welfare. Harmonization may even be welfare-improving if the tax rates were chosen efficiently. This is because there is a distinction between what is efficient for each country and what is efficient for a set of countries. If each country chooses
its tax rate to maximize national welfare, then it will not take into account spillover effects on other countries. Harmonization could move the tax rates closer to those that maximize aggregate welfare. We explore this argument in more detail below.

One form of inefficiency that has been the basis of a practical argument in favor of harmonization of VAT in the European Union is cross-border shopping. Cross-border shopping is inefficient because individual consumers are driven by tax differentials to personally transport commodities. The transport of commodities across borders by firms can exploit economies of scale; cross-border shopping by individuals does not. In addition individual transportation increases congestion and environmental damage. These costs are avoided if VAT rates are harmonized, thus removing the incentive to engage in cross-border shopping. More generally, harmonization lessens the distortion of trade patterns and enhances the efficiency of trade.

These competing perspectives show that the question of whether harmonization can be beneficial is a significant one that does not have an immediately obvious answer. Insight into the question can be obtained by considering two countries that impose different rates of VAT. This permits a harmonization of the rates to be imposed and the welfare consequences determined. The question is whether such a harmonization can raise the welfare levels of both countries.

A case for harmonization can be constructed when countries choose tax rates independently and ignore the effects on other countries. To demonstrate this argument, suppose that for two countries, A and B, and two goods, 1 and 2, good 1 is exported from country A to B while good 2 travels in the opposite direction. (The direction of trade should really be derived as the equilibrium outcome, but it is easier to impose it directly.) The price of good 1 is normalized at 1 and the price of good 2, before tax is imposed, is $p$. The tax on good 2 in country $j$ is denoted $t_j$. Neither country taxes good 1.

The equilibrium condition for the economy is that trade is balanced. Demand for good $i$ in country $j$ is a function of the price of good 2 in that country. Hence $x_i^j = x_i^j (p + t_j)$, $i = 1, 2$. The trade balance condition is

$$p\left[y_A^2 (p + t_A) - y_A^1\right] = x_B^1 (p + t_B) - y_B^1,$$

(21.41)

where $y_j^i$ is the supply of good $i$ in country $j$. This equilibrium condition determines the price of good 2 as a function of the tax rates

$$p = p(t_A, t_B).$$

(21.42)

For country B the price $p$ represents the terms of trade (the price of exports of good 2 divided by the price of imports of good 1). For country A it is the inverse of the terms of trade.
The source of the inefficiency can now be explained. The dependence of the equilibrium price on the tax rates means that domestic tax policy in one country affects the terms of trade for the other country. This is a form of externality between the countries. When the tax rates are chosen independently, this externality will not be internalized and the resulting equilibrium will be inefficient. Notice that this is a consequence of assuming that the countries are “large” in the sense that their actions have a significant effect on the world price. In contrast, if the countries are assumed to be “small,” then the terms of trade would be perceived as fixed and the externality would not exist.

When the countries act independently country $A$ chooses the tax rate to solve

$$
\max_{t_A} W_A(t_A, p(t_A, t_B)),
$$

which has necessary condition

$$
\frac{\partial W_A}{\partial t_A} + \frac{\partial W_A}{\partial p} \frac{\partial p}{\partial t_A} = 0.
$$

(21.44)

There is an equivalent condition for the choice of country $B$. Denote the taxes resulting from this optimization by $\{t^I_A, t^I_B\}$. The efficient taxes should solve the joint welfare maximization problem

$$
\max_{\{t_A, t_B\}} W = W_A(t_A, p(t_A, t_B)) + W_B(t_B, p(t_A, t_B)),
$$

(21.45)

with necessary condition

$$
\frac{\partial W_A}{\partial t_A} + \frac{\partial W_A}{\partial p} \frac{\partial p}{\partial t_A} + \frac{\partial W_B}{\partial p} \frac{\partial p}{\partial t_B} = 0.
$$

(21.46)

Denote the taxes resulting from this optimization by $\{t^J_A, t^J_B\}$. The third term in (21.46) is the terms of trade externality that is not taken into account in the independent optimization. It is this that results in the difference between the taxes $\{t^I_A, t^I_B\}$ and $\{t^J_A, t^J_B\}$.

It is now possible to explain the circumstances where the countries can gain from harmonization. The jointly determined taxes always lead to higher world welfare

$$
W_A(t^J_A, p(t^J_A, t^J_B)) + W_B(t^J_B, p(t^J_A, t^J_B)) > W_A(t^I_A, p(t^I_A, t^I_B))
$$

$$
+ W_B(t^I_B, p(t^I_A, t^I_B)).
$$

(21.47)

The comparison of welfare levels across the two countries is not as clear-cut. At least one country will gain by moving from the independent tax rates to the jointly determined
tax rates, but it is possible that one may lose. Consider first the case in which both gain so that

\[ W_A(t_A^I, p(t_A^I, t_B^I)) > W_A(t_A^I, p(t_A^I, t_B^I)), \] (21.48)

and

\[ W_B(t_B^I, p(t_A^I, t_B^I)) > W_B(t_B^I, p(t_A^I, t_B^I)), \] (21.49)

and assume that \( t_A^I \neq t_B^I \). Next define the efficient harmonized tax rate, \( t_U \), that solves

\[ \max_{t_U} \{ W = W_A(t, p(t, t)) + W_B(t, p(t, t)), \] (21.50)

with the resulting welfare levels \( W_j(t^U, p(t^U, t^U)), j = A, B \). The case for harmonization can now be made. If

\[ W_A(t^U, p(t^U, t^U)) > W_A(t_A^I, p(t_A^I, t_B^I)), \] (21.51)

and

\[ W_B(t^U, p(t^U, t^U)) > W_B(t_B^I, p(t_A^I, t_B^I)), \] (21.52)

then a move from the independent taxes to the harmonized tax will raise the welfare level in both countries. This happens because the countries move from a jointly inefficient tax system to one that is preferable. The harmonized system need not be jointly optimal, but it is better. If this is the case, a partial harmonization will also improve welfare. A form of partial harmonization is for each country to move to a weighted average of the independent tax rates

\[ \tilde{t}_A = [1 - \mu] t_A^I + \mu t_B^I, \] (21.53)

\[ \tilde{t}_B = [1 - \mu] t_B^I + \mu t_A^I. \] (21.54)

This will always ensure that the new tax rates are more similar than the old tax rates, \( |t_A^I - t_B^I| > |\tilde{t}_A - \tilde{t}_B| \), and will bring the countries closer to the position of harmonization. Unless the welfare functions have unusual properties, it will generally be the case that both countries gain by moving from the independent taxes to the partially harmonized taxes.

There are caveats to this argument. If one—or both—of the countries prefer independent taxation to the harmonized tax rates, then harmonization will not raise welfare. Note that unlike the joint taxation solution, harmonization can make both countries
worse off. We have also said little about the role of tax revenue. If tax revenue is used in a beneficial manner (e.g., to provide a public good), then it will matter how the partial harmonization affects revenue. If it leaves revenue unchanged, the argument holds as given. If it changes revenue, then the argument will need to be extended to take this into account.

An extension to the analysis makes it possible for partial harmonization to raise the welfare of both countries even when one country is worse off at the efficient harmonized taxes (one of (21.51) or (21.52) does not hold) or even at the jointly efficient taxes (one of (21.48) or (21.49) does not hold). If there is a country that gains from partial harmonization, the extension is to permit it to make a lump-sum transfer to the country that loses. Partial harmonization will then raise welfare if one country gains more than the other loses, so that a compensating transfer can be found that leaves both countries better off.

This analysis has explored why partial harmonization can raise welfare. The argument is supported if there are other inefficiencies—such as cross-border shopping—that are reduced by partial harmonization. It is also supported if there are other externalities between the countries generated by the taxes. One example could be the locational decisions of firms as described in section 21.5.1. Despite the potential benefits of partial harmonization, it is important to recognize that it is very much a second-best policy if the jointly determined taxes are not uniform. In most cases there will be some way of combatting the externality that allows the countries to retain the right to set different tax rates.

### 21.7 Tax Principles

Completion of the single EU market has raised some debate as to how to operate the VAT system. Prior to 1993 the operation of the VAT system involved commodities leaving tax free from the exporting country and being taxed in the importing country. This was achieved by a zero-rating of exports, combined with subjecting all imports to VAT. In 1993 border controls within the European Union were abolished to facilitate free trade and a level playing field between firms operating within and across the EU member states. Cross-border trade undermined the operation of the VAT system, so a search began for a system that would be consistent with the principles of a single market. The EU committed to introducing a “definitive” system by 1997, but no such system is yet in place.

As a commodity crosses a border between the place of production and the place of consumption, the question of which country should levy tax is raised. Should it be
the country where the good is produced or the country where it is consumed? VAT is fundamentally a system of consumption taxation. By design, the (formal) incidence of VAT falls on the final consumer. At each stage in the production process the VAT paid on inputs can be reclaimed so that VAT is only paid on the value added. Only the final consumer is unable to reclaim VAT.

There are two distinct principles of taxation that correspond to the choice between levying taxes on consumption or production. Under the destination principle, goods are taxed in the country of final consumption. Under the origin principle, goods are taxed in the country of production. An important practical distinction between them is that the destination principle requires borders to monitor movement of goods, whereas the origin principle does not. The EU system with its zero-rating of exports involves an application of the destination principle. There have been numerous proposals that the European Union should move to the origin principle to ensure that the tax system is compatible with the absence of internal borders. These proposals are often based on the observation that the two systems are equivalent in the sense that a switch from one to the other will not lead to any reallocation of resources. This equivalence is demonstrated next.

In a closed economy with no exports or imports the consequence of the circular flow of income is that consumption and production must be equal. Expressed formally, over any given period of time,

\[ \text{Value of consumption} = \text{Value of production}. \]  

(21.55)

The equality between these two flows implies that it does not matter whether consumption or production is taxed. A tax at a fixed rate levied on final consumption will have exactly the same economic effects as the same tax levied on production. The interesting question is how this equivalence of a consumption tax and a production tax translates to an open economy engaged in trade. The key insight is that the same logic still applies but with one extra degree of complication.

To illustrate the argument better, suppose that in the two countries (A, B) with two goods (1, 2), good 1 is produced in A, and good 2 is produced in B. The goods are produced with constant returns to scale using only labor as an input. The tax rate in country \( i \) is \( t_i \). The wage rate in \( i \) is \( w_i \) with destination taxation and \( v_i \) with origin taxation. The prices with destination taxation are given in table 21.1a. The prices with origin taxation are given in table 21.1b.

The next step is to adopt the price normalization \( w_B = 1 \) and \( v_B = 1 \), and then express the prices of the two goods relative to the wage rates \( w_A \) and \( v_A \). The real prices of the two goods are displayed in tables 21.2a and 21.2b. It is these real prices that determine resource allocation.
### Table 21.1a
Prices and wages in destination regime

<table>
<thead>
<tr>
<th>Country</th>
<th>Good 1</th>
<th>Good 2</th>
<th>Wage</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$w_A [1 + t_A]$</td>
<td>$w_B [1 + t_A]$</td>
<td>$w_A$</td>
</tr>
<tr>
<td>B</td>
<td>$w_A [1 + t_B]$</td>
<td>$w_B [1 + t_B]$</td>
<td>$w_B$</td>
</tr>
</tbody>
</table>

### Table 21.1b
Prices and wages in origin regime

<table>
<thead>
<tr>
<th>Country</th>
<th>Good 1</th>
<th>Good 2</th>
<th>Wage</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$v_A [1 + t_A]$</td>
<td>$v_B [1 + t_B]$</td>
<td>$v_A$</td>
</tr>
<tr>
<td>B</td>
<td>$v_A [1 + t_A]$</td>
<td>$v_B [1 + t_B]$</td>
<td>$v_B$</td>
</tr>
</tbody>
</table>

Consider starting with the destination principle in operation. The real prices shown in table 21.2a must be the equilibrium prices that clear the markets for the two goods and labor in each country. Now switch to the origin principle. After the switch the wage rate in country A must adjust to the new structure of taxes to ensure that equilibrium is attained. The equivalence argument is completed by observing that the wage rate with the origin principle and the wage rate with the destination principle must be related by

$$v_A = \left[ \frac{1 + t_B}{1 + t_A} \right] w_A.$$  

(21.56)

This relation must hold because it is the only one that ensures that the real commodity prices are identical in tables 21.2a and 21.2b: the initial real prices attained equilibrium, so they must remain the same to attain equilibrium after the switch. This demonstrates that switching from one of these tax principles to the other does not change the equilibrium. The wage rate instead adjusts to compensate for the change in relative tax rates so as to leave the equilibrium unchanged.

This equivalence was described by the economist Tinbergen in a 1956 report for the European Steel and Coal Community (the forerunner of the European Union). The result shows that borders can be eliminated inside a single market and a system of destination taxation replaced by a system of origin taxation. Real resource allocation will not change as a consequence of the change in the tax system. The conditions under which the equivalence results hold are much more general than those in the example. All that is required for equivalence to hold is that taxation within each country is uniform (so
Table 21.2a
Prices and wages in destination regime

<table>
<thead>
<tr>
<th></th>
<th>Good 1</th>
<th>Good 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country A</td>
<td>$1 + t_A$</td>
<td>$\frac{1+t_A}{w_A}$</td>
</tr>
<tr>
<td>Country B</td>
<td>$w_A [1 + t_B]$</td>
<td>$1 + t_B$</td>
</tr>
</tbody>
</table>

Table 21.2b
Prices and wages in origin regime

<table>
<thead>
<tr>
<th></th>
<th>Good 1</th>
<th>Good 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country A</td>
<td>$1 + t_A$</td>
<td>$\frac{1+t_B}{v_A}$</td>
</tr>
<tr>
<td>Country B</td>
<td>$v_A [1 + t_A]$</td>
<td>$1 + t_B$</td>
</tr>
</tbody>
</table>

all commodities are taxed at the same rate) and that the wage rate (or the exchange rate in a monetary model) can adjust. The insight for policy is that origin taxation provides a viable alternative to destination taxation. It would also allow EU member states the freedom to set their own tax rates without the intensity of tax competition that occurs with the destination tax. The implementation of some form of origin system (though probably not the simple form used above) remains a “long-term goal” of the European Union. There is no (publicly known) movement toward that goal at the present time. It remains a problem that has to be resolved in the construction of a “definitive” tax system.

21.8 Tariff Policy

A tariff is a charge levied as a good crosses a border into a country. The advantage of tariffs over other forms of taxation is that they are generally easy to collect. Recall that taxes are most easily collected when they are levied on observable activities that are public information. A good crossing a border is observable, particularly if the good is carried on a ship that must enter a port. Hence tariffs can be used to raise revenue even if other parts of the tax administration are weak. Tariffs also offer protection to the home industry relative to the foreign industry, since they make the price of imported goods more expensive relative to home-produced goods. This makes tariffs attractive to countries wishing to encourage development. The drawbacks in using tariffs are that they are distortionary and may be met with retaliatory tariffs from trade partners.
21.8.1 Welfare Cost of Tariffs

The imposition of a tariff on a good raises the price paid by domestic consumers, raises the price received by domestic producers, and raises some revenue for the government. Since a tariff is a form of distortion, it must be the case that it creates a deadweight loss. The allocation of this deadweight loss is more interesting, and it may even be the case that a country can gain by the introduction of a tariff.

The effects of a tariff depend on whether the country is large or small. The meaning of large is that the policy actions of the country have an effect on the prices of traded goods in world markets. Conversely, small means that the country does not affect world prices.

We begin by analyzing the effect of a tariff levied by a small country. Suppose that there are two goods. The country imports one of these goods and exports the other. The country is small, so imports are sold at the fixed world price $p_M$ and exports at the fixed world price $p_X$. The imposition of a tariff $\tau$ implies that the domestic price of the imported good is $q_M = p_M + \tau$. This is the price that is paid by domestic consumers and the price that is earned by domestic firms on domestic sales (domestic firms do not pay the tariff).

The effect of the tariff is shown in figure 21.2. The curves $D_d$ and $S_d$ represent domestic demand and supply respectively. The imposition of the tariffs increases domestic supply from $s_0$ to $s_1$ but reduces domestic consumption from $d_0$ to $d_1$. Imports fall from $d_0 - s_0$ to $d_1 - s_1$. Area $a$ is the increase in producer surplus of domestic firms and area $b$ the revenue accruing to the government from the tariff. The fall in consumer surplus is the area $a + b + c$. Hence the tariff causes a deadweight loss equal to the area $c$. This is the standard result: the imposition of a tariff by a small country will reduce welfare.

There is a caveat to this result, however. The comparison of the areas in the figure is based on the assumption that the two forms of surplus and government revenue are valued equally. Recall from our introductory discussion that tariffs are a practical form of policy to implement in a country with limited tax administration capability. In such a country it is likely that a social valuation would weight government revenue more highly than consumer surplus. An example would be a country that needs the tariff revenue to finance the provision of an essential public good. This opens the possibility that the tariff can increase a welfare measure when the use of revenue is added to the analysis.

The welfare effect of a tariff in a large country is potentially very different due to the effect of the policy on world prices. When a large country introduces a tariff on
an imported good, it encourages domestic firms to increase output. This raises world output and reduces the world price of the imported good. In addition, as domestic firms produce more of the imported good, they must produce less of the exported good. The world price of the export increases as a consequence. The changes in world prices can be beneficial to the domestic country.

This discussion can be summarized using the concept of the terms of trade. The terms of trade measure the price of the exported good relative to the price of the imported good, \( \frac{p_X}{p_M} \). An improvement in the terms of trade (meaning an increase in \( \frac{p_X}{p_M} \)) is advantageous for the domestic country because it allows more units of the import for each unit of export. The effect of the tariff is to improve the terms of trade, and this can permit the domestic economy to gain from the imposition of a tariff.

The argument is illustrated in figure 21.3. The tariff causes the world price of the import to fall from \( p_M^0 \) to \( p_M^1 \), so the price on the domestic market rises only to \( q_M = p_M^1 + \tau < p_M^0 + \tau \). The consequence of this smaller rise in price is that areas \( a, b, \) and \( c \) are all reduced in size compared to the small country case. In the small country case, government revenue from the tariff was area \( b \). In the large country case, the tariff revenue is the sum of areas \( b \) and \( d \). The total loss of consumer surplus from the tariff is \( a + b + c \). Area \( a \) is gained by domestic firms as an increase in profit and \( b \) is gained by the government as revenue. Therefore, if area \( d \) is greater than area \( c \), then the tariff will increase welfare. This is clearly the case in the figure, but it does not
have to hold since it depends on the size of the tariff and the extent to which the world price falls.

There is one clear result that can be established in this framework: a large country will always gain by introducing a small tariff. That is, starting from a position with no tariff, the country can increase welfare if it implements a tariff that is just above zero. Formally, if welfare is written as a function of the tariff, \( W = W(\tau) \), then \( \frac{dW}{d\tau}|_{\tau=0} > 0 \). This is because the terms of trade effect dominates the loss of consumer surplus for a small tariff. It follows that there must be an optimal tariff, \( \tau^* > 0 \), that maximizes the welfare of the large country. This tariff is optimal conditional on other countries not changing their policies in response.

A tariff is always a distortion in the pricing system, and so must cause a deadweight loss. Now, if a large country gains from the tariff, where does the deadweight loss arise? The answer is that the trading partners of the large country must lose. They suffer a deterioration in the terms of trade as the price of the good they import rises relative to the price of their export. The resulting welfare loss due to the introduction of a tariff is always damaging for trading partners. Since the tariff must cause a reduction in world welfare, the trading partners always lose more than the large country gains.

Figure 21.3
Effect of a tariff on a large country
21.9 Trade Agreements

In an attempt to protect the US economy during the Great Depression the Smoot–Hawley Act of 1930 introduced heavy tariffs on imports. This was met by swift retaliation from Canada and reversed many years of trade liberalization. At the conclusion of the World War II, pressure mounted for a reduction in tariffs to secure the benefits of increased international trade. There have been numerous trade agreements since 1945 that have secured significant reductions in tariffs and ushered in the current era of globalization. There is little doubt that these trade agreements have benefited developed countries, but many would argue they have disadvantaged developing countries.

A fundamental premise of economics is that voluntary trade is welfare-enhancing: if two parties choose to enter a trade, then both must gain. In the absence of market failure, this argument leads to the conclusion that unregulated trade will attain a Pareto-efficient distribution of resources. If correct, this implies that any government intervention to prohibit trade, restrict trade, or levy taxes and tariffs creates a distortion and causes a loss of efficiency. Furthermore these arguments are as relevant for trade between two countries as they are to trade between two individuals and provide the basis for seeking free trade.

The basic explanation for trade among countries is comparative advantage that is based on the rate at which one good is given up to get another. It is distinct from absolute advantage that measures quantity of inputs used to produce output. Comparative advantage exists even in the absence of absolute advantage. Trade allows countries to specialize in the goods they produce relatively efficiently. Specializing and trading increases welfare. The benefits of trade are enhanced when countries differ in endowments and so can specialize in production of the goods that use their abundant factors intensively. Increased trade also expands market size, which permits gains to be obtained from external economies of scale. Increased competition from abroad forces domestic firms to be efficient. In the longer run the rate of economic growth can be increased through FDI (foreign direct investment) and the import of innovation. The summary of these arguments is that free trade is superior to autarky (the position of no trade) and superior to an intermediate regime of trade restrictions.

It needs to be said that the benefits of free trade rely on the assumption that market failure is absent. If there are preexisting market failures, then free trade would not lead to the benefits described. Involuntary unemployment due to market failure may be made worse by liberalization if labor market rigidities hinder redeployment of labor to the production of exports. Markets may be missing so that intermediate goods are not
Table 21.3
World trade talks

<table>
<thead>
<tr>
<th>Year</th>
<th>Place/name</th>
<th>Subjects</th>
<th>Number of countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>1947–48</td>
<td>Geneva</td>
<td>Tariffs</td>
<td>23</td>
</tr>
<tr>
<td>1949</td>
<td>Annecy</td>
<td>Tariffs</td>
<td>13</td>
</tr>
<tr>
<td>1950–51</td>
<td>Torquay</td>
<td>Tariffs</td>
<td>38</td>
</tr>
<tr>
<td>1956</td>
<td>Geneva</td>
<td>Tariffs</td>
<td>26</td>
</tr>
<tr>
<td>1960–62</td>
<td>Dillon Round</td>
<td>Tariffs</td>
<td>26</td>
</tr>
<tr>
<td>1963–67</td>
<td>Kennedy Round</td>
<td>Tariffs, antidumping</td>
<td>62</td>
</tr>
<tr>
<td>1973–79</td>
<td>Tokyo Round</td>
<td>Tariffs, nontariff barriers</td>
<td>102</td>
</tr>
<tr>
<td>1986–93</td>
<td>Uruguay Round</td>
<td>Tariffs, nontariff barriers, services,</td>
<td>123</td>
</tr>
<tr>
<td></td>
<td></td>
<td>intellectual property, textiles</td>
<td></td>
</tr>
<tr>
<td>2001–</td>
<td>Doha Round</td>
<td>Under negotiation</td>
<td>142+</td>
</tr>
</tbody>
</table>


available for the production of final goods (but, conversely, free trade may allow the necessary intermediates to be imported). Countries may not have adequate financial markets to permit risk to be hedged, so liberalization could increase volatility. Such market failures undermine the argument for free trade, but they do not prove that intervention is better: the benefits of intervention need to be demonstrated.

These comments provide background to what the trade agreements have been trying to achieve and also help explain the current deadlock in the latest round. The successive rounds of talks are now described and the achievements documented.

There have been nine rounds of world trade talks (see table 21.3). The first eight rounds were conducted under the General Agreement on Tariffs and Trade (GATT). The Doha Round are the current talks and began in 2001 but have not yet reached conclusion. GATT was a draft charter for the International Trade Organization (but the charter was not ratified by the US Congress) and became a multilateral treaty in 1948. GATT set principles for the negotiation of reduction in tariffs, the reduction in other impediments to trade, and the elimination of discriminatory practices. It also provided a forum for exchange of concessions and the settlement of trade disputes. The success of GATT in liberalizing world trade is clear from the substantial increase in world trade since 1948. In this sense at least, GATT was successful in promoting the liberalization of trade. GATT was succeeded by the World Trade Organization in 1995, which is now the body that overseas trade talks and arbitrates over trade disputes.

A strong argument can be made that the world trade system that has emerged from the successive agreements is tailored to suit developed countries. The fact that developing
countries played little role in the first seven rounds is significant in this. Furthermore, Article XVIII of the 1947 agreement provided for “special and differential treatment” for developing countries. In principle, this was intended to assist developing countries by permitting them greater freedom in trade policy. In practice, the consequence has been that it has marginalized those countries in negotiations.

The purpose of GATT was to reduce barriers to trade. This was achieved by negotiating tariff reductions and by the process of tariffication: changing nontariff barriers (e.g., a quota) into an equivalent tariff that led to the same level of trade. Two arguments support the process of tariffication. First, it made the size of the barrier explicit. Second, it was believed to be easier to negotiate future reduction in tariffs than it was to negotiate reductions in nontariff barriers. The downside has been that countries can exploit the tariffication process to set a high initial tariff to obtain something that can be traded in future concessions.

There is no doubt that the successive rounds of talks have significantly reduced the level of tariffs on goods traded among developed countries. This is demonstrated in table 21.4a, which shows that the levels of tariffs are now, on average, very low after the successive rounds of trade talks. The average tariff on goods imported from other high-income countries is 0.8 percent. So, on trade among high-income countries, there is very little scope for further reductions. The position on imports from developing countries is worse but not by a great deal: the average tariff levied by high-income countries on manufactured goods from developing countries is higher, at 3.4 percent. In contrast, developing countries, on average, levy higher tariffs, and the highest average tariffs are on trade among developing countries. A different story emerges from table 21.4b, which reports tariffs on trade in agricultural products. The average tariff is much higher than it is for manufactured products, which is a reflection of the focus of the early rounds of trade talks. What the two figures have in common is that it is the developing countries that levy the highest tariffs on average.

Further details on the protection offered to the agricultural sector, and the food sector more generally, are given in table 21.5. As the table shows, there can be high levels of tariff, significant export subsidies, and production subsidies. Each one of these is an impediment to free trade. It has been argued that the trade policies of developed countries impede development. One justification for this argument is that tariffs on processed food imports are 42 percent in Canada, 65 percent in Japan, and 24 percent in the European Union. In contrast, the tariffs on the least processed food items are 3 percent, 35 percent, and 15 percent respectively. This tariff structure discourages developing countries from moving into food processing and limits the value added that they derive from agriculture. A second justification can be found in the level of
farm subsidies in developed countries. In 1986 to 1988 OECD farm subsidies were equal to 51 percent of the value of farm production. Such subsidies have proved highly persistent: in 2002 they were still 48 percent of value. Subsidies of this level greatly reduce the ability of developing countries to compete in world agricultural markets.

The Uruguay Round that began in 1986 was characterized by a significant increase in the number of countries that participated. This round also extended the scope of the discussion. It included talks on services, trade-related aspects of intellectual property, and trade-related investment measures. The Round was brought to a close in 1993 and was predicted to bring large welfare improvements for developing countries. Estimates since have suggested that developing countries actually suffered a loss as a result of the Uruguay Round. Explanations for the difference between the expected and the actual outcome were that the agreement to remove barriers to trade in textiles was backloaded, so it did not come into effect until several years after the Round was completed. It has also been suggested that the benefits from tariffication were overestimated and the implementation of the agreement proved costly. Furthermore the average OECD tariff on imports from developing countries remained four times larger than the average tariff on imports from developed countries.
Table 21.5
Protection to agriculture

<table>
<thead>
<tr>
<th></th>
<th>Import tariff</th>
<th>Export subsidy</th>
<th>Production subsidy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food grains</td>
<td>23</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Feed grains</td>
<td>97</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>Oilseeds</td>
<td>4</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Meat and livestock</td>
<td>17</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Dairy</td>
<td>23</td>
<td>27</td>
<td>2</td>
</tr>
<tr>
<td>Other agriculture</td>
<td>11</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Other food</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Beverages and tobacco</td>
<td>18</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>


The latest round of talks began in Doha in 2001. These aimed to set the mandate of the World Trade Organization (which replaced GATT in 1995) after the previous meeting in Seattle in 1999 ended in failure. The Doha Declaration outlined a framework for negotiations that focused on the promotion of economic development and led to it being given the name of “The Development Round”. In 2003 the WTO convened a meeting in Cancun in Mexico. This ended without agreement because the developing countries felt that the European Union and United States had not given ground on agricultural subsidies. The developing countries also saw significant costs arising from the implementation of new regimes concerning competition policy, investment regulations, trade and customs procedures, and intellectual property rights.

The Doha talks were suspended in July 2008. The European Union and India were blamed by the United States for trying to exclude too many agricultural products from tariff cuts. China and India refused to lower barriers to imports of subsidized agricultural goods. For several West and Central African countries the US cotton subsidy is a primary issue. In negotiations the United States refused to reduce the subsidy but offered more aid instead. This proved unacceptable. However, there are two additional arguments to consider. First, many agricultural exports from Africa are subsidized or benefit from artificially high international prices. Second, many African countries also have little to gain from a trade agreement because under the EU Everything But Arms programme African LDCs (least developed countries) have duty-free, quota-free access to the EU market. Many sub-Saharan countries have duty-free, quota-free access to markets in developed countries under Economic Partnership Agreements. These facts limit the gains to be obtained from a successful trade agreement.
The reduction of tariffs and the growth of world trade are clear evidence that the rounds of trade talks have been successful. At least part of this success has arisen from the existence of mutual gains to be exploited. These gains have been realized in the trade of developed countries. The Doha Round intentionally set out to help developing countries. The interests of developed and developing countries can be very different, so it may be hard to find mutual gains. More likely, aiding the developing countries will involve losses by the developed countries. It is therefore not surprising that the history of successful talks is now confronted with this round of talks that failed to make progress.

21.10 Conclusions

Increasing globalization of the world economy has enhanced the mobility of capital and labor. This mobility has significant implications for tax policy because a mobile tax base will quickly move to exploit international tax differentials. The ability of countries to design tax policy with no regard for what international competitors are doing is being rapidly diminished. A country that chooses to impose a punitive tax policy will quickly discover other countries reaping the benefits.

The central theme of the chapter is that tax design is confronted with significant challenges in this era of globalization. This was exemplified by the discussion of alternative bases for capital taxation. None of the possible bases was free from shortcomings created by the difficulty of allocating the tax base across countries. These problems have been addressed by international agreements on taxation. International agreements have also succeeded in liberalizing much of international trade. They are likely to play an equally important role in the future development of international tax policy.

Further Reading

An analysis whether the Diamond–Mirrlees lemma extends to international setting are found in:

The earliest paper on transfer pricing is:

An analysis that includes limits on the transfer price is:
Part VII: Multiple Jurisdictions


An extensive discussion of the choice of international tax base can be found in:


An extensive discussion of the calculation of the effective average tax rate and an empirical application is given by:


The major contribution to the analysis of harmonization is:


The equivalence of origin and destination regimes in a range of circumstances is discussed in:


Nonequivalence is explored in:


An extensive but informal discussion of trade agreements and globalization generally is found in:


---

### Exercises


21.2 How are the efficiency conditions of section 21.2.1 modified if there is a cost of $s$ per unit to transport the product between countries?

21.3 For the model of section 21.2.1 show that if one country levies a tax on capital, there will be inefficiency even with perfect mobility of capital and labor.

21.4 The European Union is a customs union. What is a customs union? How does it differ from a free-trade area?

21.5 (Transfer pricing) A firm produces $Y_A$ units of a good in country $A$ and $Y_B$ units in country $B$. These output levels are fixed. The (constant) marginal cost in country $j$, $j = A, B$, is $c_j$ and the price is $p_j$. There are no fixed costs. Let $\xi$ denote the number of units of the good
that is transferred from country A to country B. A negative value of $\xi$ represents a transfer from B to A. Assume profit is taxed at rate $t_j$ in country j.

a. What will determine the transfer price, $\phi$, at which the good is traded between the two divisions of the firm?

b. Assume each country levies a tariff at rate $\tau_j$. What will now determine the transfer price?

21.6 Consider the model of the previous exercise. Can you characterize the equilibrium when the output levels can be chosen?

21.7 Assume that there is an agreed arm’s-length price, $\hat{\phi}$. If the firm deviates from this price, it will be fined an amount $f$. However, fines are not levied with certainty (e.g., an investigation may be needed to confirm the deviation from arm’s length and not all firms can be investigated at any one time). Let the probability of being fined, $q$, be determined by the deviation of $\phi$ from $\hat{\phi}$, so that $q = q(\phi - \hat{\phi})$. The firm produces in country A, sells $X_A$ in country A at price $p_A$, and sells $X_B$ in country B at price $p_B$. There is no production in country B. The cost function is given by $C = C(X_A + X_B)$. Both countries levy a profit tax and country B levies a tariff.

a. What are the necessary conditions for choice of $X_A$ and $X_B$? Are these affected by the choice of agreed arm’s-length price?

b. What condition describes the optimal choice of transfer price? What structure for the probability function $q(\phi - \hat{\phi})$ will provide an incentive for an interior choice of transfer price?

21.8 Consider a world economy composed of two countries, A and B. There is a fixed world population of size $H$ that allocates between countries on the basis of the utility levels offered. Each consumer is characterized by a parameter $x$ that measures their attachment to country A (and by $1 - x$, which is their attachment to country B). The parameter $x$ is distributed uniformly across the population with values from 0 to 1. Let the utility level in country A of a consumer with attachment $x$ be equal to $M - x - t_A$, where $t_A$ is the tax rate in country A (and so utility in B is $M - (1 - x) - t_B$). Assume that there always exists a value of $x$ that partitions the population between those who choose country A and those who choose country B.

a. For given tax levels in the two countries, determine the value of $x$ at which the population partitions.

b. For a given partition of population, state the budget constraint for the government in each country.

c. Using the answers to parts a and b, state the decision problems of the two governments when they act as independent “leviathans” and attempt to maximize revenue. What is the equilibrium level of taxes?

d. What are the efficient taxes that maximize the sum of tax revenues?

e. What are the implications of these findings for tax policy?

21.9 Consider the formula apportionment model with demand functions $p(x^A) = a^A - b^Ax^A$ and $p(x^B) = a^B - b^Bx^B$. How does the allocation of output between the countries when formula apportionment is used compare to the allocation when the actual profit earned in each country is taxed?
21.10 (Keen and Lahiri 1998) Assume there are two identical countries, $j = A, B$, and two identical firms, indexed $i = A, B$. Firm 1 is located in country $A$ and firm 2 in country $B$. The firms produce homogeneous products and engage in Cournot competition on both markets. The marginal cost of production is fixed at 0 in both countries. The countries level specific commodity taxes $t^A$ and $t^B$.

a. Show that the profit-maximizing output levels when the destination principle operates are $x^D_{i,j} = a - t^j$, where $x^D_{i,j}$ is the amount supplied by the firm in country $i$ to country $j$.

b. Show that the optimal destination tax that maximizes national welfare $W^D_j = U^j + \pi^j + R^j$ is $t^D_j = 0$, $j = A, B$. Hence compute the maximized level of national welfare.

c. Repeat part a for the origin principle to show the output levels are $x^O_{i,j} = a - 2t^i + t^j$. Derive the optimal taxes that maximize national welfare.

d. Determine the optimal coordinated taxes that maximize the sum of welfare levels $W^E = W^A + W^B$. Constrain to the solutions to parts b and c and discuss.

21.11 (Hashimzade, Khodavaisi, and Myles 2011) Consider the exercise above embedded in a two-stage game. In the first stage, the countries choose either to use destination taxation or origin taxation. In the second stage, each country chooses its tax rate to maximize its welfare level.

a. Calculate the outputs, taxes, and welfare levels if country $A$ uses the origin principle and country $B$ the destination principle.

b. Construct a payoff matrix for the first stage of the game (strategic choice of principle). Show that the game has a dominant strategy equilibrium.

c. What is the policy implication of the dominant strategy equilibrium?

21.12 Consider a country with a demand curve given by $p = 100 - 2q$ and a domestic supply curve given by $p = 1 + q$, where $p$ is price and $q$ is quantity. Let the world price for the product be $p = 25$.

a. Assume that the country is small. Calculate the levels of consumer surplus and producer surplus when there is no tariff.

b. A tariff of $\tau = 1$ is now introduced. Calculate the levels of producer surplus, consumer surplus, and government revenue. Show that the tariff causes a deadweight loss.

c. Now assume that the country is large. How far must the world price fall when the tariff is introduced to lead to a welfare gain?

21.13 Demonstrate that the introduction of a “small” tariff will always benefit a large country. (Hint: Differentiate welfare with respect to the tariff and evaluate with the tariff equal to zero.)

21.14 A firm has to choose to locate in country $A$ or country $B$. It will earn profit $\pi$ wherever it locates, but will earn an additional location-specific rent in $A$. The location specific-rent is uncertain. It assumes the value $\theta^H$ with probability $p$ and value $\theta^L$ with probability $1 - p$, with $\theta^H > \theta^L$. Country $A$ earns income $Y^A$ if the firm chooses to locate there and levies a tax $T^A$, while country $B$ earns income $Y^B$ and levies a tax $T^B$. The countries choose their tax rates and the firm chooses its location before the location-specific rent is realized.

a. For given values of the tax rates determine the payoff of the firm in the two countries for the two values of the location-specific rent.
b. When will the firm choose to locate in country \( A \) for both values of the rent? When will it choose to locate in country \( B \)?

c. Determine the tax rate that \( B \) will set and derive a condition that ensures location in \( B \) will occur.

21.15 Assume there are two regions 1 and 2 that are located on a line of length 1. Each region covers \( \frac{1}{2} \) of the line. Consumers are uniformly distributed in both regions with density \( h \). A good is sold in each country at the before-tax price, \( p \). The tax rate in region 1 is \( t_1 \) and in region 2 is \( t_2 \), with \( t_1 < t_2 \).

a. Assume that consumers can transport the good at cost \( c \) per unit of distance. Determine the set of consumers that will cross-border shop and the total transport costs incurred.

b. Now assume instead that firms transport the good at a transport cost of \( \delta c \), with \( \delta < 1 \). Determine the set of consumers who are supplied with a good that crosses the border. Compute the total transport cost and compare with that in part b.

c. Do the countries generate more tax revenue with cross-border shopping or with firms transporting?

21.16 Toyota has the production capacity to produce 550,000 vehicles a year in Thailand. Much of this production is exported to Singapore, then re-exported from Singapore to final markets. Why?

21.17 (Cross-border shopping) International traded goods can be taxed either in the country where they are consumed (destination principle), or where they are produced (origin principle). In open economies such as European Union where border controls have been abolished, consumers can shop around for the lowest price. Under GATT rules, international commodity trade is generally taxed under the destination principle.

a. What are the tax implications for the goods that are exported and for the goods that are imported?

b. Consider two countries \( c = A, B \) and two goods \( i = 1, 2 \) with commodities tax rates \( t^c_i \) that can be differentiated across goods and countries. Show that consumer arbitrage under the destination principle will always guarantee international production efficiency even in the presence of differentiated commodity taxation \( (t^A_i \neq t^B_i) \) and international differences in tax rates \( (t^A_i \neq t^B_i) \).

21.18 Consider the previous exercise and suppose that each country levies a uniform tax rate on both goods \( (t^A_i = t^B_i) \).

a. In this case show that neither production nor consumption decisions are distorted internationally

b. Now suppose that tax rates are differentiated across goods within each country. For instance, one good (e.g., good 1) is untaxed by the two countries. Show that in this case consumption decisions are distorted by commodity tax differences.

21.19 Consider the previous exercise and suppose that commodity trade is now taxed under the origin principle.

a. What are the implications for the goods that are exported and for the goods that are imported?
b. Show that consumer arbitrage under the origin principle will always guarantee international consumption efficiency even in the presence of differentiated commodity taxation \( t_1^c \neq t_2^c \) and international differences in tax rates \( t_1^A \neq t_1^B \).

c. Show that if \( t_1^c = t_2^c \) (uniform commodity taxation in each country) but tax rates differs across countries \( t_1^A \neq t_1^B \), then the production decisions are undistorted internationally.

d. Furthermore in the case of part c, show that a general consumption tax is equivalent to a general production tax.

21.20 To combat profit-shifting by using transfer-pricing, American law and the OECD model require the use of arm’s-length prices for international transactions. Arm’s-length prices are market prices: the prices that unrelated parties transacting at “arm’s length” would use. What problems do you see with the application of this principle?

21.21 The interaction of two tax systems, each belonging to a different country, can result in double taxation. Every country seeks to tax the income generated within its territory on the basis of one or more connecting factors such as location of the source, residence of taxable entity, and maintenance of permanent establishment. Double taxation of the same income in the hands of same entity would give rise to harsh consequences and impair economic development. Double Taxation Avoidance Agreements between two countries therefore aim at eliminating or mitigating the incidence of double taxation. The object of Double Taxation Agreements is to provide for the tax claims of two governments, both legitimately interested in taxing a particular source of income, either by assigning to one of the two the whole claim or else by prescribing the basis on which tax claims is to be shared between them. What are the objectives of these Double Taxation Agreements?

21.22 (Double taxation) What methods are used to eliminate double taxation?

21.23 This exercise seeks to illustrate the optimal capital income tax problem for an economy with an open capital market and two types of capital, one of which is private, mobile, and “directly productive” (i.e., generates profits) while the other is public, immobile, and “indirectly productive” (i.e., infrastructure). The problem is how to set a level of tax so as to fund infrastructure and maximize welfare when capital can move abroad. Let us consider an economy with three production factors: immobile labor \( (L) \), mobile capital \( (K) \), and public infrastructure \( (P) \). National income \( (Y) \) is determined by

\[
Y = L^\alpha K^\beta P^\gamma, \quad 0 < \alpha, \beta, \gamma < 1.
\]

The return on capital is given by

\[
r = \frac{\partial Y}{\partial K} = \beta L^\alpha K^{\beta-1} P^\gamma.
\]

The capital stock is endogenous as capital can flow inward or outward, depending on the after-tax rate of return. At equilibrium then

\[
(1 - t)r = (1 - t^*)r^*.
\]

where \( t \) is the domestic tax rate, \( t^* \) is the foreign tax rate, and \( r^* \) is the foreign return on capital.

a. Calculate the equilibrium capital stock, and show that this will give an elasticity of the capital stock to the tax rate that is negative and greater than unity (in absolute value) if \( \beta > \frac{1-2\gamma}{1-\gamma} \).
b. Show that for a fixed level of public infrastructure \( P \) any increase in the domestic tax rate \( t \) will reduce national income \( Y \). Show also that a reduction in the foreign tax rate \( t^* \) would require a reduction in \( t \) as well to ensure national income remains unchanged. Explain.

Now suppose that tax revenues serve to finance infrastructure \( P \), so that the stock of public infrastructure is no longer independent of the tax rate but is given by

\[
P = trK.
\]

c. Show that national income is not monotonically decreasing with \( t \) and that the optimal tax rate is positive. Explain the optimal tax pattern in relation to the relative productivities of private capital and public infrastructure.

d. What does this model predict in terms of a possible race to the bottom?
VIII  ISSUES OF TIME
22.1 Introduction

Time is an essential component of economic activity. The passage of time sees the birth and death of consumers and the purchase, depreciation, and eventual obsolescence of capital. It sees new products and production processes introduced, and provides a motive for borrowing and saving. Time also brings with it new and important issues in public economics such as the benefits from the provision of social security (pensions) and the effect of government policy on economic growth. This chapter and the two that follow are devoted to exploring these issues.

The competitive economy described in chapter 2 provided a firm foundation for the discussion of economic efficiency and equity in a static setting. This economy underpinned the analysis of efficiency failures and the policy responses to them. It taught a number of important lessons about economic modeling. For all these reasons it has been one of the most influential and durable models in economics. Despite its usefulness, the model has shortcomings when it is applied to economic issues involving time. We presented the competitive economy as being atemporal—having an absence of any time structure. A temporal structure can be added by interpreting commodities as being available at different times so that the commodity “bread for delivery today” is a distinct commodity from “bread available tomorrow.” The list of commodities traded is then extended to include all commodities at all points in time. Since only the labeling of commodities has changed, all the results derived for the economy—the efficiency theorems in particular—remain valid.

Although analyzing time in this way has the benefit of simplicity, it also has one major shortcoming. This shortcoming is best understood by considering the implications of the equilibrium concept we applied to the model. Equilibrium was found by selecting a set of prices that equate supply and demand on all markets. Moreover it was assumed that no trade takes place until these equilibrium prices are announced. The implication of this structure is that all agreements to trade present commodities, and commodities to be consumed in the future, have to be made at the start of the economy. That is, contracts have to be negotiated and agreed, and equilibrium prices determined, before production and consumption can take place. This produces a poor representation of an intertemporal economy that misses the gradual unfolding of trade, which is the very essence of time. It is also an untenable interpretation: the need to make trades now
for all commodities into the future requires all consumers and firms to be present at the start of the economy and to know what all future products will be. Consequently, if issues of time are to be properly addressed, a better model is needed. We consider two alternative models of time: the overlapping generations economy, whose focus is on population structure, in this chapter, and an alternative set of models, focusing on growth through capital accumulation and technological innovation, in chapter 24.

The overlapping generations economy is one alternative to the competitive economy that introduces time in a more convincing manner. This model has the basic feature that the economy evolves over time with new consumers being born at the start of each period and old ones dying. At any point in time the population consists of a mix of old and young consumers. The life spans of these generations of young and old overlap, which gives the model its name and provides a motive for trade between generations at different points in their life cycles. This evolution of the population allows the overlapping generations economy to address many issues of interest in public economics.

Overlapping generations economies are important, not only because they give a simple yet realistic model of the life cycle but also because of their many surprising properties. Foremost among these properties, and the one that is focus of this chapter, is that the competitive equilibrium can fail to be Pareto-efficient despite the absence of any of the sources of market failure identified in part III. The potential failure of Pareto-efficiency provides an efficiency-based justification for assessing the benefits of government intervention. Among such interventions the one with the most important policy implications is social security (or pensions). Social security can transfer wealth across points in the life cycle and between the generations. Given the impending “pensions crises” that are slowly developing in many advanced economies as the elderly population increases relative to the number of workers, social security is an issue of major policy concern.

This chapter sets out the structure of a basic overlapping generations economy with production. It presents the decision problems facing the consumers and the producers in the economy. The solutions to these problems are then used to characterize the equilibrium of the economy and to determine the steady-state equilibrium in which consumption and output per capita are constant. Both Pareto-efficient steady states and the optimal Golden Rule steady state are characterized. It is then shown that the economy can settle into an inefficient steady state that is the important result from a policy perspective. The nature of the inefficiency and the reason why it arises are discussed.
22.2 Overlapping Generations

22.2.1 Time and Generations

The two features of economic activity connected with the passage of time that we wish to capture are the accumulation of capital and the fact that the life span of each individual is short relative to the life span of the economy. The model we now develop incorporates the first aspect by allowing capital to be transferable across time periods and to depreciate steadily over time. The second aspect is introduced by letting each consumer have a finite life set within the infinite life of the economy.

In the overlapping generations economy, time is divided into discrete periods, with the length of the unit time interval being equal to the time between the birth of one generation and the birth of the next. There is no end period for the economy; instead, economic activity is expected to continue indefinitely. At the beginning of each period a new generation of young consumers is born. Each consumer lives for two periods of time. The population grows at a constant rate, so, if the rate of population growth is positive, each generation is larger than the previous one. Generation \( t \) is defined as the set of consumers who are born at the start of period \( t \). Denoting the population growth rate by \( n \), if generation \( t \) is of size \( H_t \), then the size of generation \( t + 1 \), \( H_{t+1} \), is given by \( H_{t+1} = [1 + n] H_t \).

The population at any point in time is made up of young and old consumers; it is this overlapping of two consecutive generations that gives the model its name. This generational structure is shown in figure 22.1 where the solid lines represent the life span of a generation. It is the differing motives for trade for the old and the young, due to their different life-cycle positions, that give economic content to the model.

At each point in time the economy has a single good that is produced using capital and labor. This good can either be consumed or saved to be used as the capital input in the next period. (Thinking of potatoes may be helpful. When harvested, they can be either eaten or put aside to be used for planting in the next year.) The existence of capital as a store of value allows consumers to carry purchasing power from one period to the next. To simplify, we assume that capital does not depreciate during the production process. Consumers plan their consumption to maximize lifetime utility and the level of production is chosen so as to maximize profits. All markets are competitive. An allocation of production is feasible for the economy if consumption plus saving by the two generations alive at each point in time is no greater than total output.
22.2.2 Consumers

The modeling of consumers is designed to capture a very simple form of life-cycle behavior. Each consumer works only during the first period of their life and inelastically supplies one unit of labor. This unit of labor is their entire endowment. Hence the total quantity of labor in the economy is equal to the number of young consumers. In their second period of life each consumer is retired and supplies no labor. Retired from work, old consumers live off the savings they accumulated when working. They are fully aware of their own mortality and plan their consumption profile accordingly. The income earned by a consumer during the first period of their life is divided between consumption and savings. Second-period consumption is equal to savings plus interest. With the exception of their date of birth, consumers are otherwise identical.

All consumers have identical preferences over consumption in the two periods of life. For a consumer born in period $t$, these preferences are represented by the utility function

$$U = U(x^t_t, x^{t+1}_t), \quad (22.1)$$

where $x^t_t$ is consumption when young and $x^{t+1}_t$ consumption when old. There is no explicit disutility from the supply of the single unit of labor in the first period of life.

The budget constraint of a typical consumer can be constructed by noting that labor income is equal to the sum of consumption and saving. In the first period of life, consumption, $x^t_t$, and saving, $s_t$, must satisfy the budget constraint

$$w_t = x^t_t + s_t, \quad (22.2)$$
where \( w_t \) is the wage received for the single unit of labor. Savings accrue interest at rate \( r_{t+1} \) (with interest paid in period \( t+1 \)), so the value of second-period consumption \( x_{t+1}^t \) is given by
\[
x_{t+1}^t = [1 + r_{t+1}] s_t.
\] (22.3)

Combining (22.2) and (22.3) gives the life-cycle budget constraint
\[
w_t = x_t^t + \frac{x_{t+1}^t}{[1 + r_{t+1}]}.
\] (22.4)

Before proceeding to further analysis of consumer choice, it is worth emphasizing an important point: there are no financial assets in this economy. Instead, saving takes the form of investment in real capital. The interest rate is therefore equal to the return on capital, and the same interest rate guides the purchase of capital by firms.

From (22.1) and (22.4) the utility-maximizing consumption plan satisfies the first-order condition
\[
\frac{\partial U}{\partial x_t} = \frac{\partial U}{\partial x_{t+1}^t} = [1 + r_{t+1}].
\] (22.5)

In (22.5) the left-hand side is the intertemporal marginal rate of substitution between consumption in the two periods of life. The right-hand side is the intertemporal marginal rate of transformation. The solution to this choice problem is illustrated in figure 22.2.

### 22.2.3 Production

The productive sector of the economy is assumed to consist of many competitive firms all producing with the same constant-returns-to-scale production technology. These assumptions allow the firms to be aggregated into one single representative firm modeled by an aggregate production function. Using a representative firm greatly simplifies the presentation.

It has been assumed that the capital used in production does not depreciate (this simplifies matters but has no significant economic consequences). At the end of the production process in each period, the firm has (1) the (undepreciated) capital used in production and (2) new output. The sum of these is the total output of the economy, which is divided between saving (to be re-invested as capital) and consumption. To be consistent, the aggregate production function is defined to measure the *gross* output of
the firm, which is the sum of new output plus the undepreciated capital. Denote this production function by \( F(K_t, L_t) \), where \( K_t \) is the capital stock in period \( t \) and \( L_t \) is aggregate labor supply. An allocation is feasible if gross output is equal to the sum of consumption for the two generations alive at time \( t \) plus savings

\[
F(K_t, L_t) = H_t x_t^t + H_{t-1} x_{t-1}^t + H_t s_t. \tag{22.6}
\]

The representative firm chooses its use of capital and labor to maximize profits, \( \pi_t \), where

\[
\pi_t = F(K_t, L_t) - w_t L_t - r_t K_t. \tag{22.7}
\]

Note that this expression for profit values net output and the undepreciated capital equally and assigns a rental rate of \( r_t \) for the use of capital. The necessary condition for choice of the level of capital is

\[
F_K = r_t. \tag{22.8}
\]

This is just the usual statement that capital should be employed up to the point at which its marginal product, \( F_K \), is equal to its cost, \( r_t \). The first-order condition for the quantity of labor input is

\[
F_L = w_t, \tag{22.9}
\]
so labor is employed up to the point at which its marginal product, $F_L$, is equal to the wage, $w_t$.

This development of the firm’s decision problem allows the results to be related to standard results from microeconomics. However, an alternative presentation is more helpful in the context of an overlapping generations economy with a possibly variable population. In this case what matters for economic welfare is not just how much is produced but, instead, how much is produced per unit of labor. An increase in production per unit of labor (with labor equal to the number of young consumers) can allow an unambiguous increase in welfare provided that it is correctly distributed, but an increase in production, without reference to the size of population, cannot. To capture these observations, it is preferable to re-phrase the formulation of the production function.

It is now assumed that the production function satisfies constant returns to scale. This assumption makes it possible to write

$$Y_t = L_t F \left( \frac{K_t}{L_t}, 1 \right) = L_t f \left( \frac{K_t}{L_t} \right),$$  \hspace{1cm} (22.10)

where $\frac{K_t}{L_t}$ is the capital–labor ratio. Defining $y_t = \frac{Y_t}{L_t}$ and $k_t = \frac{K_t}{L_t}$, we determine net output per unit of labor by way of a function that has the capital–labor ratio as its sole argument

$$y_t = f(k_t).$$  \hspace{1cm} (22.11)

It is assumed that this function satisfies $f(0) = 0$, $f' > 0$, and $f'' < 0$ so that no output can be produced without capital and the marginal product is positive but decreasing. From (22.10) it follows that the marginal product of capital is

$$\frac{\partial Y_t}{\partial K_t} \equiv F_K = f',$$  \hspace{1cm} (22.12)

and the marginal product of labor

$$\frac{\partial Y_t}{\partial L_t} \equiv F_L = f - \frac{K_t}{L_t} f' = f - k_t f'.$$  \hspace{1cm} (22.13)

These derivatives can be used to rewrite (22.8) and (22.9) as

$$f'(k_t) = r_t$$  \hspace{1cm} (22.14)

and

$$f(k_t) - k_t f'(k_t) = w_t.$$  \hspace{1cm} (22.15)
Conditions (22.14) and (22.15) represent the optimal choice of capital and labor for the firm when the production function is expressed in terms of the capital–labor ratio. This pair of conditions characterize the choices arising from profit maximization by the firm.

22.3 Equilibrium

At an equilibrium for the overlapping generations economy it is necessary that consumers maximize utility, that the representative firm maximizes profit, and that all markets clear. Since there is a single good that can be used as capital or consumed, market clearing can be captured by the equality of demand and supply on the capital market.

Granted this fact, there are two ways in which equilibrium can be viewed. The first is to consider the intertemporal equilibrium of the economy. Here, by intertemporal, is meant a sequence of values for the economic variables that ensure that markets are in equilibrium in every time period. This intertemporal equilibrium fully determines the time path for the endogenous variables \( (x_t, x_{t+1}, k_t, w_t, \text{and } r_t) \) and hence their changes from one period to the next. The alternative form of equilibrium is to consider the steady state of the economy. The steady state is the situation where the endogenous variables remain constant over time. Such an equilibrium can be thought of as a long-run position for the economy. By definition, once the economy reaches a steady state, it never leaves it.

To describe either form of equilibrium, it is first necessary to characterize equilibrium in the capital market. Equilibrium is achieved when the quantity of capital used in production is equal to the level of savings, since capital is the only store of value for saving. By definition, saving is labor income less consumption, so \( s_t = w_t - x_t \). Hence, since there are \( H_t \) young consumers in period \( t \), the equality of total savings in period \( t \) with capital used in period \( t + 1 \) requires that

\[
H_t[w_t - x_t] = K_{t+1}. \tag{22.16}
\]

Dividing through by \( H_t \), and recalling that \( H_{t+1} = [1 + n] H_t \) and \( H_t = L_t \), we can express this in terms of the capital–labor ratio as

\[
w_t - x_t = k_{t+1}[1 + n]. \tag{22.17}
\]

When (22.17) is satisfied, there is equilibrium in the capital market.
22.3.1 Intertemporal Equilibrium

An intertemporal equilibrium is a sequence \( \{x_t^t, x_{t+1}^t, k_t, w_t, r_t\} \) of the endogenous variables that attains equilibrium in every time period \( t \). In each time period all consumers must maximize utility, the representative firm must maximize profit, and the capital market must be in equilibrium. Putting these together, we have the set of conditions that must be simultaneously satisfied for the economy to be in equilibrium:

- Utility maximization: (22.4), (22.5)
- Profit maximization: (22.14), (22.15)
- Market clearing: (22.17)

The equilibrium determined by these conditions should be seen as one where the economy develops over time. The way that this works can be understood by following the economy from its very beginning. Let the economy have an initial capital stock, \( k_1 \), in period 1. This capital stock is endowed by nature and belongs to consumers who are already in the second period of life at the start of economic activity. The level of capital and the initial labor force determine the interest rate, \( r_1 \), and wage rate, \( w_1 \), from (22.14) and (22.15). Correspondingly (22.5), (22.4), and (22.17) simultaneously determine \( x_1^1, x_1^2 \), and \( k_2 \). Starting with \( k_2 \), the process can be repeated for the next time period. Continuing forward in this way generates the entire equilibrium path of the economy.

Although the intertemporal behavior of the overlapping generations economy is of great analytical interest, it will not be pursued in detail here. Instead, our focus will be on the steady state from this point on.

22.3.2 Steady State

In a steady state all variables are constant. Consequently consumers in all generations must have the same lifetime consumption plan. The quantity of capital per worker must also remain constant. These observations suggest an interpretation of the steady state as the long-run equilibrium in which the economy has reached the limit of its development (but note that this interpretation is only strictly true if the economy converges to the steady state).

Since all variables are constant in the steady state, the notation can be simplified by dropping the subscripts referring to time. The steady-state equations determining the wage and the interest rate are \( w = f(k) - k f'(k) \) and \( r = f'(k) \), where \( k, w, \) and \( r \) are
the (constant) capital–labor ratio, wage rate and interest rate. Each consumer’s budget constraint can then be written as

\[ x^1 + \frac{x^2}{1 + f'(k)} = f(k) - kf'(k), \]  

(22.18)

where \( x^1 \) and \( x^2 \) are the steady-state consumption levels in the first period and the second period of the consumer’s life. The steady-state capital market equilibrium condition becomes

\[ f(k) - kf'(k) - x^1 = [1 + n]k. \]  

(22.19)

These equations can be used to provide a helpful means of displaying the steady-state equilibrium. Solving (22.18) and (22.19) for the consumption levels \( x^1 \) and \( x^2 \) gives

\[ x^1 = f(k) - kf'(k) - [1 + n]k, \]  

(22.20)

and

\[ x^2 = [1 + n]k \left[ 1 + f'(k) \right]. \]  

(22.21)

The interpretation of (22.20) and (22.21) is that each value of the steady-state capital–labor ratio \( k \) implies a steady-state level of first-period consumption from (22.20) and of second-period consumption from (22.21). As \( k \) is varied, this pair of equations generates a locus of \( \{x^1, x^2\} \) pairs. This is termed the consumption possibility frontier. It shows the steady-state consumption plans that are possible for alternative capital–labor ratios.

There are basic economic reasons for expecting the consumption possibility frontier to describe a non-monotonic relationship between \( x^1 \) and \( x^2 \) as illustrated in figure 22.3. When \( k = 0 \), it can be seen immediately that \( x^1 \) and \( x^2 \) are both zero—no output can be produced without capital. This is illustrated by the frontier beginning at the origin in the figure. When \( k \) becomes positive, \( x^1 \) and \( x^2 \) also become positive. As \( k \) is increased, we move further around the frontier with both \( x^1 \) and \( x^2 \) increasing. For large values of \( k \), \( x^1 \) may begin to fall; \( x^1 \) can even become zero since \( f(k) \) increases at an ever slower rate while \([1 + n]k\) increases at a constant rate. The actual shape of the frontier depends on how quickly the returns to capital decrease as the capital input is increased while holding labor input constant. What is underlying this is that low values of \( k \) do not allow much to be produced, so consumption must be low. As \( k \) increases, more consumption becomes possible. However, at high values of \( k \), decreasing returns to capital become important and consumption must be decreased in order to support the reproduction of a very high capital stock.
Chapter 22: Intertemporal Efficiency

Figure 22.3
Consumption possibilities

The importance of this construction is the following interpretation: All points on the frontier are potentially steady-state equilibria. Each point determines a pair \( \{x^1, x^2\} \) and an implied value of the capital–labor ratio, \( k \). The steady state that will actually arise as the competitive equilibrium of the economy is determined by the interaction of consumer preferences and the consumption possibility frontier. The question then arises as to the efficiency properties of the consumption pair \( \{x^1, x^2\} \), or equivalently of the value of \( k \). That is, are all values of \( k \) equally good or are some preferable to others? If some are preferable, will the competitive economic activity result in an optimal value of \( k \) in equilibrium? The answers to these questions are given in the following sections. Before proceeding to these discussions, we will look at how the competitive equilibrium is determined.

The optimal choice of the consumer is a point where the highest attainable indifference curve is tangential to the budget constraint. The budget constraint has gradient \(-[1+r]\). In a steady-state equilibrium the consumption plan \( \{x^1, x^2\} \) must also be on the consumption possibility frontier and be tangential to the budget constraint. The equilibrium is found by moving around the frontier until a value of \( k \) is reached at which the indifference curve is tangent to the budget constraint defined by the rate of interest, \( r = f'(k) \), at that level of capital. A steady-state equilibrium satisfying this condition and two nonequilibrium allocations, at \( a \) and \( b \), are shown in figure 22.4.
22.4 Optimality and Efficiency

22.4.1 The Golden Rule

The central message of the previous section was that the competitive equilibrium will occur at some point on the consumption possibility frontier. The consumer’s preferences, in conjunction with the production function, will determine precisely which point this is. Having reached this conclusion, it is now possible to determine whether any of the points on the frontier are preferable to others.

To do this, it is first necessary to clarify in what sense one point can be preferable to another. In a steady state every consumer in every generation has an identical lifetime consumption plan. Consequently there are no equity issues involved, so “preferable” will have to be stronger than just raising welfare through redistribution. If one point is to be preferred to another, it must in the sense of a Pareto improvement. But if a Pareto-preferred allocation can be found, it implies that the competitive equilibrium is not efficient—a finding that would show the First Theorem of Welfare Economics does not apply to the overlapping generations economy.

The analytical strategy that we employ is to show that there is an optimal value of the capital–labor ratio. This is the content of this section. The next step is to show that there are other values that are Pareto-inferior to the optimal value. This is undertaken in the next section.
In the construction of the consumption possibility frontier it was noted that each consumption allocation was related to a unique value of the capital–labor ratio. This observation allows the study of the efficiency of alternative consumption allocations to be reinterpreted as the study of alternative capital–labor ratios. The optimum level of the capital–labor ratio can then be taken as that which maximizes total consumption in each period. The relation that this level of capital satisfies is termed the Golden Rule and the resulting capital–labor ratio is the Golden Rule level. Rules such as this are important throughout the theory of economic growth.

The Golden Rule is derived by observing that total consumption in any period must be equal to total output less additions to the capital stock, or

$$x^t_H + x^t_{t-1}H_{t-1} = H_t f (k_t) - H_t[k_t+1 [1 + n] - k_t].$$  \hspace{1cm} (22.22)

At a steady-state equilibrium (22.22) reduces to a constraint on steady-state consumption levels given the capital stock:

$$x^1 + \frac{x^2}{[1 + n]} = f (k) - nk.$$ \hspace{1cm} (22.23)

The Golden Rule steady state consumption levels are found by maximizing the steady-state level of utility, $U(x^1, x^2)$, subject to the constraint (22.23). Substituting for $x^2$, the Golden Rule level of first-period consumption, $x^1$, and capital, $k$, solve

$$\max_{\{x^1, k\}} U(x^1, [1 + n] \left[f (k) - nk - x^1 \right]).$$ \hspace{1cm} (22.24)

The necessary conditions for $x^1$ and $k$ are

$$U_1 - [1 + n] U_2 = 0$$ \hspace{1cm} (22.25)

and

$$U_2[1 + n] \left[f' (k) - n \right] = 0.$$ \hspace{1cm} (22.26)

The first condition implies that at the Golden Rule the marginal rate of substitution between consumption in first period of life and consumption in second period of life is given by $[1 + n]$. The second condition implies that the optimal capital–labor ratio, denoted $k^*$, satisfies

$$f'(k^*) = n.$$ \hspace{1cm} (22.27)
The condition in (22.27) is called the Golden Rule and the capital–labor ratio $k^*$ is termed the Golden Rule capital–labor ratio. It is the optimal capital–labor ratio in the sense that it maximizes steady-state utility. Its relation to Pareto-efficiency is addressed in the next section.

Returning to the competitive economy, we have a simple rule for determining whether its equilibrium achieves the Golden Rule. The choice of capital by the firm ensures that $f' = r$. Combining this equality with the Golden Rule shows that if the competitive economy reaches a steady-state equilibrium with $r = n$, this equilibrium will satisfy the Golden Rule. Since no other equilibrium will, $r = n$ must be the Golden Rule rate of interest. Hence the competitive economy achieves the Golden Rule when its interest rate is equal to the rate of population growth. If this occurs, it will have a capital–labor ratio $k = k^*$.

Some further analysis provides more insight into the structure of the Golden Rule equilibrium. From (22.20) and (22.21) total differentiation shows that

$$
\frac{dx^1}{dx^2} = -\frac{1 + n}{1 + n + kf''}. 
$$

This expression is the gradient of the consumption possibility frontier at a point corresponding to a given value of $k$. At the Golden Rule capital–labor ratio $k^*$, with $f' = n$, (22.28) reduces to

$$
\frac{dx^2}{dx^1} = -[1 + n]. 
$$

To understand what this implies, recall that with $r = n$ the gradient of the consumer’s budget constraint is $- [1 + n]$. The maximal budget constraint with this gradient will thus be tangential to the consumption possibility frontier at the point corresponding to the Golden Rule capital–labor ratio. Denote the implied consumption levels at this point by $x_1^*, x_2^*$ (see figure 22.5). Therefore, for the competitive equilibrium to achieve the Golden Rule, when offered this budget constraint, the consumer must want to choose the quantities $x_1^*, x_2^*$. The solid indifference curve in figure 22.5 illustrates such an outcome. The coincidence of the Golden Rule allocation and the consumer’s choice can only happen with an unlikely combination of preferences and technology. In fact the optimal choice for the consumer with this budget constraint will almost always be somewhere other than at the Golden Rule. With the dashed indifference curve in figure 22.5 the Golden Rule will not be a competitive equilibrium.
22.4.2 Pareto-Efficiency

Having now characterized the Golden Rule capital–labor ratio and its corresponding rate of interest, it is possible to address the question of Pareto-efficiency. To do this, first note that if $k > k^*$, so that the equilibrium capital stock exceeds the Golden Rule level, then $r < n$ and the rate of interest is less than the rate of population growth. The converse is true if $k < k^*$. These relations are a simple consequence of the decision process of the firm and the concavity of the production function. In treating Pareto-efficiency, we should take the cases of $k > k^*$ and $k \leq k^*$ separately. We begin with $k > k^*$.

If the capital–labor ratio is above $k^*$, the economy has overaccumulated along its growth path. Consequently it is in a steady state with an excessive capital–labor ratio. The analysis of the Golden Rule has shown that such a steady state fails to maximize consumption per head. We now show that it is also not Pareto-efficient. This is achieved by describing a Pareto-improving reallocation for the economy.

The first point to note is that there is a single good available in the economy, so capital simply represents units of the good withheld from consumption. It is therefore feasible at any point in time to reduce the capital stock and to raise consumption simply by consuming some of the capital stock (i.e., eating the potatoes put aside for planting.) So, in an economy that has overaccumulated, the consumers alive in any period with an excessive capital stock ($k > k^*$) can consume some of the existing capital stock so
as to reduce the stock to the level $k^*$. Consumption of the excess capital stock has two consequences:

1. It raises the welfare of the existing generations because it increases their present consumption at no cost.
2. It raises the welfare of all following generations because it places the economy on the Golden Rule path and so maximizes their consumption.

Clearly, undertaking consumption of the capital stock above the Golden Rule level raises the consumption of those currently alive and of all those who follow. This is a certain Pareto improvement. Therefore any steady state with $k > k^*$ and $r < n$ is not Pareto-efficient.

When $k \leq k^*$, no such Pareto improvements can be found. In this case the economy has accumulated insufficient capital over the growth path. For the economy to move to the Golden Rule, it must accumulate additional capital. This can only be achieved if one (or more) of the generations is willing to forgo consumption. This has two effects:

1. It reduces the welfare of the generations who give up consumption to increase the capital stock.
2. It raises the welfare of all following generations because it moves the economy closer (or on to) the Golden Rule.

Consequently, since at least one generation must reduce their consumption in the transition to a Golden Rule steady state, no Pareto improvement can be made from the initial position. Therefore all states with $k \leq k^*$ are Pareto-efficient.

In summary, any steady state with $k > k^*$ and $r < n$ is not Pareto-efficient. Such states are called dynamically inefficient. Those with $k \leq k^*$ are Pareto-efficient and are termed dynamically efficient. The fact that steady states that are not Pareto-efficient can exist despite the model satisfying all the standard behavioral and informational assumptions that describe a competitive economy shows that the First Theorem of Welfare Economics cannot be extended to include overlapping generations economies. Therefore these economies demonstrate that competition does not always lead to efficiency even when none of the standard causes of inefficiency (e.g., monopoly) are present. This observation is the most fundamental to emerge out of the analysis of the overlapping generations economy. As we will see, it provides the motive for studying numerous forms of policy intervention.

The discussion has concluded that a steady-state equilibrium with $r < n$ is not Pareto-efficient despite the economy satisfying all the standard competitive assumptions.
However, it might still be suspected that to arrive at a steady state with \( r < n \) requires some unusual structure to be placed on the economy. To show that this is not so, consider the following example: The utility function of the single consumer in each generation is given by

\[
U(x^1, x^2) = \beta \log x^1 + (1 - \beta) \log x^2, \tag{22.30}
\]

and the production function is Cobb–Douglas with \( y = Ak^\alpha \). From the five equations describing a steady state, the interest rate can be calculated to be (the derivation of this result is undertaken in exercise 22.7)

\[
r = \frac{\alpha [1 + n]}{[1 - \beta][1 - \alpha]}. \tag{22.31}
\]

This will only be equal to the Golden Rule rate when

\[
n = \frac{\alpha}{[1 - \beta][1 - \alpha] - \alpha}. \tag{22.32}
\]

If preferences and production do not satisfy this condition, and there is no reason why they should, the economy will not grow on the Golden Rule growth path. This example illustrates that a Golden Rule economy will be the exception rather than the norm. A dynamically inefficient steady state occurs when \( r < n \). Using the solution for \( r \) in (22.31), we can write this inequality as

\[
\frac{\alpha}{1 - \alpha} < \frac{n}{1 + n} [1 - \beta]. \tag{22.33}
\]

From (22.33) it can been seen that inequality is most likely to arise when the following occurs:

1. The increase in output following a marginal increase in the capital–labor ratio is small (\( \alpha \) low).
2. The rate of population growth is high (\( n \) large).
3. The consumer places a high weight on second-period consumption (\( 1 - \beta \) large).

In conclusion, the efficiency of the steady-state equilibrium is dependent on the relation of the capital stock to the Golden Rule level. The economy may reach equilibrium at a dynamically inefficient steady state that is not Pareto-efficient. In such a case a Pareto improvement can be achieved by consuming some of the capital stock. A Cobb–Douglas example illustrated the factors that may lead to dynamic inefficiency.
Now that it has been demonstrated that the competitive equilibrium of the overlapping generations economy need not be Pareto-efficient, it remains to explain why. There is a significant difficulty in doing this: there is no agreed explanation for the inefficiency. To explore this further, consider a very simple variant of the economy. In this variant there is no production, and hence no capital. Instead, each young consumer is endowed with one unit of a consumption good while old consumers are endowed with nothing. Clearly, each consumer would like to even out consumption over the life span and so would trade some consumption when young for consumption when old. But such a trade is not possible. The young could give the old some consumption, but the old have nothing to trade in exchange. Therefore the only equilibrium is that no trade takes place (a position called autarky), whereas a Pareto-efficient allocation would have consumption in both periods of life.

It was in this setting that the inefficiency result was discovered. At first sight it might seem that it is just the structure of the economy—in particular, the lack of any way of transferring purchasing power across periods—that prevents the attainment of Pareto-efficiency. There are two responses to this. First, in the standard competitive economy the efficiency result holds independent of any particular details of the economy. Second, the analysis above has already shown that inefficiency can hold even if consumers are able to hold savings that transfer purchasing power across periods. Inefficiency usually arises when the market provides the wrong price signals. This is the case, for example, with monopoly and externalities. Here one might suspect the inefficiency to be due to an interest rate that provides the wrong signal for investment. But this cannot be the explanation because in a model without production there is no interest rate.

There is one point that is agreed on. Because the overlapping generations economy has no end, it can have an infinite number of consumers and, counting the good in each period as a different good, an infinite number of goods. The inefficiency only arises if there is a double infinity of consumers and goods. We have already seen that the competitive economy with a finite number of goods and consumers is Pareto-efficient. If the number of consumers is infinite but the number of goods finite, we have the idealized competitive model with each consumer being insignificant relative to the market, and efficiency again holds. Finally, with a finite number of consumers but an infinite number of goods, the economy is again efficient.
22.5 Testing Efficiency

The Golden Rule, and the characterization of dynamic efficiency, provide conditions that are very simple to evaluate. Before this can be done credibly, there is an important issue concerning the assumptions describing the economy that needs to be addressed. The critical assumption is that of a constant growth rate in the population. The importance of constant growth lies in the fact that the Golden Rule is determined by the equality of the interest rate to the growth rate of population. If the growth rate is not constant, then this simple condition cannot be used. To provide a general means of testing efficiency, an extension must be made to the analysis.

A more general condition can be motivated as follows: In the economy we have described, the growth rate of capital is equal to the growth rate of population in the steady state. Observing this, the new investment in each period is \( nK \). The total payments to the owners of capital are \( rK \). The difference between these, \( rK - nK \), measures the total flows out of the firm—which we can call dividend payments. The economy is dynamically efficient if \( r \geq n \), which implies that dividend payments are positive so that funds are flowing out of the firm to the consumption sector. Conversely, the economy is dynamically inefficient if \( r < n \), so funds are flowing to the firm. The logic of looking at the flows in or out of firms provides a more general method of testing efficiency than comparing the interest rate to the population growth rate, since it holds under much less restrictive assumptions.

The general version of the test is to look at the difference between gross profit (the generalization of \( rK \)) and investment (the generalization of \( nK \)). The value of this difference, as a proportion of GDP, for a selection of countries is presented in table 22.1. All the values in the table are positive, which is clear evidence that the countries are dynamically efficient. However, given the high values reported in the table, these countries remain at some distance from achieving the Golden Rule.

<table>
<thead>
<tr>
<th>Year</th>
<th>England</th>
<th>France</th>
<th>Germany</th>
<th>Italy</th>
<th>Japan</th>
<th>United States</th>
</tr>
</thead>
<tbody>
<tr>
<td>1965</td>
<td>9.4</td>
<td>13.6</td>
<td>8.5</td>
<td>22.9</td>
<td>15.2</td>
<td>6.9</td>
</tr>
<tr>
<td>1970</td>
<td>7.5</td>
<td>11.8</td>
<td>7.8</td>
<td>18.9</td>
<td>11.6</td>
<td>5.6</td>
</tr>
<tr>
<td>1975</td>
<td>6.0</td>
<td>10.9</td>
<td>12.4</td>
<td>16.6</td>
<td>6.8</td>
<td>14.4</td>
</tr>
<tr>
<td>1980</td>
<td>10.1</td>
<td>8.3</td>
<td>8.4</td>
<td>12.9</td>
<td>7.5</td>
<td>10.2</td>
</tr>
<tr>
<td>1984</td>
<td>13.9</td>
<td>12.9</td>
<td>13.8</td>
<td>17.3</td>
<td>9.4</td>
<td>6.7</td>
</tr>
</tbody>
</table>

Source: Abel et al. (1989).
22.6 Conclusions

The overlapping generations economy provides a very flexible representation of how an economy evolves through time. It captures the natural features that consumers’ lives are short relative to the life span of the economy and that consumers allocate consumption in a rational way over their life cycle. The concept of the steady state also gives a description of equilibrium that is simple to apply.

The most interesting feature of the economy is that its lack of an ending means that there is a double infinity of goods and consumers. The double infinity is responsible for creating a potential inefficiency of the competitive equilibrium. Note the complete contrast to the static model. The chapter has characterized both efficient and optimal steady-state equilibria. The characterization of efficiency produces a simple test of dynamic efficiency that, as the evidence suggests, is met by a range of economies.

Further Reading

The classic paper that introduced the overlapping generations economy is:

It should be noted that the focus of Samuelson’s paper is on providing an intertemporal model that determines the interest rate endogenously. The inefficiency result, which has generated an immense literature since, is an almost unintentional by-product of this paper.

The model used here was introduced in:

Two interesting discussions (but note the first is very technical) of the inefficiency result are:

For a study of the role of money in overlapping generations, see:

The empirical analysis of efficiency is taken from:
Chapter 22: Intertemporal Efficiency

Exercises

22.1 “The interest rate is just the rental price of capital. Therefore competition will ensure a price that leads to economic efficiency.” True or false?

22.2 A consumer has preferences described by $U = \log(x_1) + \delta \log(x_2)$, where $x_t$ denotes consumption in period $t$ and $0 < \delta < 1$. Assume that the price of consumption is 1 in both periods and that the interest rate is $r$. If the consumer has income $M > 0$ in period 1 and no income on period 2, find her optimal level of saving and consumption plan. How is saving affected by changes in the interest rate and the discount factor $\delta$? Explain your results.

22.3 If a firm’s production function is $Y = K^{\alpha} L^{1-\alpha}$, show that it can earn at most zero profit. Use this production function to express the output–labor ratio as a function of the capital–labor ratio. Discuss the properties of the function derived.

22.4 Why might the purchase of capital (instead of the rental of capital) affect a firm’s profit-maximization decision?

22.5 The production technology of an economy is given by $y_t = k_t^{\alpha}$, with $0 < \alpha < 1$.

a. Verify that the equilibrium wage rate must be $w_t = [1 - \alpha]k_t^{\alpha}$ and the interest rate $r_t = \alpha k_t^{\alpha-1}$.

b. Show that capital market equilibrium in a steady state requires $x_1 = [1 - \alpha]k^{\alpha} - [1 + n]k$.

c. Hence use the consumer’s budget constraint to show that $x_2 = [1 + n]k[1 + \alpha k^{\alpha-1}]$.

d. Use these relationships to calculate the consumption possibility frontier. Sketch this frontier and locate the Golden Rule capital–labor ratio.

22.6 If a consumer has preferences $U = x_1 x_2$ over the steady-state levels of consumption in the two periods of life, show that the utility-maximizing choices satisfy $\frac{x_2}{x_1} = [1 + r]$. Use this result to calculate the capital–labor ratio in the steady-state equilibrium given the consumption possibility frontier from exercise 22.5. What is the effect on the capital–labor ratio of an increase in $n$ and $\alpha$? Explain.

22.7 Let each consumer have preferences described by the utility function

$$U(x^1, x^2) = \beta \log(x^1) + [1 - \beta] \log(x^2),$$

and let the production function be given by

$$y = k^{\alpha}.$$

a. Demonstrate that utility maximization results in demands that satisfy

$$\frac{x^2}{x^1} = [1 + r] \frac{1 - \beta}{\beta}.$$

b. Using the result in part a, the consumer’s budget constraint and the capital market equilibrium condition, show that the steady-state value of $k$ satisfies

$$\frac{1 - \beta}{1 - \alpha} = [1 + n]k^{1-\alpha}.$$
c. Employing the factor price condition \( r = \alpha k^{\alpha - 1} \), show that the steady-state interest rate is

\[
r = \frac{\alpha (1 + n)}{[1 - \beta (1 - \alpha)]}.
\]

22.8 Governments frequently manipulate the interest rate as part of economic policy. Is this a method for ensuring that the Golden Rule is achieved?

22.9 Economic inefficiency arises through market failure. What is the market failure in the overlapping generations economy?

22.10 A possible explanation for the inefficiency might be that the consumers are not all alive at the same time and therefore some mutually advantageous trades cannot occur. Consider an economy where consumer \( t \) receives an endowment of one unit of the single consumption good at time \( t \) and obtains utility only from consumption at times \( t \) and \( t + 1 \). All consumers meet at time 0 to trade. What is the equilibrium? Is efficiency restored?

22.11 Consider an economy with one consumer in each generation. Each consumer has an endowment of one unit of the consumption good when young and nothing when old. There is no production and the consumption good cannot be stored.

a. What are the consumption levels in the two periods of life if there is no trade? If preferences are given by \( U = \max\{x_t, x_{t+1}\} \), what level of utility is achieved?

b. At any point in time, what are the feasible consumption allocations between the young and the old consumers who are alive at that point? Given the preferences, which allocation is optimal?

c. Can the optimal allocation be reached by trade?

d. Is the inefficiency in the production model a consequence of the existence of capital?

22.12 Assume instead that the economy in exercise 22.11 lasts only for two periods. In the first period, there is only a young consumer. In the second period, there is one old consumer and a new young consumer. At the end of the second period, the economy terminates.

a. What is the equilibrium allocation for this economy?

b. Is it efficient?

c. What is the fundamental difference between this economy and that of exercise 22.11?

22.13 Assume that a consumer born in \( t \) has preferences represented by \( U = x_t^\alpha x_{t+1}^{\alpha - 1} \) and that the production technology is described by \( y_t = k_t^\alpha \).

a. Show that \( x_t^\alpha = \frac{y_t}{\alpha} \) and \( s_t = \frac{y_t}{\alpha} \). Hence demonstrate that the dynamics of the capital stock are given by \( \frac{1 - \alpha}{2} k_t^\alpha = (1 + n)k_{t+1} \).

b. Setting \( \alpha = 0.5 \), \( n = 0.05 \), and \( k_0 = 0.01 \), calculate the values of \( k_1, \ldots, k_{20} \). How quickly is the steady state reached? How is the level of the steady state \( k \) affected by an increase in \( n \)?

22.14 Obtain data on population growth and the real interest rate. Do the data indicate dynamic efficiency or inefficiency?

22.15 Consider an economy where there is one consumer born at the start of each time period. Each consumer lives for two periods and receives an endowment of one unit of the consumption
good when young. At the start of the economy there is a consumer who is already old. This consumer owns one unit of money but has no endowment of the consumption good. Money has no intrinsic value.

a. Can money be valuable in a finite economy (one that has a known end point)?
b. Can money be valuable in an infinite economy?
c. Can money allow efficiency to be attained?
23 Social Security

23.1 Introduction

A typical social security system provides income during periods of unemployment, poor health or disability, and financial support, in the form of pensions, to the retired. Although the generosity of systems varies among countries, these elements are present in all developed economies. The focus of this chapter is the economic implications of financial assistance to the retired. The overlapping generations economy proves to be ideal for this purpose.

In economic terms, the analysis of the part of the social security system that provides assistance during unemployment or poor health is concerned with issues of uncertainty and insurance. Specifically, unemployment and poor health can be viewed as events that are fundamentally uncertain, and the provision of social security is insurance cover against bad outcomes. In contrast, retirement is an inevitable outcome, or at least an option, once the retirement age has been reached. Insurance is therefore not the main issue (except for the problem of living for longer than accumulated wealth can finance). Instead, the issues that are raised with pensions are the potential transfers of resources between generations and the effect on savings behavior in the economy. Both issues require a treatment that is set within an explicitly intertemporal framework.

The pensions systems in many developed economies are coming under pressure in a process that has become known as the “pensions crisis.” The roots of this crisis can be found in the design of the systems and the process of change in population structure. The potential extent of this crisis provides strong ground for holding the view that reform of the pension system is currently one of the most pressing economic policy challenges.

After describing alternative forms of pension systems, the nature of the pensions crisis is described. This introduces the concept of the dependency ratio and how this ratio links pensions and pension contributions. The economic analysis of social security begins with a study of its effect on the equilibrium of the economy. Chapter 22 introduced the overlapping generations economy and showed how its competitive equilibrium may be inefficient. The potential for inefficiency opens up the possibility of efficiency-enhancing policy interventions. From this perspective we consider whether social security can be used to secure a gain in efficiency. The fact that a social security program may enhance efficiency can be understood from the effect of social security
on the level of the capital stock. If a social security program has the form of forced saving, so that consumers are provided with greater second-period income than they would naturally choose, then the program will raise the capital stock through the increased savings it generates. This will be beneficial in an undercapitalized economy. Conversely, if the program simply transfers earnings from those who are working to those who are retired, savings will fall and hence the level of capital. These observations motivate the search for a social security program that can guide the economy to the Golden Rule.

The fall in the birth rate is one of the causes of the pensions crisis. It is an interesting question to consider how a change in the birth rate affects the level of welfare at the steady state of an overlapping generations economy. We pursue this issue by considering how the birth rate affects the structure of the consumption possibility frontier, both in the absence and in the presence of a social security program. Social security may be beneficial for the economy, but there are issues of political economy connected with the continuation of a program. The introduction of a program with the structure observed in practice results in a transfer of resources toward the first generation of retired (they receive but do not contribute) and away from some of the generations that follow. This raises the question of how such a program is ever sustained, since each generation has an incentive to receive but not to contribute. The final analytical issue is to review the concept of Ricardian equivalence and its implications for social security. Ricardian equivalence is the observation that by changing their behavior, consumers are able to offset the actions of the government. We show the consequences this can have for social security and address the limitations of the argument. Finally, after having completed the analytical material, we return to address some of the proposals that have been made for the reform of social security programs.

23.2 Types of System

One defining characteristic of a social security system is whether pensions are paid from an accumulated fund or from current tax contributions. This feature forms the distinction between fully funded and pay-as-you-go social security systems. The economic effects, both in terms of efficiency and distribution, between these two polar forms of system are markedly different.

In a pay-as-you-go social security program the current contributions through taxation of those in employment provide the pensions of those who are retired. At any point in time the contributions to the system must match the pension payments
made by the system. The social security systems presently in operation in the United States, the United Kingdom, and numerous other countries are broadly of this form. The qualifier “broadly” is used because, for example, although the US system owns some assets and could afford a short-term deficit, the assets would fund only a very short period of payments. At each point in time a pay-as-you-go system satisfies the equality

\[
\text{Benefits received by retired} = \text{Contributions of workers.} \tag{23.1}
\]

This equality can be expressed in terms of the number of workers and pensioners by

\[
\beta R = \tau E, \tag{23.2}
\]

where \(\tau\) is the average social security contribution of each worker, \(\beta\) is the average pension received, \(E\) the number of workers in employment, and \(R\) the number of retired. If there is a constant rate of growth of population, so that the workforce is a constant multiple of the retired population, then \(E = (1 + n)R\). Using this in (23.2) yields

\[
\beta R = \tau (1 + n) R, \text{ or } \beta = (1 + n) \tau. \tag{23.3}
\]

This relationship implies that the tax paid when young earns interest at rate \(n\) before being returned as a pension when old. Hence in a pay-as-you-go pension system the return on contributions is determined by the growth rate of population.

In a fully funded system each worker makes contributions toward social security via the social security tax, and the contributions are invested by the social security program. The program therefore builds up a pension fund for each worker. The total pension benefits received by the worker when retired are then equal to their contribution to the program plus the return received on the investment. Such a program satisfies the equalities

\[
Pensions = \text{Social security tax plus interest} = \text{Investment plus return.} \tag{23.4}
\]

The implication of this constraint is that the fund earns interest at rate \(r\), so the pension and the tax are related by

\[
\beta = (1 + r) \tau. \tag{23.5}
\]

A fully funded social security system forces each worker to save an amount at least equal to the tax they pay. It remains possible for workers to save more if they choose to do so. If, in the absence of social security, all workers chose to save an amount in excess of the taxed levied by the program, then, holding all else constant, a fully
funded system will simply replace some of the private saving by an equivalent amount of public saving. In this case a fully funded system will have no effect on the equilibrium outcome. We explore this observation further when we discuss Ricardian equivalence in section 23.8. In more general settings with a variety of investment opportunities, the possibility must be considered that the rate of return on private savings may differ from that on public savings. When it does, a fully funded system may affect the equilibrium. This point arises again in the analysis of pension reform.

Contrasting these two forms of system, it can be observed that a pay-as-you-go system leads to an intergenerational transfer of resources, from current workers to current retired, whereas a fully funded system can at most cause an intertemporal reallocation for each generation. This observation suggests that the two systems will have rather different welfare implications; these will be investigated in the following sections. In addition the pay-as-you-go system has a return of \( n \) on contributions and the fully funded system has a return of \( r \). These returns will differ unless the economy is at the Golden Rule allocation.

Systems that fall between these two extremes are termed non–fully funded. Such systems make some investments, but the payments made in any given period may be greater than or less than the revenue, composed of tax payments plus return on investment, received in that period. The difference between payments and revenue will comprise investment, or disinvestment, in the pension fund.

### 23.3 The Pensions Crisis

Many countries face a pensions crisis that will require that their pensions systems be significantly reformed. This section identifies the nature and consequences of this crisis. Once the analysis of social security is completed, we return in section 23.9 to review a range of proposals for reform of the system in the light of this crisis.

The basis of the pensions crisis is threefold. First, the birth rate has fallen in most developed economies. Although immigration has partially offset the effect of this in some countries, there has still been a net effect of a steady reduction in the addition of new workers. The second effect is that longevity is increasing since people are, on average, living longer. For any given retirement age, this is increasing the number of retired. Third, there is also a tendency for the retirement age to fall.

The net effect of these three factors is that the proportion of retired in the population is growing, and it is this increase that is problematic. In general terms, as the proportion of the population that is retired rises, the output of each worker must support an ever
larger number of people. Output per capita must rise just to keep consumption per capita constant. If output does not rise quickly enough, then productivity gains will be diluted and output per capita will fall. Furthermore, supporting the retired at a given standard of living will impose an increasing burden on the economy.

The size of this effect can be seen by looking at forecasts for the dependency ratio. The dependency ratio measures the relative size of the retired population and is defined as the size of the retired population relative to the size of the working population. Table 23.1 reports the dependency ratio for a range of countries over the recent past and forecasts for its development into the future. The countries in the table are typical with the dependency ratio forecast to increase substantially—in all cases the ratio more than doubles from 1980 to 2040. This means that those working have to support an increasing proportion of retired. In some cases, Japan for instance, the forecast increase in the dependency ratio is dramatic.

The consequence of the increase in the dependency ratio can be expressed in more precise terms by looking at the relationship between the contributions to pay for social security and the resulting level of social security. Using the identity (23.2) for a pay-as-you-go system and dividing through by $E$, we can write the relationship between social security tax, pension, and dependency ratio as

$$\tau = \beta D,$$

(23.6)

where $D$ is the dependency ratio, $\frac{R}{E}$. Hence as $D$ rises, $\tau$ must increase if the level of the pension $\beta$ is to be maintained. Alternatively, the pension decreases as $D$ increases if the tax rate is held constant. If some combination of such changes is not made, then the social security system will go into deficit if the dependency ratio increases. Neither a deficit, a falling pension, or an increasing tax is not an attractive option for governments to present to their electors.
These factors can be seen at work in forecasts for the future path of the US social security program as predicted by the Board of Trustees of the Federal Old-Age and Survivors Insurance and Disability Insurance Trust Funds (OASDI). Figure 23.1 shows the forecast deficit for the US Old-Age and Survivors Insurance Fund (but does not include the Disability Insurance Fund). The income rate is defined as the ratio of income from payroll tax contributions to the OASDI taxable payroll (effectively the average tax rate for social security contributions), and the cost rate is the ratio of the cost of the program to the taxable payroll. The projections are based on the structure of the social security program remaining much as it is today (in terms of the rate of tax and the value of benefits). As the figure shows, the fund is forecast to go into deficit in 2018 and remain in deficit unless some significant reform is undertaken.

To avoid such deficits, what these facts imply is that governments face a choice between maintaining the value of pension payments, but with an ever-increasing tax rate, or allowing the value of pensions to erode so as to keep the tax rate broadly constant. For example, the UK government has reacted to this situation by allowing the real value of the state pension to steadily erode. As shown in table 23.2, the value of the pension has fallen from almost 40 percent of average earnings in 1975 to 26 percent in 2000, and it is expected to continue to fall, especially since the pension is now indexed to prices rather than earnings. These reductions have taken the value of the pension well below the subsistence level of income. Consequently pensioners with no other source of income receive supplementary state benefits to take them to the subsistence level. This reduction in the state pension has been accompanied by government encouragement of the use of private pensions. We return to this in the discussion of reforms in section 23.9.


Table 23.2
Forecasts for UK basic state pension

<table>
<thead>
<tr>
<th>Date</th>
<th>Rate as percentage of average earnings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1975</td>
<td>39.3</td>
</tr>
<tr>
<td>1980</td>
<td>39.4</td>
</tr>
<tr>
<td>1985</td>
<td>35.8</td>
</tr>
<tr>
<td>1990</td>
<td>29.1</td>
</tr>
<tr>
<td>1995</td>
<td>28.3</td>
</tr>
<tr>
<td>2000</td>
<td>25.7</td>
</tr>
</tbody>
</table>


In conclusion, the basis of the pensions crisis has been identified, and it has been shown how this impacts on the state pensions that will be paid in the future. The depth of this crisis shows why social security reform is such an important policy issue. The chapter now proceeds to look at the economic effects of social security as a basis for understanding more about the arguments behind the alternative reforms that have been proposed.

23.4 The Simplest Program

Having set out the issues connected with social security programs, the focus is now placed on their economic effects. The fundamental insight into the effect of social security on the economy can be obtained using the simple model of section 22.4.2. In this economy there is no production but only the exchange of endowments. Although simple, this economy is still capable of supporting a role for social security.

In the economy under analysis, each consumer receives an endowment of one unit of the single consumption good in the first period of their life but receives no endowment in the second period. To simplify, the population is assumed to be constant. As already noted in chapter 22 the equilibrium of this economy without any government intervention has the endowment entirely consumed when young so that there is no consumption when old. This has to be the equilibrium, since the old have nothing to offer the young in trade. This autarkic equilibrium is not Pareto-efficient, since all consumers would prefer a more even distribution of consumption over the two periods of life.

How can a social security program improve on the autarkic equilibrium? Consider a pay-as-you-go program that taxes each young consumer half a unit of consumption...
and transfers this to an old consumer. The lifetime consumption plan for every consumer then changes from the autarkic equilibrium consumption plan of \( \{1, 0\} \) to the new consumption plan of \( \{\frac{1}{2}, \frac{1}{2}\} \). Provided that the preferences of the consumers are convex, the new allocation is preferred to the original allocation. Since this applies to all generations, the social security system has achieved a Pareto improvement. This argument is illustrated in figure 23.2. The Pareto improvement from the social security system is represented by the move from the lowest indifference curve to the central indifference curve.

In fact a far stronger conclusion can be obtained than just the ability of social security to achieve a Pareto improvement. To see this, note that the assumption of a constant population means that the per capita consumption possibilities for the economy lie on the line joining \( \{1, 0\} \) to \( \{0, 1\} \). In the same way that the Golden Rule was defined for the economy with production, the Golden Rule allocation can be defined for this economy as that which maximizes utility subject to the first- and second-period consumption levels summing to 1. Denote this allocation by \( \{x^{1*}, x^{2*}\} \). The Golden Rule allocation can then be achieved by a pay-as-you-go social security program that transfers \( x^{2*} \) units of the consumption good from the young consumer to the old consumer.

These arguments show how social security can achieve a Pareto improvement and, for the simple exchange economy described, even achieve the Golden Rule allocation. The social security program is effective because of the intergenerational transfer that it
engineers and the consequent revision in the consumption plans. The optimality result was built upon the use of a pay-as-you-go program. In contrast, a fully funded program cannot be employed, since there is no commodity that can be used as an investment vehicle. The form in which these conclusions extend to the more general overlapping generations economy with production is now discussed.

23.5 Social Security and Production

It has already been shown how social security can obtain a Pareto improvement in an overlapping generations economy with no production. When there is production, a wider range of effects can arise, since social security affects the level of savings and hence capital accumulation. These additional features have to be accounted for in the analysis of social security.

The concept of the Golden Rule and its associated capital–labor ratio was introduced in chapter 22. This showed that the optimal capital stock is the level that equates the rate of interest to the rate of population growth. If the capital stock is larger than this, the economy is dynamically inefficient and a Pareto improvement can be made by reducing it. When it is smaller, the economy is dynamically efficient, so no Pareto improvement can be made, but the economy is not in an optimal position. These observations then raise the questions: How does social security affect capital accumulation? Can it be used to move a nonoptimal economy closer to the Golden Rule?

To answer these questions, consider a social security program that taxes each worker an amount \( \tau \) and pays each retired person a pension \( \beta \). The program also owns a quantity \( K^s_t \) of capital at time \( t \). Equivalently it can be said to own \( k^s_t \), \( k^s_t = \frac{K^s_t}{L_t} \), of capital per unit of labor. A social security program will be optimal if the combination of \( \tau \), \( \beta \), and \( k^s_t \) is feasible for the program and ensures that the economy achieves the Golden Rule.

A feasible social security program must satisfy the budget identity

\[
\beta L_{t-1} = \tau L_t + r_t k^s_t L_t - \left[ k^s_{t+1} L_{t+1} - k^s_t L_t \right],
\]

which states that pension payments must be equal to tax revenue plus the return on capital holdings less investment in new capital. Since the population grows at rate \( n \), in a steady state the identities \( L_{t-1} = \frac{L_t}{1+n} \), \( L_{t+1} = [1+n]L_t \), and \( k^s_{t+1} = k^s_t \equiv k^s \) can be used in (23.7) to generate the steady-state budget identity

\[
\frac{\beta}{1+n} = \tau + [r - n] k^s.
\]

(23.8)
Noting that the pension, $\beta$, which is received in the second period of life, is discounted in a consumer’s budget constraint (since $x^1 + s = w - \tau$ and $[1 + r]s + \beta = x^2$, it follows that $s = \frac{x^2 - \beta}{1 + r}$), the budget constraint under the program can be written as

$$x^1 + \frac{x^2}{1 + r} = w - \tau + \frac{\beta}{1 + r},$$

(23.9)

The condition describing consumer choice remains

$$\frac{U_1(x^1, x^2)}{U_2(x^1, x^2)} = 1 + r.$$  

(23.10)

Equilibrium on the capital market requires that private savings are equal to total capital less the capital owned by the social security program. This condition can be expressed as

$$w - x^1 - \tau = [1 + n] [k - k^s].$$

(23.11)

The choices of the representative firm do not change, so the conditions relating factor prices to capital still apply with

$$f'(k) = r,$$

(23.12)

$$f(k) - kf'(k) = w.$$  

(23.13)

The steady-state equilibrium with the pension program is the solution to equations (23.8) to (23.13).

The aim now is to investigate the effect that the social security policy can have on the equilibrium. To see why it may be possible to design a program that can achieve the Golden Rule, it should be noted that the failure of the competitive equilibrium without intervention to achieve efficiency results from the savings behavior of individuals leading to over- or underaccumulation of capital. With the correct choice of social security program the government can effectively force-save for individuals. This alters the steady-state level of the capital stock and hence the growth path of output.

In equations (23.8) to (23.13) there are five private-sector choice variables ($k, x^1, x^2, w,$ and $r$) that are treated as endogenous, plus the three variables ($\beta, \tau,$ and $k^s$) that describe the social security program. Given that there are six equilibrium conditions, the pension system can choose any two of the variables describing the program with the third determined alongside the endogenous variables. To analyze the system, it is simplest to treat $\beta$ as endogenous and $\tau$ and $k^s$ as exogenous.

The method of analysis is to assume that the Golden Rule is achieved and then to work back to the implications of this assumption. Consequently let $r = n$. From the
firm’s choice of capital, the Golden Rule is consistent with a capital stock that solves
\( f'(k^*) = n \) and hence a wage rate that satisfies \( w = f(k^*) - k^* f'(k^*) \). The important
observation is that with \( r = n \), the budget constraint for the social security program
collapses to
\[
\frac{\beta}{1 + n} = \tau + [r - n] k^* = \tau,
\]
so a program attaining the Golden Rule must have the form of a pay-as-you-go system
with \( \beta = [1 + n] \tau \). It is important to observe that any value of \( k^* \) is consistent with
(23.14) when \( r = n \), including positive values. This observation seems to conflict with
the definition of a pay-as-you-go program. These comments are rationalized by the fact
that we are working with the steady state of the economy. The social security program
may own a stock of capital, \( k^* > 0 \), but in operating the pay-as-you-go-system, it does
not add to or subtract from this level of capital. Instead, the return on the capital it owns is
just sufficient to maintain it at a constant level. It remains true that along any growth path,
including the steady state, a pay-as-you-go system cannot increase its capital holdings.
The values of the tax and capital stock of the program required to support the Golden
Rule can now be found by using the fact that the program is pay-as-you-go to reduce
the consumer’s budget constraint to
\[
x^1 + \frac{x^2}{1 + r} = w.
\]
Combining this constraint with the condition describing consumer choice indicates that
the demand for first-period consumption must depend only on the wage rate and the
interest rate, so \( x^1 = x^1(w, r) \). Using the conditions for the choice of the firm, we
have that the wage rate and interest rate depend on the level of capital, so demand for
first-period consumption can be written as
\[
x^1 = x^1(w, r) = x^1(f(k) - k f'(k), f'(k)) = x^1(k).
\]
The capital market-clearing condition can then be written as
\[
w - x^1(k) - \tau = [1 + n] [k - k^*].
\]
Using the conditions for the choice of the firm and evaluating at the Golden Rule level
generates
\[
\tau = \left[ f(k^*) - k^* f'(k^*) - x^1(k^*) - [1 + n] k^* \right] + [1 + n] k^*.
\]
Condition (23.18) determines pairs of values \( \{\tau, k^s\} \) that will achieve the Golden Rule. Since \( k^s \) is fixed the relationship between \( \tau \) and \( k^s \) can be written more compactly as

\[
\tau = a + bk^s. \tag{23.19}
\]

Any pair \( \{\tau, k^s\} \) that satisfies (23.19) will generate the Golden Rule provided that the capital stock held by the program is not negative. For instance, if the program holds no capital, so that \( k^s = 0 \), then the value of the social security tax will be

\[
\tau = a. \tag{23.20}
\]

Although the discussion to this point has implicitly been based on the tax, \( \tau \), being positive, it is possible that the optimal program may require it to be negative. If \( k^s = 0 \), it follows from (23.20) that \( \tau < 0 \) when \( a < 0 \). When \( \tau \) is negative, the social security program involves a transfer from the old to the young (rather than a pension paid to the old from the output of the young). As an example, if \( x^1 (w, r) = \frac{w}{2} \) and \( f (k) = k^\alpha \), then \( k^s = \left(\frac{\alpha}{n}\right)^{1/(1-\alpha)} \) (see exercise 23.4 for the details of this derivation). Substituting these values into (23.20) gives

\[
\tau = \left(\frac{\alpha}{n}\right)^{1/(1-\alpha)} \left[ \frac{1 - \alpha}{2\alpha} n \right] - [1 + n]. \tag{23.21}
\]

If the rate of population growth is 5 percent, then the tax will be negative whenever

\[
\frac{1}{43} < \alpha. \tag{23.22}
\]

For this example the tax rate is positive only for very small values of \( \alpha \).

The results have shown that attainment of the Golden Rule requires a pay-as-you-go social security system. By implication, a fully funded program will fail to attain the Golden Rule. In fact an even stronger result can be shown: a fully funded program will have no effect on the equilibrium. To demonstrate this result, observe that a fully funded program must satisfy the identity that the value of pension paid must equal the value of tax contributions plus interest, or

\[
\beta L_{t-1} = \tau L_{t-1} [1 + r_t] = k^s L_t [1 + r_t]. \tag{23.23}
\]

Evaluated at a steady state,

\[
\beta = \tau [1 + r] = k^s [1 + n] [1 + r]. \tag{23.24}
\]

The substitution of (23.24) into the equilibrium conditions (23.8) to (23.13) shows that they reduce to the original market equilibrium conditions described in (22.18) to
The fully funded system therefore replaces private saving by public saving and does not affect the consumption choices of individual consumers. It has no real effect on the equilibrium, and if the initial steady state were not at the Golden Rule, the fully funded social security program would not restore efficiency. This observation is discussed further in section 23.8.

This analysis has demonstrated how a correctly designed social security program can generate the Golden Rule equilibrium, provided that it is not of the fully funded kind. A fully funded system does not affect the growth path. In contrast, a pay-as-you-go system can affect the aggregate levels of savings and hence the steady-state capital–labor ratio. This allows it to achieve the Golden Rule.

23.6 Population Growth

The fall in the rate of population growth is an important factor in the pensions crisis. While operating a simple pay-as-you-go program, a decreasing population size makes it harder to sustain any given level of pension. Observing this fact raises the general question of how the level of welfare is related to the rate of population growth. This section addresses this issue both with and without a social security program.

Assume first that there is no social security program in operation. Recall that the consumption possibility frontier is defined by a pair of consumption levels \(x_1\) and \(x_2\) that satisfies the conditions

\[
x_1 = f(k) - k f'(k) - [1 + n] k \tag{23.25}
\]

and

\[
x_2 = [1 + n] k [1 + f'(k)]. \tag{23.26}
\]

The effect of a change in the population growth rate can be determined by calculating how it modifies this consumption possibility frontier. For a given value of \(k\), it follows that \(\frac{\partial x_1}{\partial n} = -k\) and \(\frac{\partial x_2}{\partial n} = k [1 + f'(k)].\) Consequently, holding \(k\) fixed, an increase in the growth rate of population reduces the level of first-period consumption but raises the second-period level. This moves each point on the consumption possibility frontier inward and upward. Furthermore, when evaluated at the Golden Rule capital–labor ratio, these changes in the consumption levels satisfy

\[
\frac{\partial x_2}{\partial x_1} \frac{\partial x_2}{\partial n} = - \left[ 1 + f'(k^*) \right] = - [1 + n]. \tag{23.27}
\]
Hence, for a small increase in \( n \), the point on the frontier corresponding to the Golden Rule equilibrium must shift upward along a line with gradient \( -[1 + n] \). The consequence of these calculations is that the shift of the consumption possibility must be as illustrated in figure 23.3.

How the level of welfare generated by the economy is affected by an increase in \( n \) then depends on whether the initial equilibrium level of capital is above or below the Golden Rule level. If it is below, then welfare is reduced by an increase in the population growth rate—the capital stock moves further from the Golden Rule level. The converse occurs if the initial equilibrium is above the Golden Rule. This is illustrated in figure 23.4 where the initial equilibrium is at \( e^0 \) with a capital–labor ratio below the Golden Rule. The equilibrium moves to \( e \) following an increase in \( n \). It can also be seen in the figure that if the initial equilibrium had been at a point on the frontier above the Golden Rule, then the upward shift in the frontier would imply that the new equilibrium moves onto a higher indifference curve.

Now introduce a social security system and assume that this is adjusted as population growth changes to ensure that the Golden Rule is satisfied for all values of \( n \). For a small change in \( n \), the Golden Rule allocation moves along the line with gradient \( -[1 + n] \), as noted above. However, for large increases in \( n \), the gradient of this line becomes steeper. This moves the Golden Rule equilibrium as shown in figure 23.5 to a point below the original tangent line. As a consequence the increase in population growth must reduce the per capita level of consumption \( x^1 + \frac{x^2}{1+n} \). Therefore, even with an optimal social
security scheme in operation, an increase in population growth will reduce per capita consumption.

The effects of changes in the rate of population growth are not as clear as the simple equilibrium identity for a pay-as-you-go program suggests. As well as the mechanics of the dependency ratio, a change in population growth also affects the shape of the consumption possibility frontier. How welfare changes depends on whether a social
security program is in operation and on the location of the initial equilibrium relative to the Golden Rule. If an optimal program is in operation, then an increase in population growth must necessarily reduce the level of per capita consumption.

### 23.7 Sustaining a Program

In the simple economy without production, a social security program involving the transfer of resources between generations achieves a Pareto improvement. This raises the obvious question of why such a program will not always be introduced.

The basic nature of the pay-as-you go pension program described above is that the young make a transfer to the old without receiving anything directly from those old in return. Instead, they must wait until their own old age before receiving the compensating payment. Although these transfers do give rise to a Pareto improvement, it can be argued that it is not in the young consumer’s private interest to make the transfer provided that they expect to receive a transfer. (Think of the generations playing a game. Giving a transfer cannot be a Nash equilibrium strategy.) If the young consumers do not give their transfer but still expect to receive their pensions, then their consumption level will be increased. Clearly, this makes them better off, so they will not wish to make the transfer. Since the social security system is not individually rational, how can the young be persuaded to consent to the imposition of the social security program?

Two different answers to this question will be considered. The first answer is based on altruism on the part of the young—they are willing to provide the transfer because they care about the old. This rationalizes the existence of a social security program but only by making an assumption that moves outside the standard economic framework of individual self-interest. The second answer works with the standard neoclassical model of self-interest but shows how the program can be sustained by the use of “punishment strategies” in an intertemporal game. It should be stressed that the fact that participation in a social security program is mandatory is not by itself a valid explanation of the existence of the program. All programs have to have willing participants to initiate them (so they must be individually rational at their introduction) and need continuing support to sustain them.

Altruism refers to feelings of concern for others beside oneself. It is natural to think that altruism applies to close family members, but it may also apply to concern for people generally.

Although the existence of altruism takes us outside the standard perspective of behavior driven by narrow self-interest, it need not affect the tools we employ to analyze
behavior. What is meant by this is that altruism alters the nature of preferences but does not affect the fact that a consumer will want to achieve the highest level of preference possible. Consequently, given a set of altruistic preferences, the consumer will still choose the action that best satisfies those preferences subject to the constraints placed on their choices. The standard tools remain valid but operate on different preferences.

There are numerous ways to represent altruism, but one of the simplest is to view it as a consumption externality. Writing the utility of a consumer in generation \( t \) in the form

\[
U_t = U(x_t^t, x_t^{t+1}, x_{t-1}^t),
\]

(23.28)
gives an interpretation of altruism as concern for the consumption level, \( x_{t-1}^t \), achieved by a member of the earlier generation (which is usually interpreted as the parent of the consumer). A very similar alternative would be to assume that

\[
U_t = U(x_t^t, x_t^{t+1}, U_{t-1}),
\]

(23.29)
so that altruism is reflected in a concern for the utility of the member of the earlier generation.

Both of these forms of altruism provide a motive for a social security program that transfers resources from the young to the old. Consider (23.28). A consumer with this utility function can be thought of as choosing her personal consumption levels \( x_t^t \), \( x_t^{t+1} \), and a transfer, \( \tau \), to the old consumer. The effect of the transfer is to raise the consumption level \( x_{t-1}^t \), since the budget constraint of the old consumer is

\[
x_{t-1}^t = [1 + r_t] s_{t-1} + \tau.
\]

(23.30)
Provided that the marginal utility generated by an increase in \( x_{t-1}^t \) is sufficiently high, the consumer will willingly choose to make a positive transfer. In this sense the provision of social security has become individually rational because of altruism.

The second reason why transfers may be sustained is now considered. A rational explanation for participating in a social security program can be found in the fact that each young person expects a similar transfer when he is old. Young persons can then be threatened with having this removed if they do not themselves act in the appropriate manner. This punishment can sometimes (but not always) be sufficient to ensure that compliance with the social security program is maintained.

To give substance to these observations, it is best to express the argument using the language of game theory. The analysis so far has shown that the strategy to provide a transfer is not a Nash equilibrium. Recall that in the determination of a Nash equilibrium
each individual holds the strategies of all others constant as he considers his own choice. So, if all others are providing transfers, it will be a better strategy not to do so but to still receive. If others are not transferring, then it is also best not to do so. Therefore not providing a transfer is a dominant strategy, and the individually rational Nash equilibrium must be for no transfers to take place.

These simple Nash strategies are not the only ones that can be played. To motivate what else can be done, it is best to think about repeated games and the more sophisticated strategies that can be played in them. A repeated game is one where the same “stage” game is played once each period for an endless number of periods by the same players. The prisoners’ dilemma, given in the matrix in figure 23.6, has the general features of the social security model. It is not exactly the same, since the social security model has many generations of consumers and not just the two given in the game.

If both players contribute to social security, then a payoff of 5 is attained. If neither contributes, the payoff is only 2. This reflects the fact that the social security equilibrium is Pareto-preferred to the equilibrium without. However, the highest payoff is obtained if a player chooses not to contribute but the other does. When played a single time, the unique Nash equilibrium is for both players to choose Don’t contribute—if the other contributes, then it pays not to. This reasoning applies to both players and hence the equilibrium. This equilibrium is inefficient and is Pareto-dominated by \{Contribute, Contribute\}.

The situation is completely changed if the game is repeated indefinitely. Doing so allows the efficient equilibrium \{Contribute, Contribute\} to be sustained. A strategy that supports this is for each player to choose Contribute until their opponent chooses Don’t contribute. Once this has happened, they should continue to play Don’t contribute from that point on.
To evaluate the payoffs from this strategy, assume that the discount factor between periods is $\delta$. The payoff from always playing Contribute is then

$$5 + 5\delta + 5\delta^2 + 5\delta^3 + \ldots = 5 \left[ \frac{1}{1 - \delta} \right]. \quad (23.31)$$

Alternatively, if Don’t contribute is played unilaterally a temporary gain will be obtained but the payoff will then fall back to that at the Nash equilibrium of the single-period game once the other player switches to Don’t contribute. This gives the payoff

$$10 + 2\delta + 2\delta^2 + 2\delta^3 + \ldots = 10 + 2 \left[ \frac{\delta}{1 - \delta} \right]. \quad (23.32)$$

Contrasting these, playing Contribute in every period will give a higher payoff if

$$5 \left[ \frac{1}{1 - \delta} \right] > 10 + 2 \left[ \frac{\delta}{1 - \delta} \right] \quad (23.33)$$

or

$$\delta > \frac{5}{8}. \quad (23.34)$$

That is, $\{\text{Contribute, Contribute}\}$ will be an equilibrium if the players are sufficiently patient. The reason behind this is that a patient player will put a high value on payoffs well into the future. Therefore the reduction to a payoff of 2 after the first period will be very painful. For a very impatient player, only the payoff of 10 will really matter and they are driven to Don’t contribute.

The strategy just described is known as a “punishment strategy”: the deviation from Contribute is punished by reversion to the inefficient Nash equilibrium. Although the punishment will hurt both players, the point is that it will not happen in equilibrium, since the optimal play with these strategies is always to choose Contribute when players are patient. In summary, in an infinitely repeated game, punishment strategies can be used to support efficient equilibria.

The same line of reasoning can be applied to the analysis of social security. What is different in this context is that the same players do not interact every period. Instead, it is a different pair of old and young consumers that meet in each period. However, the punishment strategy can still be employed in the following way: each consumer when young will provide a transfer of size $x$ to the old consumer that overlaps with them only if that old person alive at the same time provided a transfer in the previous period;
otherwise no transfer is provided. If all generations of consumers play according to this strategy, then the transfers can be made self-supporting.

There remains one important limitation to this use of punishment strategies in the social security environment. To implement the strategy, each young consumer must know whether the transfer was made in the period before they were alive. This issue does not arise in the standard application of punishment strategies, since the players are alive in all periods—they need only remember what happened in the previous period. Consequently some form of verification device is necessary to support the punishment strategy. Without the verification, the only equilibrium is for there to be no transfers, which is a Pareto-inferior outcome.

This discussion of pay-as-you-go social security has shown how such a system can be sustained even when there is a short-run incentive for consumers not to make the required transfers. The basis for this claim is that social security in an overlapping generations economy has the nature of a repeated game so that strategies that punish the failure to provide a transfer can be employed. What this analysis shows is that an apparent act of generosity—the gift of a transfer to the older generation—can be made to be rational for each individual. So the provision of social security may occur not through altruism but through rationality.

23.8 Ricardian Equivalence

Ricardian equivalence refers to the proposition that the government can alter an economic policy and yet the equilibrium of the economy can remain unchanged. This occurs if consumers can respond to the policy by making off-setting changes in their behavior that neutralize the effect of the policy change. In terms of the present chapter, Ricardian equivalence holds when the government introduces, or changes, a social security system and yet the changes in individual behavior render the policy change ineffectual.

Such equivalence results have already featured twice in the text. On the first occasion, in the analysis of the private purchase of public goods, it was shown that by changing their purchases, the individuals could offset the effect of income redistribution. Furthermore it was also rational for the individuals to make the off-setting changes. The second case of equivalence arose in the derivation of the optimal social security program where it was noted that a fully funded system would not affect the capital–labor ratio. The explanation for this equivalence was that consumers react to a fully funded social security program by making a reduction in their private savings that ensures that total savings is unchanged.
The common feature of these examples is that the effect of the policy change and the offsetting reaction involves the same individuals. So they have a direct incentive to modify their behavior. Clearly, this is characteristic only of a social security system that is fully funded with a return equal to that on private savings. If social security is anything but fully funded, a change in the system will affect a number of generations, since the system must be redistributive over time. In the case of pay-as-you-go, social security involves purely intertemporal redistribution. A change in a program can therefore affect consumers in different generations who need not be alive at the time the program is changed nor even be alive at the same time. At first sight, this would seem to mean that it cannot be possible for equivalence to hold. This argument is in fact correct given the assumptions made so far.

To obtain a basis for eliminating the effect of policy, it is necessary to link the generations across time so that something that affects one generation directly somehow affects all generations indirectly. The way that this can be done is to return to the idea of altruism and intergenerational concern. Intuitively we can think of each consumer as having familial forebears and descendents (or parents and children in simple language). This time we assume that each parent is concerned with the welfare of their children, and that their children are concerned with the welfare of the grandchildren. Indirectly, although they are not alive at the same time in the model, this makes the parents concerned about the grandchildren. What effect does this have? It makes each family act as if it was a dynasty stretching through time, and its decisions at any one moment take into account all later consequences. A change in a social security program then causes a reaction right through the decision process of the dynasty.

To provide some details, let the utility of the generation born at time \( t \) be

\[
U_t = U \left( x_t^t, x_{t+1}^t, \bar{U}_{t+1} \right). \tag{23.35}
\]

It is the term \( \bar{U}_{t+1} \) that represents the concern for the next generation. Here \( \bar{U}_{t+1} \) is defined as the maximum utility that will be obtained by the children, who are born at \( t + 1 \), of the parent born in \( t \). The fact that the family will act as a dynasty can then be seen by substituting for \( \bar{U}_{t+1} \) to give

\[
U_t = U \left( x_t^t, x_{t+1}^{t+1}, U \left( x_{t+1}^{t+1}, x_{t+1}^{t+2}, \bar{U}_{t+2} \right) \right). \tag{23.36}
\]

If this substitution is continually repeated, then the single parent born at \( t \) ultimately cares about consumption levels in all future time periods.

By this fact it is now possible to demonstrate that Ricardian equivalence applies to social security in these circumstances. Consider an initial position with no social...
security program and no population growth (so $n = 0$). The consumer at $t$ reflects his concern for the descendent by making a bequest of value $b^t$. Hence the consumption level in the second period of life is

$$x_{t+1}^{t+1} = s_t [1 + r_{t+1}] - b_t,$$  \hspace{1cm} (23.37)

and that of his descendent is

$$x_{t+1}^{t+1} = w_{t+1} + b_t - s_{t+1}.$$  \hspace{1cm} (23.38)

Assume that a social security program is now introduced and that each consumer has one descendent. Under the terms of the program, young consumers are taxed an amount $\tau$ to pay a pension of equal value to old consumers. Then the consumption level of each parent satisfies

$$x_{t+1}^{t+1} = s_t [1 + r_{t+1}] + \tau - \hat{b}_t,$$  \hspace{1cm} (23.39)

and that of his descendent

$$x_{t+1}^{t+1} = w_{t+1} + \hat{b}_t - \tau - s_{t+1}.$$  \hspace{1cm} (23.40)

But note that if the bequest is changed so that $\hat{b}_t = b_t + \tau$, the same consumption levels can be achieved for both the parent and the child as for the case with no pension. Furthermore, since these consumptions levels were the optimal choice initially, they will still be the optimal choice. The old consumers will make this change to their bequest, and the social security scheme will have no effect.

The conclusion of this analysis is that the change in the bequest can offset the intertemporal transfer caused by a social security system. Although this was only a two-period system, it can easily be seen that the same logic can be applied to any series of intergenerational transfers that the government might implement. All that the dynasty has to do is adjust each bequest to offset the effect of the government transfer between any two generations. This will always be possible if the initial bequests are sufficiently large. The outcome is that the policy has no effect. This is the basic point of Ricardian equivalence.

It must be noted that there are limitations to this argument. First, it is necessary that there be active intergenerational altruism. Without this altruism, there is no dynastic structure, and the offsetting changes in bequests will not occur. In addition the argument only works if the initial bequest is sufficiently large that it can be changed to offset the policy without becoming negative. Does it apply in practice? We clearly observe bequests, but many of these may be unintentional and occur due to premature death.
The concept of Ricardian equivalence can be extended into other areas of policy. Closely related to social security is the issue of government debt, which is also an intergenerational transfer (but from children to parents), and its effects on the economy. This was the initial area of application for Ricardian equivalence, with changes in bequests offsetting changes in government debt policy. Furthermore, if links are made across households, it becomes possible for changes in household choices to offset a policy that causes transfers between households. This has lead to the question of whether “everything is neutral.” The answer depends on the extent of the links.

23.9 Social Security Reform

The basic nature of the pensions crisis facing a range of economies was identified in section 23.3: increasing longevity and the decline in the birth rate are combining to increase the dependency ratio. Without major reform or an unacceptably high increase in tax rates, the pension programs will either go into deficit or pay a much reduced pension. A variety of reforms have been proposed in response to this crisis. Some of these are now briefly reviewed.

Underlying the crisis is the fact that the pension systems are essentially of the pay-as-you-go form. With such a structure an increase in the dependency ratio will always put pressure on the pension system. The reform most often discussed in the United States is for the social security system to move toward a fully funded structure. Once the system reaches the point of being fully funded, pensions are paid from the pension fund accumulated by each worker. This breaks the identity relating pensions to the dependency ratio. A fully funded system can operate either as a government-run scheme or on the basis of private pensions. We comment on this choice below. For now, we note that as well as reducing the real value of the pension, the UK government has moved in the direction of a fully funded program by encouraging the use of private pensions. The difficulty with this approach is that it relies on workers making adequate provision for their retirement—and there is much evidence that this is not the case.

If an economy were to reform its pension system, it would take some time to transit from the pay-as-you-go system to the fully funded system. The reform requires that a capital fund be established that requires a period of investment. Furthermore the pay-as-you-go system cannot be terminated abruptly. Those already retired will still require the provision of their pensions, and those close to retirement will have too little time to invest in a pension fund and so will require the continuation of the pay-as-you-go element. These facts imply that those who are working during the transition process
will have to both pay the pensions of the retired and pay to finance their own pension fund. In simple terms, they are paying for two sets of pensions and fare badly during the reform process. At the very least, this suggests that there could be significant political pressure against the proposed reform.

It is interesting to consider the extent to which social security provision is determined by political considerations. Evidence on this is provided by Mulligan, Gil, and Sala-i-Martin (2002) in their analysis of social security and democracy. Their key finding is that social security has little to do with the voting process because countries without voting still supply public insurance in the same way. They even observe for Chile that most of the growth in social security spending occurred under nondemocratic regimes, and payroll taxes reached extremely high levels under General Augusto Pinochet. In fact, they report on nine dynamic case studies—Greece, Portugal, Spain, Italy, Argentina, Brazil, Chile, Peru, and Uruguay—for the period 1960 to 1990. The countries were selected on the basis of their extreme political changes over this period. With the exception of Greece, it is found that formerly nondemocratic countries do not, relative to their democratic neighbors, change their social security programs after experiencing democracy (in terms of the amount of public insurance spending, and the design of tax and benefit formulas). Similarly, formerly democratic countries do not change their program when they become nondemocratic. Furthermore, multiple regression studies of the determinants of public insurance spending, controlling for population age and per capita income, find neither a significant partial correlation between democracy and social insurance spending (relative to GDP) nor a significant interaction between democracy and the other variables in a spending regression. These results suggest that the role of political constraints on social security may sometimes be overstated.

It is useful to stress a classical error that often accompanies discussion of switching to a fully funded system. The error arises from comparing the likely rates of return on personal accounts with those paid under the current pay-as-you-go system. The proposition that suggests switching to the fully funded system to benefit from the opportunity for higher rates of return is a fallacy. Compare first the real rate of return delivered by the existing social security over the last decades (about 2 percent per year) with the risk-free rate of return of 3 to 4 percent that personal accounts could guarantee by holding inflation-indexed US Treasury securities. The return in the existing system is only 2 percent because of the arithmetic of the pay-as-you-go system.

Suppose that all workers contribute a fixed fraction of their incomes to social security. The key point is that today’s contributions cover the pension benefits of today’s retirees, who were the previous generation of workers that contributed. The total return corresponds to the growth of overall wage income (population plus productivity growth
Thus the real rate of return in an ongoing system is 2 percent if the economy grows at that rate in the long run.

There is a fallacy to the argument that 3 to 4 percent yield on personal accounts is better. The fallacy is that the return on the existing system is low because workers start with a liability to provide for the retirees of the previous generation. If the workers could defect from their liability to the current elderly, they could earn a rate much higher than 2 percent, even if no personal accounts were introduced. But, of course, no one wants to cut the benefits of the elderly who contributed to the system throughout all their working lives. To put it differently, the opportunity of a higher rate of return with personal accounts comes from the misleading feature that they come with no obligation to raise the pensions of the current elderly. This is the feature that accounts for the differences in returns. Moreover the higher expected return is offset by at least the perception of greater risk. This is not to say that the returns in the existing system are risk-free. The major risk in the present system is probably that pension benefits paid in the future are subject to the political whims of future governments.

The distributional effects of a reform from a pay-as-you-go system to a fully funded system are illustrated by the simulation reported in table 23.3. This simulation determines the growth path of an economic model for a reference case in which the state pension is held constant. Applied to the United Kingdom, the model assumes that the value of the pension is 20 percent of average earnings. For the application to Europe, the value is taken to be 40 percent. A reform is then considered where an announcement

### Table 23.3
Gains and losses in transition

<table>
<thead>
<tr>
<th>Age in 1997</th>
<th>United Kingdom</th>
<th>Europe</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;57</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>50–57</td>
<td>−0.09</td>
<td>−0.6</td>
</tr>
<tr>
<td>40–50</td>
<td>−1.1</td>
<td>−2.3</td>
</tr>
<tr>
<td>30–40</td>
<td>−3.0</td>
<td>−5.7</td>
</tr>
<tr>
<td>20–30</td>
<td>−3.8</td>
<td>−7.2</td>
</tr>
<tr>
<td>10–20</td>
<td>−2.3</td>
<td>−4.2</td>
</tr>
<tr>
<td>0–10</td>
<td>0.7</td>
<td>1.7</td>
</tr>
<tr>
<td>−10–0</td>
<td>3.95</td>
<td>9.2</td>
</tr>
<tr>
<td>−20–−10</td>
<td>6.5</td>
<td>15.7</td>
</tr>
<tr>
<td>−40–−30</td>
<td>7.4</td>
<td>18.7</td>
</tr>
<tr>
<td>&lt;−40</td>
<td>7.2</td>
<td>18.9</td>
</tr>
</tbody>
</table>

is made in 1997 (the year the research was conducted) that the state pension will be steadily reduced from the year 2020 until being phased out in 2040. The aim of the long period between announcement and reduction is to allow for adjustment in private behavior. The removal of the state pension implies that private savings will have to increase to compensate.

The negative ages in the first column of table 23.3 refer to consumers who had not yet been born in 1997, so a consumer with age $-10$ in 1997 will be born in 2007. The numbers in the second and third columns shows the percentage by which the lifetime wage of that age group would need to be changed in the reference case to give the same level of welfare as in the reform case. Hence the value of $-1.1$ for the age group 40 to 50 in the United Kingdom shows that this group is worse off with the reform—a reduction of 1.1 percent of their wage in the base case would give them the same welfare level as in the reform case.

The values in table 23.3 show that the pension reform hurts those early in life who must pay the pensions of the retired and pay into their own retirement fund. Ultimately the reform benefits consumers in the long run. The long-run gain comes from the fact that the reduction in the pension leads to an increase in private saving. Private saving has to be invested, so there is also an increase in the capital stock. The consequence of this capital stock increase depends on the initial level of capital compared to the Golden Rule level. In the simulations, capital is initially below the Golden Rule level and remains so throughout the transition. But since this is moving the economy closer to the Golden Rule, there is ultimately a gain in welfare for later generations. The structure of the gains and losses also illustrates the political problem involved in implementing the reform: those who must vote in favor of its implementation are those who lose the most. This political problem will be exacerbated by the aging of the electorate that is expected over the next 50 years. Estimates of the age of the median voter are given in table 23.4. These estimates reveal that the age of the median voter is likely to rise from the midforties to the midfifties. So the electorate will become dominated by the age group that will lose most if the pension system reform is undertaken.

It has already been noted that a fully funded scheme run by the government is equivalent to a system of private pension provision. This is only strictly true in an economy, like the overlapping generations model we have studied, that has a single capital good. In a more practical setting with a range of investment assets, the equivalence will only hold if the same portfolio choices are made. Moving from a pay-as-you-go system to a fully funded system run by the government raises the issue of the portfolio of investments made by the pension fund. In the United States the assets of the fund are invested entirely in long-term Treasury debt. Such debt is very low risk, but as a consequence
it also has a low return. This is not a portfolio that any private sector institution would choose, except one that is especially risk-averse. Nor is it one that many private investors would choose. Permitting the social security fund to invest in a wider portfolio opens the possibility for a higher return to be obtained but introduces questions about the degree of investment risk that the pension fund could accept. In addition changing the portfolio structure of the social security fund could have significant macroeconomic consequences because of its potential size.

A further issue in the design of a pensions system is the choice between a defined contributions system and a defined benefits system. In a defined contribution scheme, social security contributions are paid into an investment fund, and at the time of retirement the accumulated fund is annuitized. What annuitized means is that the fund purchases an annuity, which is a financial instrument paying a constant income to the purchaser until his date of death. In a defined benefits scheme, contributions are made at a constant proportion of income and the benefit is a known fraction of income at retirement (or some average over income levels in years close to retirement).

The consequences of these differences are most apparent in the apportionment of risk under the two types of system. With a defined contributions system, the level of payment into the pension fund is certain for the worker. What is not certain is the maturity value of the pension fund, since this depends on the return earned on the fund, or the pension that will be received, since this depends on the rate offered on annuities at retirement.
All risk therefore falls upon the worker. With a defined benefits system, the risk is placed entirely on the pension fund, since it must meet the promises that have been made. The pension fund receives contributions that it can invest, but it runs the risk that the returns on these investments may not meet pension commitments. This is currently the situation of the US fund where the forecast deficit is a consequence of the defined benefits it has promised.

Assuming that a defined contributions scheme is chosen, there is a further reform that can be made. In the discussion of the simulation it was noted that the reform involved a move from a state pension scheme to private pension schemes. In a defined contributions system there is no real distinction between state and private schemes in principle. When put into practice, distinctions will arise in the choice of investment portfolio, the returns earned on the portfolio and the transactions costs incurred in running the scheme. If moving to a fully funded system pensions, the choice between state and private become a real issue. One option is to use a public fund, either directly administered or run privately after a competitive tendering process. Alternatively, a limited range of approved private funds could be made available. Both choices would lead to a problem of monitoring the performance of the schemes given the fundamentally uncertain nature of financial markets. In addition, seeking low transactions costs could prove detrimental to other areas of performance. A final option is to make use of an open selection of private investment funds. Doing so relies on investors making informed choices between the providers and between the funds on offer to ensure that the risk characteristics of the fund match their preferences. Such a scheme will not work with poorly informed investors and may run foul of high transactions costs. Both of these have been significant problems in the United Kingdom where “mis-selling”—the selling of pensions plans with inappropriate risk characteristics for the purchasers—and high costs have accompanied the move toward the private financing of pensions.

The reform of pensions systems is an issue with much current policy relevance. A range of reforms have been suggested to cope with the forecast change in the dependency ratio. Some of these represent adjustments to the structure of pension schemes, whereas others seek a major reorganization of pension provision.

23.10 Conclusions

Social security in the form of pensions is important both in policy relevance and for its effect on the economy. The generosity of a pension scheme has implications for
individual saving behavior and, in the aggregate, for capital accumulation. Since an economy may reach an inefficient steady state, the designs of pension schemes have an impact on economic efficiency.

Demographic changes and changes in employment behavior are currently putting existing state pension schemes under pressure because of their fundamentally pay-as-you-go nature. Reform proposals have focused on a move to a fully funded system, but such a reform can be detrimental to the welfare level of consumers living during the transition period.

Further Reading

The data on the dependency ratio is taken from:

A survey of policy issues in social security is:

The original analysis of an optimal program can be found in:

Policy neutrality is analyzed in:

The analysis of social security and democracy is conducted in:


The forecasts on the effect of pension reform are taken from:

The forecasts of the age of the median voter are from:
Extensive discussion of reform proposals are in:


Exercises

23.1 Some economists argue that immigration will overwhelm the welfare system, others that immigration will avert a pensions crisis. Which view do you support, and why?

23.2 In Dickens’s novel *David Copperfield*, the character of Mr Micawber has a memorable understanding of intertemporal economics: “Income twenty shillings, expenses nineteen shillings and sixpence—result happiness. Income twenty shillings, expenses twenty shillings and sixpence—result, misery.” Discuss this observation.

23.3 If you work for 30 years and wish to retire for 15 years on 50 percent of your working income, how much of your income must be saved when working? (Assume that the interest rate and income when working are constant, and that there are no taxes.)

23.4 Assume that all consumers have preferences represented by $U = x_t^\alpha x_{t+1}^{1-\alpha}$. If the budget constraint is $x_t + \frac{x_{t+1}^{\gamma}}{1+\gamma} = y_t - \tau + \frac{\beta}{1+\gamma}$, determine the relationship between the level of savings and the parameters $\tau$ and $\beta$ of the social security program. Assuming that $\gamma_t = k_t^\alpha$, find the steady-state level of the capital–labor ratio. Solve for the social security programs that lead to the Golden Rule. Show that none of these programs is fully funded. What is the form of the pay-as-you-go system that achieves the Golden Rule?

23.5 For the economy described in exercise 23.4, relate the structure of social security programs achieving the Golden Rule to dynamic efficiency and inefficiency.

23.6 A common policy is to make pension contributions tax deductible and to insist that the pension fund be annuitized on retirement. Explain the logic behind this policy.

23.7 To avoid a pensions crisis, the UK government is reducing the real value of the guaranteed state pension. Assuming that much of the population is unaware of this, is the policy credible?

23.8 Consider an economy with a single consumer. The government gives the consumer a bond with a face value of $1,000. Has the wealth of the consumer increased? Alternatively, the government levies a tax of $1,000 on the consumer and promises to pay a pension. Has the wealth of the consumer decreased?

23.9 “The operation of a pay-as-you-go pension system is like a hotel with an infinite number of rooms: the hotel can never be full, since a vacancy can be obtained by making all occupants move along one room. Both are theoretical constructs, and neither has practical value.” Discuss.

23.10 Consider a consumer with true preferences $U = [x_t^\alpha x_{t+1}^{1-\alpha}]^{1-\alpha}$. Rather than acting on the basis of these preferences, the consumer is myopic and does not realize the true value of
second-period consumption. The myopic preferences are given by
\[ U = \left[ x_t^\alpha \right]^{\frac{1}{1-\alpha}} \rho x_{t+1}^{1-\alpha}, \]
r < 1.

a. Determine how the level of saving depends on ρ.
b. How does the level of welfare measured by true preferences depend on ρ?
c. Suppose that a population of \( H \) consumers act according to these myopic preferences and that the equilibrium interest rate is \( r_{t+1} = a - b s_t \), where \( s_t \) is the total level of savings in the economy. Can myopia ever increase the consumers’ true utilities?
d. Does this form of myopia provide a justification for social security?

23.11 For the myopia model, suppose a pay-as-you-go pension system. The consumers overestimate the generosity of the pension scheme and believe that the pension, \( β \), and the social security tax, \( τ \), are related by \( β = (1 + \phi)τ \), where \( \phi > 0 \). There is no population growth, so the true value of the pension is \( β = τ \). What effect does an increase in \( \phi \) have on savings? Does welfare increase or decrease in \( \phi \)? Should we have the social security program when consumers have this from of myopia?

23.12 Consider an economy where individuals live for two periods only. Their utility function over consumption in periods 1 and 2 is given by \( U = 2 \log(C_1) + 2 \log(C_2) \), where \( C_1 \) and \( C_2 \) are period 1 and period 2 consumption levels respectively. They have labor income of $100 in period 1 and labor income of $50 in period 2. They can save as much of their income in period 1 as they like in bank accounts, earning interest rate of 5 percent per period. They have no bequest motive, so they spend all their income before the end of period 2.

a. What is each individual’s lifetime budget constraint? If they choose consumption in each period so as to maximize their lifetime utility subject to their lifetime budget constraint, what is the optimal consumption in each period? How much do the consumers save in the first period?
b. Suppose that the government introduces a social security system that will take $10 from each individual in period 1, put it in a bank account, and transfer it back to them with interest in period 2. What is the new lifetime budget constraint? What is the effect of this social security system on private savings? How does the system affect total savings in society?

23.13 Consider the previous exercise and suppose that the introduction of social security induces the individuals to retire in period 2. So they receive no labor income in period 2.

a. What is the new optimal consumption in each period? How much do the consumers save? How does it compare with previous exercise? Explain.
b. Now building on this example, should the actual social security system lead to early retirement? Why or why not? What is the evidence on the impact of social security on the retirement decision in the United States and elsewhere?

23.14 The pensions crisis arises because the number of working people is falling relative to the number of retired people. Many politicians argue that the crisis can be solved by accumulating greater pension savings when working. Is this a real solution?

23.15 Consider an individual who lives for two periods and has utility of lifetime consumption \( U = \log(C_1) + \frac{1}{1-\delta} \log(C_2) \), where \( C_1 \) and \( C_2 \) are the consumption levels in the first and second period respectively, and \( \delta, 0 < \delta < 1 \), denotes the per period discount rate. Suppose
that the individual has an income of $Y_1 > 0$ in the first period and no income in the second period, so $Y_2 = 0$. He can transfer some income to the second period at a before-tax rate of return of $r$, so saving $S$ in the first period gives $[1 + r]S$ in the second period. The government levies a capital tax at rate $\tau$ on capital income received in the second period. The tax proceeds are paid as a lump-sum transfer to the following generation. The present generation does not care about the next one.

a. What is the lifetime consumption profile of this individual? What is his lifetime indirect utility function expressed as a function of $Y_1$ and $(1 - \tau) r$?

b. Evaluate the change in initial income $Y_1$ that is required to compensate the individual for the welfare loss due to the capital income tax $\tau$.

c. What is the impact of a tax rate change on consumption level in the first period? And in the second period? What conclusion about the welfare cost of capital income taxation can you draw from your analysis?

23.16 A government accumulates public capital for a social security scheme according to the budget constraint

$$K_t^s = (1 - \delta)K_{t-1}^s + T_t - G,$$

where $K_t^s$ is the level of public capital, $T_t$ the taxes raised, and $G$ the level of general government expenditure. In year 1, the government decides implement a permanent one-unit reduction in tax revenue.

a. How much does the stock of public capital fall in year 1 to sustain the tax cut? And in year $t$?

b. How is the steady-state value of $K^s$ affected by the tax cut?

c. Using the Ricardian equivalence argument, discuss the impact of the tax change on individual behavior.

23.17 a. How do the results of the previous exercise affect the choice between contributory and funded pension schemes?

b. How likely is Ricardian equivalence to hold in this circumstance? (You are welcome to apply arguments of behavioral economics.)

23.18 Consider an economy where individuals live for two periods only. They have the utility function over consumption in period 1 ($C_1$) and period 2 ($C_2$) given by $U = 2 \log(C_1) + 2 \log(C_2)$. The labor income of each individual in period 1 is fixed at $10$, and there is no labor income in period 2. They can save as much of their income in period 1 as they like in bank accounts, earning interest rates of 200 percent per period (recall, a period is the entire active life). The income tax rate is 50 percent, which is used to pay back the public debt inherited from the past generation.

a. Derive the optimal lifetime consumption profile of this consumer. What would be the consumption profile without income tax?

b. Suppose that a “retirement saving program” is introduced allowing each consumer to save up to 20 percent in the first period in a tax-free account. Compare the lifetime budget constraints with and without the retirement saving program.
Chapter 23: Social Security

c. Derive the optimal lifetime consumption profile with the retirement saving program. Explain the impact of this program on private saving.

d. Now suppose that the retirement saving program in part b is replaced by a new savings program taxing investment income on the first 50 percent of savings and exempting any saving in excess of 50 percent from taxation. Draw the budget set associated with this program, and find the optimal lifetime consumption profile. Explain the difference with the program in part b.

e. If the threshold for tax-exempt savings in part b is increased from 50 to 51 percent, how would this affect private savings? How does this affect total savings in society?

23.19 A sixty-year-old widow is considering claiming her social security widow benefit now. The primary insurance amount (PIA) from her deceased spouse is $15,000. If she claims them now, she will receive only 72 percent of her PIA for the rest of her life. If she claims her benefits at age 65, she will receive 100 percent of her PIA for the rest of her life. Suppose that her life expectancy is 80.

a. What is the present discounted value of her benefits when claiming at either age 60 or age 65? At what age should she claim her benefits?

b. Suppose now that her life expectancy is longer. How would it change your recommendation in part a?

c. If she expected to get remarried at age 60, how might your answer change? Will her remarriage affect her benefits anyway?

23.20 “Social security reduces private savings, since individuals who retire earlier have fewer earnings to save.” True or false?

23.21 Suppose that the government seeks to raise savings and is considering expanding the Individual Retirement Accounts (IRAs) to do this. US evidence suggests that IRA holders save more than non-IRA holders.

a. How can the theory explain that IRAs increase saving? What are the income effects and the substitution effects? What about the replacement effect of substituting IRAs for savings already done outside the program?

b. Is the evidence-based prediction that IRAs increase savings necessarily correct? Why or why not? How might the distinction between private and national savings affect the analysis?

23.22 What are the advantages and problems related to a reform of social security that consists of switching to individual annuitized accounts?

23.23 Social security requires mandatory participation in most countries. Milton Friedman (1999) states that “The fraction of a person’s income that is reasonable for him or her to set aside for retirement depends on that person’s circumstances and values. It makes no more sense to specify a minimum fraction of income that must be spent on housing or transportation.” Discuss the possible implications of making participation in social security voluntary. You can also use arguments of adverse selection and moral hazard covered in chapter 10.
24 Economic Growth

24.1 Introduction

Economic growth is the basis of increased prosperity. Growth comes from the accumulation of capital (both human and physical) and from innovation that leads to technical progress. These advances raise the productivity of labor and increase the potential for consumption. The rate of growth can be affected by policy through the effect that taxation has on the return to investments. Taxation can also finance public expenditures that enhance productivity. In most developed countries the level of taxes has risen steadily over the course of the last century: an increase from about 5 to 10 percent of gross domestic product at the turn of the twentieth century to between 30 and 40 percent at present is typical. Such significant increases raise serious questions about the effect taxation has on economic growth.

Until recently economic models that could offer convincing insights into this question were lacking. Much of the growth literature focused on the long-run equilibrium where output per head was constant or modeled growth through exogenous technical progress. By definition, when technical progress is exogenous, it cannot be affected by policy. The development of endogenous growth theory has overcome these limitations by explicitly modeling the process through which growth is generated. This allows the effects of taxation to be traced through the economy and predictions made about its effects on growth.

The chapter begins with a review of exogenous growth models. The concept of the steady state is introduced, and it is shown why growth is limited unless there is some external process of technical progress. The exogenous growth model is employed to prove the important result that the optimal long-run tax rate on capital income should be zero. Actual tax systems are some way from this ideal position, so the welfare cost of the nonoptimality is also addressed.

Endogenous growth models are then considered. A brief survey is given of the various ways in which endogenous growth has been modeled. The focus is placed on endogenous growth arising from the provision of a public input for private firms. It is shown that there is an optimal level of public expenditure that maximizes the growth rate of consumption. This model provides a positive role for government in the growth process. The optimality of a zero tax on capital extends to endogenous growth models with human capital as an input. With this result in mind, a range of simulation
experiments has assessed the effect on the growth rate of changes in the tax structure in this setting. The differences in structure and parameter values between the experiments provides for some divergent conclusions.

The analytical results and the simulations reveal that economic theory provides no definitive prediction about how taxation affects economic growth. The limitations of the theory places an increased reliance on empirical evidence to provide clarification. We look at a range of studies that have estimated the effect of taxation on economic growth. Some of these studies find a significant effect, and others do not. We discuss the many issues involved in interpreting these results.

24.2 Exogenous Growth

The exogenous growth theory that developed in the 1950s and 1960s viewed growth as being achieved by the accumulation of capital and increases in productivity via technical progress. The theory generally placed its emphasis on capital accumulation, so the source of the technical progress was not investigated by the theory. It was assumed instead to arise from some outside or exogenous factors.

The standard form of these growth models was based on a production function that had capital and labor (with labor measured in manhours) as the inputs into production. Constant returns to scale were assumed, as was diminishing marginal productivity of both inputs. Given that the emphasis was on the level and growth of economic variables rather than their distribution, the consumption side was modeled by either a representative consumer or a steadily growing population of identical consumers.

Our analysis begins with the simplest of these growth models, which assumes that both the rate of saving and the supply of labor are constant. This model is a special case of the general Solow growth model. Although the assumption of a constant saving rate eliminates issues of consumer choice, the model is still able to teach important lessons about the limits to growth and the potential for efficiency of the long-run equilibrium. The key finding is that if growth occurs only through the accumulation of capital, there has to be a limit to the growth process if there is no technical progress.

24.2.1 Constant Saving Rate

The fact that there are limits to growth in an economy when there is no technical progress can be easily demonstrated in a setting where consumer optimization plays no role. It is assumed instead that a constant fraction of output is invested in new capital goods.
This assumption may seem restrictive, but it allows a precise derivation of the growth path of the economy. In addition the main conclusions relating to limits on growth are little modified even when an optimizing consumer is introduced.

Consider an economy with a population that is growing at a constant rate. Each person works a fixed number of hours and capital depreciates partially when used. There is a single good in the economy that can be consumed or saved. The only form of saving is investment in capital. Under these assumptions the output that is produced at time \( t \), \( Y_t \), must be divided between consumption, \( C_t \), and investment, \( I_t \). In equilibrium the level of investment must be equal to the level of saving.

With inputs of capital \( K_t \) and labor \( L_t \) employed in production, the level of output is

\[
Y_t = F(K_t, L_t).
\]  

(24.1)

It is assumed that there are constant returns to scale in production. Output can be either consumed or saved. The fundamental assumption of the model is that the level of saving is a fixed proportion \( s, 0 < s < 1 \), of output. As saving must equal investment in equilibrium, at time \( t \) investment in new capital is given by

\[
I_t = sF(K_t, L_t).
\]  

(24.2)

The use of capital in production results in its partial depreciation. We assume that this depreciation is a constant fraction \( \delta \), so that the capital available in period \( t+1 \) is given by new investment plus the undepreciated capital, or

\[
K_{t+1} = I_t + [1 - \delta] K_t
\]

\[
= sF(K_t, L_t) + [1 - \delta] K_t.
\]  

(24.3)

This equation is the basic capital accumulation relationship that determines how the capital stock evolves through time.

The fact that the population is growing makes it preferable to express variables in per capita terms. This can be done by exploiting the assumption of constant returns to scale in the production function to write \( Y_t = L_t F\left(\frac{K_t}{L_t}, 1\right) = L_t f(k_t) \), where \( k_t = \frac{K_t}{L_t} \). After we divide (24.3) through by \( L_t \), the capital accumulation relation becomes

\[
\frac{K_{t+1}}{L_t} = s f(k_t) + \frac{[1 - \delta]}{L_t} K_t.
\]  

(24.4)

We denote the constant population growth rate by \( n \), so that labor supply grows according to \( L_{t+1} = (1 + n) L_t \). Using this growth relationship, we obtain the capital accumulation relation, which shows that the dynamics of the capital–labor ratio are
The relation in (24.5) will trace the development of the capital stock over time from an initial stock \( k_0 = K_0 L_0 \). To see what this implies, consider an example where the production function has the form \( f(k_t) = k_t^\alpha \). The capital–labor ratio must then satisfy

\[
(24.6)
\]

For \( k_0 = 1 \), \( n = 0.05 \), \( \delta = 0.05 \), \( s = 0.2 \), and \( \alpha = 0.5 \), figure 24.1 plots the first 50 values of the capital stock. It can be seen that starting from the initial value of \( k_0 = 1 \), the capital stock doubles in 13 years. After this the rate of growth slows noticeably, and even by the fiftieth year it has not yet doubled again. The figure also shows that the capital stock is tending to a long-run equilibrium level that is called the \textit{steady state}. For the parameters chosen, the steady-state level is \( k^* = 4 \), which is achieved at \( t = 328 \), though the economy does reach a capital stock of 3.9 at \( t = 77 \). It is the final part of the adjustment that takes a long time.

The steady state is achieved when the capital stock is constant with \( k_{t+1} = k_t \). Denoting the steady-state value of the capital–labor ratio by \( k \), we have from the capital accumulation condition that \( k \) must satisfy

\[
(24.7)
\]
The solution to this equation is called the steady-state capital–labor ratio, and it can be interpreted as the economy’s long-run equilibrium value of $k$. The solution of this equation is illustrated in figure 24.2. The steady state occurs where the curves $sf(k)$ and $[n + \delta]k$ intersect. If this point is achieved by the economy, the capital–labor ratio will remain constant. Since $k$ is constant, it follows from the production function that $\frac{Y}{L_t}$ will remain constant as will $\frac{C_t}{L_t}$. (However, it should be noted that as $L$ is growing at rate $n$, then $Y$, $K$, and $C$ will also grow at rate $n$ in the steady state.) It is the constancy of these variables that shows there is a limit to the growth achievable by this economy. Once $\frac{C_t}{L_t}$ is constant, the level of consumption per capita will remain constant over time. In this sense a limit is placed on the growth in living standards that can be achieved. The explanation for this limit is that capital suffers from decreasing returns when added to the exogenous supply of labor. If excessive capital is employed, the return will fall so low that the capital stock is unable to reproduce itself.

Although we have not yet included any policy variables, this analysis of the steady state can be used to reflect on the potential for economic policy to affect the equilibrium. Studying figure 24.2 reveals that the equilibrium level of $k$ can be raised by any policy that engineers an increase in the saving rate, $s$, or an upward shift in the production function, $f(k)$. However, any policy that leads only to a one-off change in $s$ or $f(k)$.\[ sf(k) - (n + \delta)k = 0. \] (24.8)
cannot affect the long-run growth rate of consumption or output. By definition, once the new steady state is achieved after the policy change, the per capita growth rates of the variables will return to zero. Furthermore any policy that only increases $s$ cannot sustain growth, since $s$ has an upper limit of 1 that must eventually be reached. If policy intervention is to result in sustained growth, it has to produce a continuous upward movement in the production function. A mechanism through which policy can achieve this is studied in section 24.3.2.

A means for growth to be sustained without policy intervention is to assume that output increases over time for any given level of the inputs. Input gains can be achieved by labor or capital (or both) becoming more productive over time for exogenous reasons summarized as “technical progress.” A way to incorporate this in the model is to write the production function as $f(k, t)$, where the dependence on $t$ captures the technical progress that allows increased output. Technical progress results in the curve $f(k, t)$ in figure 24.2 continuously shifting upward over time, thus raising the steady-state levels of capital and output. The drawback of this approach is that the mechanism for growth, the “growth engine,” is exogenous, so preventing the model from explaining the most fundamental factor of what determines the rate of growth. This deficiency is addressed by the endogenous growth models of the next section, where we explore the mechanisms that can drive technical progress.

Returning to the basic model without technical progress, we have by condition (24.8) that the steady-state capital–labor ratio is dependent on the savings rate $s$. This raises the question as to whether some saving rates are better than others. To address this question, we note first that for each value of $s$ there is a corresponding steady-state capital–labor ratio at the intersection of $sf(k)$ and $[n + \delta]k$. It is clear from figure 24.2 that for low values of $s$, the curve $sf(k)$ will intersect the curve $[n + \delta]k$ at low values of $k$. As $s$ is increased, $sf(k)$ shifts upward and the steady-state level of $k$ rises. The relationship between the capital–labor ratio and the saving rate implied by this construction is denoted by $k = k(s)$. We have observed that $k(s)$ is an increasing function of $s$ up until the maximum value of $s = 1$.

Employing the link between $s$ and $k$ allows the level of consumption per capita to be written as

$$c(s) = (1 - s)f(k(s)) = f(k(s)) - (n + \delta)k(s), \quad (24.9)$$

where the second equality follows from definition (24.8) of a steady state. What is of interest are the properties of the saving rate that maximizes consumption. The first-order condition for defining this savings rate can be found by differentiating $c(s)$ with respect to $s$. Doing so gives
\[
\frac{dc(s)}{ds} = \left[ f'(k(s)) - [n + \delta] \right] k'(s) = 0. \tag{24.10}
\]

Since \( k'(s) \) is positive, the saving rate, \( s^* \), that maximizes consumption is defined by

\[
f'(k(s^*)) = n + \delta. \tag{24.11}
\]

The savings rate \( s^* \) determines a level of capital \( k^* = k(s^*) \), which is called the Golden Rule capital–labor ratio. If the economy achieves this capital–labor ratio at its steady state, it is maximizing consumption per capita. The same logic applies here as it did in the derivation of the steady state in chapter 22 (though \( \delta \) was assumed to be zero in the overlapping generations economy).

The nature of the Golden Rule is illustrated in figure 24.3. For any level of the capital–labor ratio, the steady-state level of consumption per capita is given by the vertical distance between the curve \( [n + \delta]k \) and the curve \( f(k) \). This distance is maximized when the gradient of the production function is equal to \( [n + \delta] \), which gives the Golden Rule condition. The figure also shows that consumption will fall if the capital–labor ratio is either raised or lowered from the Golden Rule level. In line with the definitions of chapter 22, an economy with a steady-state capital stock below the Golden Rule level, \( k^* \), is dynamically efficient—it requires a sacrifice of consumption now in order to raise \( k \)—so a Pareto improvement cannot be found. An economy with a capital stock in excess of \( k^* \) is dynamically inefficient, since immediate consumption of the excess would raise current welfare and place the economy on a path with higher consumption.

![Golden Rule](image)
For an example of these calculations, let the production function be given by \( y = k^\alpha \), with \( \alpha < 1 \). For a given savings rate \( s \), the steady state is defined by the solution to

\[
sk^\alpha = (n + \delta)k.
\]

(24.12)

Solving this equation determines the steady-state capital–labor ratio as

\[
k = \left( \frac{s}{n+\delta} \right)^{1/(1-\alpha)}.
\]

From this solution the per capita level of consumption follows as

\[
c(s) = k^\alpha - (n + \delta)k = \left( \frac{s}{n+\delta} \right)^{\alpha/(1-\alpha)} - (n + \delta) \left( \frac{s}{n+\delta} \right)^{1/(1-\alpha)}.
\]

(24.13)

We adopt the parameter values \( n = 0.025, \delta = 0.025, \) and \( \alpha = 0.75 \), and plot in figure 24.4 the level of consumption as a function of \( s \). In the figure note that consumption rises with \( s \) until the optimal saving rate is reached, at which point the equilibrium capital stock is equal to the Golden Rule level, and then falls again for higher values.

Formally, the fact that the savings rate is fixed leaves little scope for policy analysis. However, studying the effect of changes in the saving rate reveals the factors that would be at work in a more general model in which the level of savings is a choice variable that can be affected by policy. By definition, the per capita level of the variables is constant once the steady state has been achieved. The living standards in the economy reach a limit and then cannot grow any further unless the production function is continually raised. Changes in the saving rate affect the level of consumption but not its growth rate.
24.2.2 Optimal Taxation

The analysis of the fixed saving model has touched on some of the potential consequences of policy intervention. As a tool for policy analysis, the model is very limited given the lack of choice variables that can be affected by policy. This shortcoming is now overcome by studying a variant of the Ramsey growth model in which a representative consumer chooses an intertemporal consumption plan to maximize lifetime utility. Using this model, we analyze the optimal taxes on labor and capital income.

The Ramsey model has a single representative consumer who chooses the paths of consumption, labor, and capital over time. The single consumer assumption is adopted to eliminate issues concerning distribution among consumers of differing abilities and tastes, and to place the focus entirely on efficiency. For simplicity, it is also assumed that the growth rate of labor, \( n \), is zero. There is a representative firm that chooses its use of capital and labor to maximize profits. Given that the market must be in equilibrium, the choices of the consumer drive the rest of the economy through the level of savings, and hence capital, that they imply. The supply of labor and capital from the consumer combine with the factor demands of the firm to determine the equilibrium factor rewards.

The aim is to characterize the optimal tax structure in this economy. We assume there is a government that requires revenue of amount \( g_t \) at time \( t \). It raises this revenue through taxes on capital and labor, which are denoted by \( \tau^K_t \) and \( \tau^L_t \) respectively. The government chooses these tax rates in the most efficient manner.

Before proceeding, it is helpful to discuss the choice of tax instruments. A question that is frequently analyzed in the literature is whether it is better to tax consumption or to tax income. The point we now wish to emphasize is that this question is not a clear-cut as many presentations suggest. This can be seen by taking a simple one-period budget constraint with two goods. With a tax \( t \) on income the consumption choices \( x_1 \) and \( x_2 \) must satisfy

\[
(1 - t)M = p_1x_1 + p_2x_2. \quad (24.14)
\]

Observe that the real constraint remains the same if (24.14) is rewritten

\[
M = \frac{1}{1 - t} p_1x_1 + \frac{1}{1 - t} p_2x_2 = [1 + \tau] p_1x_1 + [1 + \tau] p_2x_2, \quad (24.15)
\]

where \( \tau = \frac{t}{1 - t} \). In words, the two budget identities show that an income tax at rate \( t \) is equivalent to a consumption tax at rate \( \frac{t}{1 - t} \). Since the two taxes are equivalent, there is no question of one being better or worse than the other. This equivalence between
the income tax and the consumption tax can be extended to the Ramsey growth model. Assume, for ease of exposition, that the tax rates are constant over time. The equivalence is stated formally in the proposition.

**Proposition 24.1** (Tax system equivalence) A tax on labor income, \( \tau^L > 0 \), with an exemption for interest income, \( \tau^K = 0 \), is equivalent to a consumption tax \( \tau^C > 0 \).

The point of the proposition is that the difference between an income tax system—which would tax all flows of income equally—and a consumption tax is entirely captured in the tax treatment of interest income. Hence, if we find that the optimum tax system has \( \tau^L = \tau^K > 0 \), then an income tax is optimal. Alternatively, if the optimum is \( \tau^L > 0, \tau^K = 0 \), then a consumption tax is optimal. Finally, if the optimum is \( \tau^L > 0, \tau^K > 0, \tau^L \neq \tau^K \), then the system is neither an income tax or a consumption tax. These points should make the value of the following analysis clear.

In the Ramsey growth model the choices of the consumer are made to maximize the discounted sum of the flow of utility. Letting \( 0 < \beta < 1 \) be the discount factor on future utility, the consumer’s preferences are described by

\[
U = \sum_{t=0}^{\infty} \beta^t U(C_t, L_t). \tag{24.16}
\]

The specification of the utility function implies that the consumer has an infinite life. This can be justified by treating the consumer as a dynasty with concern for descendents. Further discussion of this assumption can be found in section 23.8.

As there is a single consumer, the capital stock is equal to the savings of this consumer. This observation allows the budget constraint for the consumer to be written as

\[
C_t + K_{t+1} = \left[ 1 - \tau^L_t \right] w_t L_t + \left[ 1 - \delta + \left[ 1 - \tau^K_t \right] r_t \right] K_t. \tag{24.17}
\]

The utility maximization decision for the consumer involves choosing the time paths of consumption, labor supply, and capital for the entire life span of the economy. The formal decision problem is

\[
\max_{\{C_t, L_t, K_t\}} \sum_{t=0}^{\infty} \left[ \beta^t U(C_t, L_t) + \beta^t \lambda_t \left[ \left[ 1 - \tau^L_t \right] w_t L_t + \left[ 1 - \delta + \left[ 1 - \tau^K_t \right] r_t \right] K_t - C_t - K_{t+1} \right] \right], \tag{24.18}
\]

where \( \lambda_t \) is the multiplier on the budget constraint at time \( t \).
In solving this optimization, it is assumed that the representative consumer takes the factor rewards $w_t$ and $r_t$ as given. This captures the representative consumer as a competitive price taker. (It is helpful to note that when we consider the government optimization below, the dependence of the factor rewards on the choice of capital and labor is taken into account by the government. This is what distinguishes the consumer who reacts to the factor rewards and the government that manipulates the factor rewards.) With fixed factor rewards the necessary conditions for the choices of $C_t$, $L_t$, and $K_{t+1}$ are

$$U_{C_t} - \lambda_t = 0,$$  
(24.19)

$$U_{L_t} + \lambda_t \left[ 1 - \tau_t^L \right] w_t = 0,$$  
(24.20)

and

$$\beta \lambda_{t+1} \left[ 1 - \delta + \left[ 1 - \tau_{t+1}^K \right] r_{t+1} \right] - \lambda_t = 0.$$  
(24.21)

Using the first condition to substitute for $\lambda_t$ in the second condition gives

$$U_{L_t} + U_{C_t} \left[ 1 - \tau_t^L \right] w_t = 0.$$  
(24.22)

Stepping the first condition one period ahead and then substituting for $\lambda_{t+1}$ in the third gives

$$\beta U_{C_{t+1}} \left[ 1 - \delta + \left[ 1 - \tau_{t+1}^K \right] r_{t+1} \right] - U_{C_t} = 0.$$  
(24.23)

Conditions (24.22) and (24.23) describe utility maximization by the consumer. To interpret these, it should be observed that there are two aspects to the consumer’s decision. First, within each period the consumer needs to optimize over the levels of consumption and labor supply. The efficient solution to this within-period decision is described by (24.22), which ensures that the marginal utilities are proportional to the relative prices. Second, the consumer has to allocate her resources efficiently across time. Condition (24.23) describes efficiency in this process by linking the marginal utility of consumption in two adjacent periods to the rate at which consumption can be transferred through time via investments in capital. Taken together for every time period $t$, these necessary conditions describe the optimal paths of consumption, labor supply, and capital investment for the consumer.

The representative firm is assumed to maximize profit by choosing its use of capital and labor. Since the firm rents capital from the consumer, it makes no irreversible
decisions, so it need do no more than maximize profit in each period. The standard efficiency conditions for factor use then apply and equate marginal products to factor rewards. Hence the interest rate and the wage rate satisfy

\[ F_K = r_t \quad (24.24) \]

and

\[ F_L = w_t. \quad (24.25) \]

Following these preliminaries, it is possible to state the government optimization problem. The sequence of government expenditures \( \{g_t\} \) is taken as given. It is assumed that these expenditures are used for a purpose that does not directly affect utility. Formally, the government chooses the tax rates and the levels of consumption, labor supply, and capital to maximize the level of utility. The values of these variables must be chosen for each point in time, so government decisions form a sequence \( \{\tau^K_t, \tau^L_t, C_t, L_t, K_t\} \). The choices of \( C_t, L_t \) and \( K_t \) must be identical to what would be chosen by the consumer given the tax rates \( \tau^K_t \) and \( \tau^L_t \). This can be achieved by imposing conditions (24.22) and (24.23) as constraints on the optimization. When these constraints are satisfied, it is as if the consumer were making the choice. As already noted, the government explicitly takes into account the endogenous determination of the factor rewards.

The optimization also has to be constrained by the budget constraints of the consumer and government, and by aggregate production feasibility. However, if any two of these constraints hold, the third must also hold. Therefore one of them need not be included as a separate constraint for the optimization. In this case the consumer’s budget constraint is dropped. The government budget constraint that taxes must equal expenditure is given by

\[ \tau^K_t r_t K_t + \tau^L_t w_t L_t = g_t. \quad (24.26) \]

In addition the aggregate production condition for the economy is that

\[ C_t + g_t + I_t = F(K_t, L_t). \quad (24.27) \]

By the definition of investment, this becomes

\[ C_t + g_t + K_{t+1} = F(K_t, L_t) + (1 - \delta) K_t. \quad (24.28) \]

From the determination of the factor prices (24.24) and (24.25), the government optimization problem that determines the efficient taxes is
The complete set of first-order necessary conditions for this optimization involves the derivatives of the Lagrangian with respect to all the choice variables at every point in time plus the derivatives with respect to the multipliers at every point in time. However, to demonstrate the key result concerning the value of the optimal capital tax, only the necessary conditions for the tax rates and for capital are required. The other first-order conditions will add further information to the solution but do not bear on the determination of the capital tax.

The necessary condition for the choice of \( \tau_t^K \) is

\[
\psi_t F_t K_t - \mu_{2t-1} U_t C_t F_t K_t = 0,
\]

(24.30)

for \( \tau_t^L \) the necessary condition is

\[
\psi_t F_t L_t - \mu_{4t} U_t C_t F_t L_t = 0,
\]

(24.31)

and for \( K_t \) it is

\[
\psi_t \left( \tau_t^K (F_t K_t + K_t F_t K_t) + \tau_t^L F_t L_t K_t + \theta_t (F_t K_t + 1 - \delta) - \frac{1}{\beta} \theta_{t-1} \right) + \mu_{1t} U_t C_t (1 - \delta) + \mu_{2t-1} U_t C_t (1 - \tau_t^K) F_t K_t = 0.
\]

(24.32)

The two conditions for \( \tau_t^K \) and \( \tau_t^L \) can be used to substitute for \( \mu_{1t} \) and \( \mu_{2t-1} \) in the condition for \( K_t \). Canceling terms and using the fact that constant returns to scale implies \( K_t F_t K_t + L_t F_t L_t K_t = 0 \), condition (24.32) reduces to

\[
\psi_t \tau_t^K F_t K_t + \theta_t \left[ F_t K_t + 1 - \delta \right] - \frac{1}{\beta} \theta_{t-1} = 0.
\]

(24.33)

Along the growth path of the economy this equation is only one part of the complete description of the outcome induced by the optimal policy. However, by focusing on the steady state in which all the variables are constant, it becomes possible to use the information contained in this condition to determine the optimal tax on capital.

Consequently the analysis now moves to consider the steady state that is reached under the optimal policy. To be in a steady state, it must be the case that the tax rates and the level of government expenditure remain constant over time. In addition the
levels of capital, consumption, and labor supply will be constant. Moreover being in a steady state also implies that \( \theta_t = \theta_{t-1} \). Using these facts, we have in the steady state the necessary condition for the choice of the capital stock as

\[
\psi \tau^K F_K + \theta (F_K + 1 - \delta) - \frac{1}{\beta} \theta = 0. \tag{24.34}
\]

This condition can be simplified further by observing that in the steady state the choice condition for the consumer (24.23) reduces to

\[
\beta \left( 1 - \delta + (1 - \tau^K) F_K \right) - 1 = 0. \tag{24.35}
\]

We use (24.35) to substitute for \( \beta \), and the final condition for the choice of the capital stock is

\[
[\psi + \theta] \tau^K F_K = 0. \tag{24.36}
\]

Given that the resource constraints are binding, implying that \( \psi \) and \( \theta \) are positive, and that the marginal product of capital, \( F_K \), is positive, the solution to (24.36) has to be \( \tau^K = 0 \). This is the well-known result (due originally to Chamley 1986 and Judd 1985) that the long-run value of the optimal capital tax has to be zero.

The analysis has concluded that in the steady state, which we can interpret as the long-run equilibrium, income from interest on capital should not be taxed. Several points can be made about this result. First, note that the result does not say that the tax should be zero when we are on the growth path to the steady state—it was derived for the steady state and so applies only to that situation. This does not prevent the tax from being positive (or negative) along the growth path. Second, the zero tax on capital income implies that all taxation must fall on labor income. If labor were a fixed factor, this conclusion would not be a surprise, but here labor is a variable factor. In light of our earlier discussion the result can be seen as supporting the optimality of a consumption tax in the steady state. Finally the reason for avoiding the taxation of capital is that the return on capital is fundamental to the intertemporal allocation of resources by the consumer. The result shows that it is optimal to leave this allocation undistorted to focus distortions on the choice between consumption and labor within periods.

Since the optimal tax rate is zero, any other value of the tax rate must lead to a reduction in welfare compared to what is achievable. An insight into the extent of the welfare cost of deviating from the optimal solution is given in table 24.1. These results are derived from a model with a Cobb–Douglas production function and a
utility function with a constant elasticity of intertemporal substitution (see equation 24.47 below). The policy experiment calculates what would happen if a tax on capital is replaced by a lump-sum tax. The increases in consumption and the welfare cost are measured by comparing the steady state with the tax to the steady state without. When a tax rate of 30 percent on capital income is replaced by a lump-sum tax, consumption increases by 3.3 percent and the welfare cost of the distortionary tax is measured at 11 percent of tax revenue. The increases in consumption and the welfare cost are both higher for an initial 50 percent tax rate.

In summary, the optimal tax policy is to set the long-run tax on capital to zero. This outcome is explained by the wish to avoid intertemporal distortions. As a consequence all revenue must be raised by taxation of labor income. This will cause a distortion of choice within periods but does not affect the intertemporal allocation. The conclusion is very general and does not depend on any restrictive assumptions. Simulations of the welfare cost of nonoptimal policies show that these can be a significant percentage of the revenue raised.

### 24.3 Endogenous Growth

Decreasing returns to capital have already been identified as the source of the limit on growth in the exogenous growth model. The removal of this limit requires the decreasing marginal product of capital to be circumvented in a way that is, ideally, determined by choices made by the agents in the economy. Models that allow both sustained growth and explain its source are said to generate “endogenous growth.” There have emerged in the literature four basic methods by which endogenous growth can be achieved. All four approaches achieve the same end—that of sustained growth—but by different routes. We briefly review these four approaches and then focus attention on government expenditure as the source of endogenous growth.
24.3.1 Models of Endogenous Growth

The first, and simplest, approach to modeling endogenous growth, the AK model, assumes that capital is the only input into production and that there are constant returns to scale. This may seem at first sight to simply remove the problem of decreasing returns by assumption, but we will later show that the AK model can be given a broader interpretation. Under these assumptions the production function is given by $Y_t = AK_t$, hence the model’s name. Constant returns to scale ensures that output grows at the same rate as the capital stock.

To show that this model can generate continuous growth, it is simplest to return to the assumption of a constant saving rate. With a saving rate $s$ the level of investment in time period $t$ is $I_t = sAK_t$. Since there is no labor, the capital accumulation condition is just

$$K_{t+1} = sAK_t + (1-\delta)K_t = (1+sA-\delta)K_t. \quad (24.37)$$

Provided that $sA > \delta$, the level of capital will grow over time at rate $sA - \delta$. Output will grow at the same rate, as will consumption. The model is therefore able to generate continuous growth.

The second approach is to match increases in capital with equal growth in other inputs. One way to do this is to consider human capital as an input rather than just raw labor time. The level of the human capital input is then the product of the quality of labor and labor time. Doing so allows labor time to be made more productive by investment in education and training, which raises human capital. Technical progress is then embodied in the quality of labor. The model requires two investment processes: one for investment in physical capital and another for investment in human capital. There can either be one sector, with human capital produced by the same technology as physical capital, or two sectors with a separate production process for human capital. The standard form of the production function for such a model would be

$$Y_t = F(K_t, H_t), \quad (24.38)$$

where $H_t$ is the level of human capital. If the production function has constant returns to scale in human capital and physical capital jointly, then investment in both can raise output without limit even if the quantity of labor time is fixed.

The one-sector model with human capital actually reduces to the AK model—this is the broader interpretation of the AK model referred to above. To see this, note that under the one-sector assumption output can be used for consumption, or invested in physical capital, or invested in human capital. This means that the two capital goods are
perfect substitutes for the consumer in the sense that a unit of output can become one unit of either. The perfect substitutability implies that in equilibrium the two factors must have the same rate of return. Combining this with the constant returns to scale in the production function results in the two factors always being employed in the same proportions. Therefore the ratio \( \frac{H}{K} \) is constant for all \( t \). Denoting this constant value by \( \frac{H}{K} \), we write the production function as

\[
Y_t = K_t f \left( \frac{H}{K} \right) = AK_t,
\]

where \( A \equiv f \left( \frac{H}{K} \right) \). This returns us to the \( AK \) form.

A two-sector model can have different production functions for the creation of the two types of capital goods. This eliminates the restriction that they be perfect substitutes and moves away from the \( AK \) setting. In a two-sector model different human and physical capital intensities can be incorporated in the production of the two types of capital. This can make it consistent with the observation that human capital production tends to be more intensive in human capital—for example, through the requirement for skilled teaching staff, and so forth.

The next two approaches focus on inputs other than labor. If output depends on labor use and a range of other inputs, technological progress can take the form of the introduction of new inputs into the production function without any of the old inputs being dropped. The additional inputs allow production to increase, since the expansion of the input range prevents the level of use of any one of the inputs becoming too large relative to the labor input. An alternative view of technological progress is that it takes the form of an increase in the quality of inputs. Expenditure on research and development results in better quality inputs that are more productive. Over time old inputs are replaced by new inputs, and total productivity increases. Firms are driven to innovate in order to exploit the position of monopoly that goes with ownership of the latest innovation. This is the process of “creative destruction” that was seen by Schumpeter as a fundamental component of technological progress.

A special case of this approach, and the one on which we will focus, is to use a public good as the additional input in the production function. This can allow for constant returns in the private inputs to production and also constant returns to private capital, provided that the level of the public good is raised to match. The analytical details of this model are described below because it is a useful vehicle for thinking about the channels through which public expenditure can have an impact on growth.

A final approach to endogenous growth is to assume that there are externalities among firms that operate through learning by doing. Investment by a firm leads to
parallel improvements in the productivity of labor as new knowledge and techniques are acquired. Moreover this increased knowledge is a public good so the learning spills over into other firms. Spillovers make the level of knowledge, and hence labor productivity, dependent on the aggregate capital stock of the economy. Decreasing returns to capital for a single firm (for a given use of labor) then translate into constant returns for the economy.

The common property of these models of endogenous growth is that there are growth-related choices that can be influenced by policy. The government can encourage (or discourage) investment in human capital through subsidies to training or the tax treatment of the returns. Subsidies to research and development can encourage innovation, as can the details of patent law. From among these many possibilities, for the remainder of the chapter we focus on the interaction of taxation and economic growth.

24.3.2 Government Expenditure

Endogenous growth can arise when capital and labor are augmented by additional inputs in the production function. One case of particular interest for understanding the link between government policy and growth is when the additional input is a public good financed by taxation. The existence of a public input provides a positive role for public expenditure and a direct mechanism through which policy can affect growth. This opens a path to an analysis of whether there is a sense in which an optimal level of public expenditure can be derived in a growth model.

A public input can be introduced by assuming that the production function for the representative firm at time $t$ takes the form

$$ Y_t = AL_t^{1-\alpha} K_t^\alpha G_t^{1-\alpha}, $$

(24.40)

where $A$ is a positive constant and $G_t$ is the quantity of the public input. The structure of this production function ensures that there are constant returns to scale in $L_t$ and $K_t$ for the firm given a fixed level of the public input. Although returns are decreasing to private capital as the level of capital is increased for fixed levels of labor and public input, there are constant returns to scale in public input and private capital together. For a fixed level of $L_t$, this property of constant returns to scale in the other two inputs permits endogenous growth to occur.

It is assumed that the public input is financed by a tax on output. Assume that capital does not depreciate in order to simplify the derivation. The profit level of the firm is

$$ \pi_t = [1 - \tau] AL_t^{1-\alpha} K_t^\alpha G_t^{1-\alpha} - r_t K_t - w_t L_t $$

(24.41)
where $r_t$ is the interest rate, $w_t$ the wage rate, and $\tau$ the tax rate. From this specification of profit, the choice of capital and labor by the firm satisfy

$$[1 - \tau] \alpha AL_t^{1-a} K_t^{a-1} G_t^{1-a} = r_t$$

(24.42)

and

$$[1 - \tau][1 - \alpha] AL_t^{-a} K_t^a G_t^{1-a} = w_t.$$

(24.43)

The government budget constraint requires that tax revenue equal the cost of the public good provided, so that

$$G_t = \tau Y_t.$$

(24.44)

Now assume that labor supply is constant at $L_t = L$ for all $t$. Without the public input, it would not be possible, given this assumption, to sustain growth because the marginal product of capital would decrease as the capital stock increases. With the public input, growth can now be driven by a joint increase in private and public capital, even though labor supply is fixed. From (24.40) and (24.44) the level of public input can be written as

$$G_t = [\tau A]^{1/\alpha} L^{[1-a]/\alpha} K_t.$$

(24.45)

This result can be substituted into (24.42) to obtain an expression for the interest rate as a function of the tax rate:

$$r_t = [1 - \tau] \alpha A^{1/\alpha} [L \tau]^{[1-a]/\alpha}.$$

(24.46)

The economy’s representative consumer is assumed to have preferences described by the utility function

$$U = \sum_{t=1}^{\infty} \beta^t C_t^{1-\sigma} - 1 \over 1 - \sigma.$$

(24.47)

This specific form of utility is adopted to permit an explicit solution for the steady state. The consumer chooses the path $\{C_t\}$ over time to maximize utility. The standard condition for intertemporal choice must hold for the optimization, so the ratio of the marginal utilities of consuming at $t$ and at $t+1$ must equal the gross interest rate. Hence

$$\frac{\partial U/\partial C_t}{\partial U/\partial C_{t+1}} \equiv \frac{C_t^{-\sigma}}{\beta C_{t+1}^{-\sigma}} = 1 + r_{t+1}.$$

(24.48)
By solving for $\frac{C_{t+1}}{C_t}$ and then subtracting $\frac{C_t}{C_t}$ from both sides of the resulting equation, this optimality condition can be written in terms of the growth rate of consumption as

$$\frac{C_{t+1} - C_t}{C_t} = \left[\beta \left(1 + r_{t+1}\right)\right]^{1/\sigma} - 1. \quad (24.49)$$

Finally, substituting the solution (24.46) for the interest rate allows the growth rate of consumption to be related to the tax rate by

$$\frac{C_{t+1} - C_t}{C_t} = \beta^{1/\sigma} \left[1 + [1 - \tau] \alpha A^{1/\alpha} \left[L \tau \right]^{1-\alpha}/\alpha \right]^{1/\sigma} - 1. \quad (24.50)$$

The result in (24.50) demonstrates the two channels through which the tax rate affects consumption growth. First, taxation reduces the growth rate of consumption through the term $1 - \tau$, which represents the effect on the marginal return of capital reducing the amount of capital used. Second, the tax rate increases growth through the term $\tau^{(1-\alpha)/\alpha}$, which represents the gains through the provision of the public input.

Further insight into these effects can be obtained by plotting the relationship between the tax rate and consumption growth. This is shown in figure 24.5 under the assumption that $A = 1, L = 1, \alpha = 0.5, \beta = 0.95, \text{ and } \sigma = 0.5$. The figure displays several notable features. First, for low levels of the public input, growth is negative, so a positive tax rate is required for there to be consumption growth. Second, the relationship between growth and the tax rate is nonmonotonic: growth initially increases with the tax rate, reaches a maximum, and then decreases. Finally, there is a tax rate that maximizes the growth rate of consumption. Differentiating (24.50) with respect to $\tau$ gives the tax rate that maximizes consumption growth as

$$\tau = 1 - \alpha. \quad (24.51)$$

For the values in the figure, this optimal tax rate is $\tau = 0.5$. To see what this tax rate implies, observe that

$$\frac{\partial Y_t}{\partial G_t} = [1 - \alpha] \frac{Y_t}{G_t} = 1, \quad (24.52)$$

since $G_t = \tau Y_t$ and $\tau = 1 - \alpha$. Hence the tax rate that maximizes consumption growth ensures that the marginal product of the public input is equal to 1, which is also its marginal cost.

This model reveals a positive role for government in enhancing growth through the provision of a public input. It illustrates a sense in which there can be an optimal level of government. Also, if the size of government becomes excessive, it reduces the rate
24.4 Policy Reform

The analysis of section 24.2.2 has demonstrated the surprising and strong result that the long-run tax rate on capital should be zero. Although the derivation was undertaken for an exogenous growth model, the result also applies when growth is endogenous. The basic intuition that the intertemporal allocation should not be distorted applies equally in both cases. This is an important conclusion, since it contrasts markedly with observed tax structures. For example, in 2002 the top corporate tax rate was 40 percent in the United States, 30 percent in the United Kingdom, and 38.4 percent in Germany. Although Ireland was much lower at 16 percent, the OECD average was 31.4 percent.

This divergence of the observed tax rates from the theoretically optimal rate raises the possibility that a reform of the actual system can raise the rate of economic growth and the level of welfare. This question has been tested by simulating the response of
model economies to policy reforms involving changes in the tax rates on capital and labor. Such studies have provided an interesting range of conclusions that are worth close scrutiny.

Before discussing these results, it is helpful to clarify the distinction between the effect of a change in taxation on the level of output and its effect on the rate of growth of output. This distinction is illustrated in figure 24.6, which shows three different growth paths for the economy. Paths 1 and 2 have the same rate of growth—the rate of growth is equal to the gradient of the growth path. Path 3, which has a steeper gradient, displays a faster rate of growth.

Assume that at time $t_0$ the economy is located at point $a$ and, in the absence of any policy change, will grow along path 1. Following this path, it will arrive at point $b$ at time $t_1$. The distinction between level and growth effects can now be described. Consider a policy change at time $t_0$ that moves the economy to point $c$ with consequent growth along path 2 up to point $d$ at time $t_1$. This policy has a level effect: it changes the level of output but not its rate of growth. Alternatively, consider a different policy that causes the economy to switch from path 1 to path 3 at $t_0$, so at time $t_1$ it arrives at point $e$. This change in policy has affected the rate of growth but not (at least initially) its level. Of course, output eventually achieves a higher level because of the higher growth rate. This second policy has a growth effect but no level effect. Most policy changes will have some combination of level and growth effects.
Chapter 24: Economic Growth

The basic setting for the simulation analysis is an endogenous growth model with both physical and human capital entering the production function. The consumption side is modeled by a single, infinitely lived representative consumer who has preferences represented by the utility function

\[ U = \frac{1}{1-\sigma} \sum_{t=0}^{\infty} \beta^t \left[ C_t L_t^\alpha \right]^{1-\sigma}, \]  

where \( C_t \) is consumption and \( L_t \) is leisure. Alternative studies adopt different values for the parameters \( \alpha \) and \( \sigma \). The second area of differentiation between studies is the range of inputs into the production process for human capital, in particular, whether it requires only human capital and time or whether it also needs physical capital. The analytical process is to specify the initial tax rates, which usually take values close to the actual position in the United States, and then calculate the initial growth path. The tax rates are changed and the new steady-state growth path calculated. The two steady states are then contrasted with a focus placed on the change in growth rate and in levels of the variables.

Figure 24.7 summarizes some of the policy experiments and their consequences. The experiment of Lucas (1990) involves elimination of the capital tax with an increase in the labor tax to balance the government budget. This policy change has virtually no growth effect (it is negative but very small) but a significant level effect. In contrast, King and Rebelo (1990) and Jones et al. (1993) find very strong growth and level effects. King and Rebelo consider the effect of an increase in the capital tax by 10 percent, whereas Jones et al. mirror Lucas by eliminating the capital tax. What distinguishes the King and Rebelo analysis is that they have physical capital entering into the production of human capital. Jones et al. employ a higher value for the elasticity of labor supply than other studies. The model of Pecorino (1993) has the feature that capital is a separate commodity to the consumption good. This permits different factor intensities in the production of human capital, physical capital, and the consumption good. Complete elimination of the capital tax raises the growth rate, in contrast to the finding of Lucas.

The importance of each of the elements in explaining the divergence between the results is studied in Stokey and Rebelo (1995). Using a model that encompasses the previous four, they show that the elasticity of substitution in production matters little for the growth effect but does have implications for the level effect—with a high elasticity of substitution, a tax system that treats inputs asymmetrically will be more
Author Features Utility parameters Initial tax rates and growth rate Final position Additional observations

Lucas (1990) Production of human capital did not require physical capital \[ \sigma = 2 \] \[ \alpha = 0.5 \] Capital 36% Labor 40% Growth 1.50% Capital 0% Labor 46% Growth 1.47% 33% increase in capital stock 6% increase in consumption

King and Rebelo (1990) Production of human capital requires physical capital (proportion = 1/3) \[ \sigma = 2 \] \[ \alpha = 0 \] Capital 20% Labor 20% Growth 1.02% Capital 30% Labor 20% Growth 0.50% Labor supply is inelastic

Jones, Manuelli and Rossi (1993) Time and physical capital produce human capital \[ \sigma = 2 \] \[ \alpha = 4.99 \] \[ \alpha = \text{calibrated given } \sigma \] Capital 21% Labor 31% Growth 2.00% Capital 0% Labor 0% Growth 4.00% 10% increase in capital stock 29% increase in consumption

Pecorino (1993) Production of human capital requires physical capital \[ \sigma = 2 \] \[ \alpha = 0.5 \] Capital 42% Labor 20% Growth 1.51% Capital 0% Labor 0% Growth 2.74% Capital and consumption different goods, consumption tax replaces income taxes

Figure 24.7 Growth effects of tax reform

distortionary. The elimination of the distortion then leads to a significant welfare increase. The important features are the factor shares in production of human capital and physical capital, the intertemporal elasticity of substitution in utility, and the elasticity of the labor supply. Stokey and Rebelo conclude that the empirical evidence provides support for values of these parameters that justify Lucas’s claim that the growth effect is small.

A range of estimates have been given for the effects of taxation on growth involving several different policy experiments. Some of the models predict that the growth effect is insignificant; others predict it could be very significant. What distinguishes the models are a number of key parameters, particularly, the share of physical capital in human capital production, the elasticities in the utility function, and the depreciation rates. In principal, these could be isolated empirically, and a firm statement of the size of the growth effect given. To do so, and thus claim an “answer,” would be to overlook several important issues about the restrictiveness of the model. Moreover it would not be justifiable to provide an answer without consulting the empirical evidence. Tax rates have grown steadily over the last century in most countries, so there should be ample
evidence for determining the actual effect. Consequently the next section considers empirical evidence on the effect of taxation.

24.5 Empirical Evidence

We have presented two theoretical perspectives on the link between taxation and growth. The endogenous growth model with a public good as an input provided a positive channel through which taxation could raise growth. The relationship was not monotonic because increases in the tax rate above the optimum would reduce the growth rate. In practice, economies could be located on either side of the optimum. Similarly the evidence from the simulations provides a wide range of estimates for the effect of taxation on economic growth from negligible to significant. Since the theory is so inconclusive, it is natural to turn to the empirical evidence.

At first glance a very clear picture emerges from this: tax revenue as a proportion of gross domestic product has risen significantly in all developed countries over the course of the last century, but the level of growth has remained relatively stable. This suggests the immediate conclusion that in practice, taxation does not affect the rate of growth. Data to support this claim is displayed in figures 24.8 and 24.9. Figure 24.8 plots the growth rate of US gross domestic product and federal government tax revenue as a percentage of gross domestic product since 1930. Trend lines have been fitted to

![Figure 24.8](image-url)

**Figure 24.8**
US tax and growth rates
the data series using ordinary least squares regression to show the trend over time. The
two trend lines show a steady rise in taxation (the upper line) and a very slight decline
in the growth rate (the lower line). Although the variance of the growth process reduces
after 1940, statistical tests on US data have found no statistical difference between the
average rate of growth prior to 1942 and after 1942. The data for the United Kingdom
in figure 24.9 tell a very similar story. The trend lines show an increase in taxation but,
in contrast to the United States, an increase in the rate of growth.

The message of these figures is compelling but should be considered carefully. First,
a contrast between tax rate and growth rate across time cannot answer the counterfactual
question: If taxes had been lower, would growth have been higher? To do so requires a
study involving countries with different regimes. Second, there are substantive issues
that have to be resolved about the definition of the tax rate as should be used in any
such comparison.

To understand the problem of definition, consider figure 24.10 which illustrates a
typical progressive income tax. There is an initial tax exemption up to income level \( Y_1 \),
then a band at tax rate \( t_1 \), and a final band at rate \( t_2, t_2 > t_1 \). Notice in the figure how the
marginal rate of tax differs from the average rate of tax. For instance, at income \( \hat{Y} \) the
marginal rate is one minus the gradient of the graph while the average rate is one minus
the gradient of the ray to the graph (shown by the dashed line). With a progressive tax
system, the marginal rate is always greater than the average rate.
The data displayed in figures 24.8 and 24.9 uses tax revenue as a fraction of gross domestic product to measure the tax rate. This measure captures the average rate of tax. However, what matters for economic behavior is the marginal tax rate—the decision on whether to earn additional income depends on how much of that income can be retained. This suggests that the link between growth and taxation should focus more on how the marginal rate of tax affects growth.

The difficulty with undertaking the analysis comes in determining what the marginal tax rate actually is. Figure 24.10 illustrates this problem: the marginal tax rate is 0, \( t_1 \), or \( t_2 \) depending on the income level of the consumer. In practice, income tax systems have several different levels of exemption (e.g., married and single persons allowances) and several marginal rates, and they interact with social security taxes and with the benefit system. All this makes it difficult to assign any unique value to the marginal rate of a tax. The same comments apply equally to corporation tax, which has exemptions, credits, and depreciation allowances, and to value-added taxation, which has exemptions, zero-rated goods, and lower rated goods. In brief, the rate of growth should be related to the marginal rate of tax, but the latter is an ill-defined concept.

Given these preliminaries, it is now possible to review the empirical evidence. The strongest empirical link between taxation and growth was reported by Plosser (1993). Plosser regressed the rate of growth of per capita gross domestic product on the ratio of
income taxes to gross domestic product for the OECD countries and found a significant negative relationship. The limitation of this finding is that the OECD countries differ in their income levels and income has been found to be one of the most significant determinants of growth. Taking account of this, Easterly and Rebelo (1993) showed that the negative relationship all but disappears when the effect of initial income is accounted for.

Easterly and Rebelo extend this analysis by using several different measures of the marginal tax rate in regressions involving other determinants of growth, notably initial income, school enrollments, assassinations, revolutions, and war casualties. In response to some of the difficulties already noted, four different measures of the marginal tax rate are used: statutory taxes, revenue as a fraction of gross domestic product, income-weighted marginal income tax rates, and marginal rates from a regression of tax revenue on tax base. From a number of regressions involving these variables, Easterly and Rebelo conclude: “The evidence that tax rates matter for economic growth is disturbingly fragile.”

A very similar exercise is undertaken in Mendoza, Milesi-Ferretti, and Asea (1997). The clear finding is that when initial GDP is included in the regressions, the tax variable is insignificant. Evidence contrary to this is presented in Leibfritz, Thornton, and Bibbee (1997). Their regression of average growth rates for OECD countries over the period 1980 to 1995 against three measures of the tax rate (average tax rate, marginal tax rate, and average direct tax rate) showed that a 10 percent increase in tax rates would be accompanied by a 0.5 percent reduction in the rate of growth, with direct taxation reducing growth marginally more than indirect taxation.

One possible route out of the difficulties of defining the appropriate tax rate is to adopt a different method of determining the effect of fiscal policy. Engen and Skinner (1996) label the regressions described above as “top-down,” since they work with aggregate measures of taxation. Instead of doing this, they propose a “bottom-up” method that involves calculating the effect of taxation on labor supply, investment, and productivity, and then summing these to obtain a total measure. Doing this suggests that a cut of 5 percent in all marginal rates of tax and 2.5 percent in average rates would raise the growth rate by 0.22 percent.

An alternative line of literature—Barro (1991), Dowrick (1993), and de la Fuente (1997)—has considered the more general issue of how fiscal policy has affected growth. In particular, the relationship of growth to the composition and level of public sector spending is investigated. The results of de la Fuente show that if public spending (measured as the share of total government expenditure in gross domestic product) increases, growth is reduced (an increase in government spending of 5 percent of
gross domestic product reduces growth by 0.66 percent), whereas an increase in public investment will raise growth. There are four significant points to be made about these findings. First, government spending may just be a proxy for the entire set of government non-price interventions—including, for example, employment legislation, health and safety rules, and product standards—and that it may be these, not expenditure, that actually reduce growth. Second, since the share of public spending in gross domestic product is very closely correlated to the average tax rate, it is not clear which hypothesis is being tested.

The final points are more significant. Levine and Renelt (1992) have shown that the finding of a negative relationship is not robust to the choice of conditioning variables. Finally, as noted by Slemrod (1995), the method of the regressions is to use national income, $Y$, as the left-hand-side variable and government expenditure, $G$, as the right-hand-side variable. In contrast, economic theory usually views the causality as running in the opposite direction: government expenditure is seen as being determined by the preferences of the population as expressed through the political system. An extreme version of this view is captured in Wagner’s law, which relates government expenditure to national income via the income elasticity of demand for government-provided goods and services. If $Y$ (or the growth of $Y$) and $G$ are related via an equilibrium relationship, then a simple regression of $Y$ on $G$ will not identify this.

This review of the empirical evidence leads to the following observation. A visual inspection of tax rates and growth rates suggests that there is little relationship between the two. This is weak evidence, but it does find support in some more detailed investigations in which regression equations that include previously identified determinants of growth, especially initial income, reveal that tax rates are insignificant as an explanatory variable. Other regressions find a small but significant tax effect. All these results are hampered by the difficulties in actually defining marginal rates of tax and in their lack of an equilibrium relationship.

24.6 Conclusions

Growth is important. Without growth, living standards will stagnate, if not decline. The effect of even small changes in the growth rate can be dramatic. With a growth rate of 2 percent it takes 35 years to double the level of income, but at a rate of 5 percent it takes just 14 years. Even if economic policy only succeeds in increasing the growth rate from 2 to 3 percent, it will reduce the time taken to double income by 12 years. The cumulative effect of a policy that affects the growth rate will eventually dominate anything achieved by a policy that affects only the level of economic variables.
In an exogenous growth model the economy must eventually reach a limit to its growth unless there is technical progress. The effects of policy are limited in this form of model because in the long run policy cannot affect the growth rate. Nevertheless, these models provide some insight into what policy must achieve in order for it to have a lasting effect on economic growth. In particular, the exogenous growth model provides a simple setting for demonstrating the important result that efficiency requires that the tax rate on capital income be zero in the long run.

The limitations of the exogenous growth model led to the development of theories of endogenous growth. The literature on endogenous growth has provided a range of mechanisms by which taxation can affect economic growth. This chapter has described the range of models and has discussed the results that have been obtained. In quantitative terms, a wide range of theoretical predictions arise for the size of the tax effect, from the insignificant to considerable. The size of the growth rate effect depends just about equally on the structure of the model and on the parameter values within the model. The production process for human capital is also critical, as are the elasticities in the utility function and the rates of depreciation. A fair summary is that the theoretical models introduce a range of issues that must be considered but do not provide any convincing or definitive answers.

The conclusions of the empirical evidence are not quite as diverse as for the theory. Although there are some disagreements, the picture that emerges is that the effect of taxation, if there is any at all, is relatively minor. However, the estimates have to be judged by taking account of the difficulty of defining the appropriate measure of the tax rate and the choice of appropriate regressors. When these problems have been overcome the effect of taxation on growth may prove to be significant but that is unlikely. As far as policy is concerned, this is a reassuring conclusion because it removes the need to be overly concerned about growth effects when tax reforms are planned.

Further Reading

The classic summary of exogenous growth theory can be found in:
The data in table 24.1 are taken from:
Three detailed surveys of growth theory that differ in their emphasis are:
Chapter 24: Economic Growth


The proof that the optimal capital tax rate is zero can be found in:


The model of endogenous growth with a public input was first proposed by:


The simulations of the effect upon growth of changes in the tax rate are taken from:


Some of the data in figure 24.9 are from:


For empirical evidence on the effect of taxation on economic growth, see:


### Exercises

24.1 What is the difference between exogenous growth and endogenous growth? Why does the latter give a larger role for public policy?

24.2 What are the main features of a steady state in the exogenous growth model?

24.3 In the exogenous growth model what are the effects on the steady-state consumption and capital–labor ratio of an increase in the savings rate and of an increase in the population growth rate? Give a graphical illustration of each effect on the steady state.

24.4 What are the determinants of the Golden Rule capital–labor ratio in the exogenous growth model?

24.5 What are the three determinants of growth in the exogenous growth model?

24.6 In the long-run steady state of the (Solow) exogenous growth model, growth in aggregate output, consumption, and investment is determined by exogenous growth in the labor force. True or false?

24.7 Explain how in the (Solow) exogenous growth model, output and consumption per worker converge in the long run to steady-state levels. Show that output per worker increases in the long run when the savings rate increases or when the population growth rate decreases. Check if these two predictions are consistent with the data.

24.8 In the Solow growth model, assume that the aggregate production function is \( Y = zF(K, L) \), where \( Y \) is total output, \( z \) is total productivity, \( K \) is the total capital input, and \( L \) is the total labor input. Suppose that the Cobb–Douglas form for the production function is \( F(K, L) = K^\alpha L^{1-\alpha} \), where \( 0 < \alpha < 1 \).

    a. Find an expression for the output per worker as a function of the capital–labor ratio.
    
    b. Suppose that the marginal product of capital increases for each level of capital, given the labor input level. What is the effect of this capital productivity increase on the aggregate production function? Show graphically the impact of this productivity increase on the steady-state capital–labor ratio and output per worker. Explain briefly.
24.9 Consider the Solow growth model with Cobb–Douglas production function
\[ Y = z[K^\alpha L^{1-\alpha}] \]
where \(Y\) is total output, \(z\) is total productivity, \(K\) is the total capital input, and \(L\) is the total labor input.

a. Show that in the competitive equilibrium, \(\alpha\) is the fraction of national income that goes to the remuneration of capital and \(1 - \alpha\) is the fraction that goes to the remuneration of labor.

b. Use the (average) labor share in national income from your own country for the period 1990 to 2000 to find the value of \(\alpha\).

c. Use your answer to part b to calculate the “Solow residual” \(z\) using the aggregate output, capital, and labor (\(Y\), \(K\), and \(L\)) from your country for each year in the period 1990 to 2000. How has this measure of total productivity evolved over time in your country? Is there a productivity slowdown? Explain the possible reasons of this evolution (inventions, government regulations, price of energy, etc.).

24.10 Consider an economy with population size \(N_t\) at time \(t\) and full employment \(L_t = N_t\). The consumption levels of a person born at time \(t\) when young and old are given respectively by
\[ c^1_t = \lambda w\, , \]
\[ c^2_t = (1 + r_{t+1})y^1_t, \]
where \(\lambda \in (0, 1)\). Production is determined by a Cobb–Douglas technology
\[ Y_t = A_t F(K_t, L_t) = A_t K_t^\alpha L^{1-\alpha}. \]

Technological progress and demographic growth are given by
\[ A_{t+1} = (1 + g)A_t, \]
\[ N_{t+1} = (1 + n)N_t. \]

a. Let \(k_t = K_t / N_t\) be the capital stock per worker and \(y_t = Y_t / N_t\) be the output per worker. Calculate the equilibrium stock of capital per worker \(k^*_t(r_t)\) for a competitive economy.

b. Show that the equilibrium wage is given by
\[ w^*_t = (1 - \alpha)A_t \left( \frac{\alpha A_t}{r_t} \right)^{\alpha/(1-\alpha)} = (1 - \alpha)y_t. \]

What are the effects of a change in the interest rate \(r_t\) and the productivity factor \(A_t\) on the equilibrium wage rate \(w^*_t\) and the equilibrium stock of capital \(k^*_t\)?

c. Show that output per capita is given by
\[ y_t = w_t + r_t k_t. \]

d. Consider a steady state with \(r_t = r\). Calculate the growth rate of capital \(g_k = \frac{k_{t+1} - k_t}{k_t}\) and the growth rate of output \(g_y = \frac{Y_{t+1} - Y_t}{Y_t}\). Derive the growth rate of the wage.
e. Suppose \( n g = 0 \). Using the consumption decisions given above, calculate the aggregate saving \( S_1^t \) of the young generation at time \( t \) and the dissaving \( S_{t-1}^2 \) of the old generation at time \( t \). Show the following equivalence:

\[
S_1^t > S_{t-1}^2 \iff n + g > r.
\]

Explain this equivalence.

24.11 Assume that output per unit of labor is related to the capital–labor ratio by

\[
y_t = \rho \log(k_t),
\]

and that saving is a fixed proportion, \( s \), of output.

a. What is the capital accumulation condition?

b. Derive the equation characterizing the steady-state capital–labor ratio.

c. For the parameter values \( k_0 = 0.5, s = 0.1, \delta = 0.05, n = 0, \) and \( \rho = 2 \), plot the capital–labor ratio and the consumption–labor ratio as functions of time.

d. What is the effect of an increase in \( \rho \) on the capital–labor ratio?

e. Repeat part c, starting the economy above the steady-state capital–labor ratio. Explain the observed growth path.

24.12 If the production function is given by \( y_t = \log(k_t) \), \( n = 0.05 \), and \( \delta = 0.1 \), and if saving is a fixed fraction of output, calculate the steady-state level of consumption per unit of labor. Plot the steady-state level of consumption per unit of labor as a function of \( s \). Show that it is maximized when \( \frac{1}{k} = 0.15 \).

24.13 Consider the Solow growth model as described in the preceding exercise and include government spending. The government purchases \( G = Lg \) units of consumption good in the current period with \( g \) a positive constant. The government finances its public purchases with a lump-sum tax on consumers, where \( T = G \) denotes the balanced budget total taxes. Suppose that consumers spend a fraction \( 1 - s \) of their disposable income (where \( 0 < s < 1 \)) so that

\[
C = [1 - s](Y - T).
\]

a. Show in a diagram how the capital–labor ratio is determined and what its value is in a steady state.

b. How many steady states are possible?

c. Concentrate on the steady state with the highest capital–labor ratio. What is the impact of an increase in public spending per worker \( g \) on the capital–labor ratio and output per worker in the steady state? What are the effects on aggregate output, consumption, and investment? What is the effect on the growth rate? Explain briefly.

24.14 Can the Solow growth model account for the persistence in income inequality across countries due to persistent difference in population growth rates? Explain and discuss.

24.15 Let \( N_t \) denote the number of workers. There are two groups of workers, \( A \) and \( B \), with \( N_t = N_t^A + N_t^B \). Every worker inelastically supplies one unit of labor. Every worker in group \( A \) (resp. \( B \)) achieves an income of \( w^A \) (resp. \( w^B \)) and

\[
N_t^A w_t^A + N_t^B w_t^B = N_t w_t.
\]
Chapter 24: Economic Growth

The wage ratio is defined as \( \frac{w^A_t}{w^B_t} = \phi \in (0, \infty) \). The proportion of group A in the population is \( \mu \in (0, 1) \).

a. Show that the wage of each group is given by

\[
\begin{align*}
    w^A_t &= \frac{\phi w_t}{\mu \phi + (1 - \mu)} , \\
    w^B_t &= \frac{w_t}{\mu \phi + (1 - \mu)} .
\end{align*}
\]

b. Assume a Cobb–Douglas production function

\( y_t = f(k_t) = k_t^\alpha \),

where \( \alpha \in (0, 1) \), \( y_t \) denotes output per worker, and \( k_t \) is the capital stock per worker. Equilibrium wage and interest rate are set equal to productivities. Show that average wage \( w_t \) increases with \( k_t \).

c. Assume that the saving rates \( \lambda^i \) are constant for both groups with \( 0 < \lambda^A < \lambda^B < 1 \). The accumulation of capital is given by

\( K_{t+1} = S_t \),

where \( S_t \) is the aggregate savings of both groups of consumers when young. Show that the steady-state stock of capital per worker is

\[
    k^* = \left[ \left( \frac{1 - \alpha}{1 + n} \right) \lambda^\frac{1}{(1 - \alpha)} \right].
\]

where

\[
    \lambda = \frac{\phi \mu \lambda^A + (1 - \mu) \lambda^B}{\mu \phi + (1 - \mu)} .
\]

d. Show that the distribution of income does not affect \( k^* \) if \( \lambda^A = \lambda^B \). Explain this result.

e. The social planner seeks to maximize the average steady-state income \( w \) by redistributing income to change \( \phi \). What is the optimal redistribution rule (in terms of \( \phi \)) as a function of the saving rates \( \lambda^A, \lambda^B \)? Explain the result.

24.16 Consider the previous exercise and suppose that saving rates of the two groups of consumers satisfy

\[
    0 < \lambda^A < \left( \frac{\alpha}{1 - \alpha} \right) \left( \frac{1 + n}{n} \right) < \lambda^B < 1 .
\]

Show that there exists a unique redistribution rule \( w^A_t / w^B_t = \phi^* \in (0, \infty) \) that achieves the Golden-rule stock of capital. You can use a graphical argument.

24.17 Consider the Ramsey growth model with utility function

\[
    U = \sum_{t=0}^{\infty} \beta^t \left[ \log(C_t) + \log(1 - L_t) \right]
\]
and production function

\[ Y_t = K_t^\alpha L_t^{1-\alpha}. \]

Assume that there are no taxes.

a. Derive the conditions describing a steady state.
b. Solve these conditions to obtain the steady-state values of \( C, K, \) and \( L. \)
c. What is the effect on the steady state of an increase in \( \beta? \)

24.18 Repeat the previous exercise but include a tax on interest income. Graph the steady-state capital–labor ratio against the tax rate. Comment on the results.

24.19 Consider the endogenous growth model with a public input as described in the text with the parameter values \( A = 1, L = 1, \beta = 0.95, \) and \( \sigma = 1. \) Solve for the growth rate of consumption as a function of \( \alpha \) and graph for various values of \( \alpha. \) What is the effect of increasing \( \alpha? \) Comment on your findings.

24.20 What are the determinants of growth in production and consumption in the endogenous growth model?

24.21 “If the government can increase the efficiency of human capital accumulation, it could persistently raise the rate of economic growth in the endogenous growth model.” True or false? Explain your answer.

24.22 “Profit is the incentive to innovate. Government intervention can only reduce it, and therefore stifle the innovation process.” Discuss the expected relation between taxation and growth in the light of this statement.

24.23 Consider a tax system where the first $5,000 of income is tax free, the next $10,000 is taxed at 25 percent, and all income above $15,000 is taxed at 50 percent. There is a population of consumers with who earn wages between $2 and $32 per hour. Wages are uniformly distributed in the population. All consumers have preferences described by \( U = \log(C) - \frac{L}{1000}. \)

a. Find the mean average tax rate for the population.
b. Find the mean marginal tax rate for the population. How do these mean rates differ from the actual rates?

24.24 Using the results of the previous exercise, find the mean labor-supply response to an increase in the 25 percent rate to 26 percent. Now consider a consumer of mean wage rate facing the mean tax rate found in part b above. What increase in this mean rate of tax has the same effect as the 1 percent increase in the 25 percent rate? Does the use of the mean marginal rate provide a good indication of the labor supply response?

24.25 In the endogenous growth model explain why per capita income levels do not necessarily converge across countries, even when countries are initially identical except for human capital levels.
25.1 Introduction

The Flood Control Act of 1936 established a commitment by the US federal government to protect people and property on approximately 100 million acres of land. The act required that the US Corps of Engineers carry out projects for the improvement of the waterway system when the total benefits of a project exceed the costs of that project. Subsequently the US Congress has authorized the Corps of Engineers to construct hundreds of miles of levees, flood walls, and channel improvements and approximately 375 major reservoirs. According to Arnold’s 1988 history of the 1936 Act, these projects “have saved billions of dollars in property damage and protected hundreds of thousands of people from anxiety, injury, and death. They stand today as one of the more significant marks of our technical skill and humane spirit.” The requirement that projects have to pass the test of offering benefits in excess of costs has meant that the Corps of Engineers had to create systematic methods for measuring the items in the calculation.

A project involves a flow of benefits and a flow of costs. Cost–benefit analysis (CBA) estimates and totals up the monetary values of the benefits and costs to the community of projects to establish whether they are worthwhile. CBA has traditionally been applied to public projects such as new motorways, by-passes, dams, tunnels, bridges, flood relief schemes, and new power stations. Cost–benefit analysis is widely used in project and policy evaluations in many OECD countries, including Australia (where CBAs are termed Regulatory Impact Assessments), the United States, the United Kingdom, and the Nordic countries. The European Commission conducts CBA for all new EU Directives, and the World Bank and the regional development banks in Asia, Africa, and Latin America use CBA in their project evaluations.

It is possible to view most forms of policy choice as the selection of a project. The idea of a project can be interpreted as any policy that affects the economy. This allows the basic principles of CBA to be applied to a very wide range of projects or programs. Some usual examples are the introduction of a public health program (e.g., the mass immunization of children using new drugs), an investment in a new rail safety system, or the opening of a new railway line. CBA can also be used in assessing the costs and benefits of introducing congestion charges for motorists in a metropolitan area, or the costs and benefits of an employment program designed to reduce long-term unemployment. CBA was even used during the recent inquiry into genetically
modified foods. Increasingly the principles of cost–benefit analysis are being applied to evaluate the returns on investment in environmental projects such as wind farms and the development of other sources of renewable energy.

Because financial resources are scarce, CBA allows different projects to be ranked according to those that provide the highest expected net gains in social welfare. This is particularly important given the limitations of government spending. The idea of this economic accounting originated with Jules Dupuit, a French engineer whose 1848 article is still worth reading. The British economist, Alfred Marshall, formulated some of the formal concepts that are at the foundation of CBA in his *Principles of Economics* (1890). The practical development of CBA in the United States began in the 1930s, but it wasn’t until about twenty years later in the 1950s that economists tried to provide a rigorous, consistent set of methods for measuring benefits and costs in deciding whether a project is worthwhile.

A formal theory of cost–benefit analysis has now been developed. There are some technical issues of CBA that are not entirely resolved—and these are discussed later—but the fundamental theory presented in this chapter is well established and widely accepted. The chapter begins with the basic principles that guide the practical application of CBA. These principles are straightforward, but CBA application can be tricky and controversial. Most of the controversy can be related to valuing human life and intangible environmental benefits in the estimation of costs and benefits. The choice of the discount rate to balance the consequences for current and future generations when the project has long-lasting effects and the relative weights given to poor and rich when a project has distributional implications are also controversial points. The discussion of the many points of difficulty for CBA is followed by a more comprehensive analysis of welfare measurement, valuation of life, and choice of discount rate. Underlying the argument in favor of CBA as a policy tool is that its use ensures only projects that raise welfare will be implemented. The final part of the chapter explores the theory that links CBA to welfare, and focuses on the use of shadow prices and their relationship to market prices.

25.2 What Is Cost–Benefit Analysis?

Cost–benefit analysis is basically an appraisal technique that aims to place monetary values on all the benefits and costs arising from a project, and then determine whether the total value of the project is positive or negative. Essentially the process of CBA is a comparative one, so that judgments can be made about which projects from a limited
choice should be given the go-ahead. It has numerous potential applications though there are inherent difficulties in assigning monetary valuations to some of the effects of a project.

25.2.1 Simple Example

Let us start with an example. Suppose that you are the production manager of a firm and you are proposing the purchase of a $1 million machine to increase output. You know you need some facts to support your proposal before you can present it to the firm’s president, so you decide to run the numbers and do a cost–benefit analysis. You list the additional benefits that the machine will deliver: with the new machine, 100 more units could be produced each hour, the three workers currently doing the work by hand could be replaced, all units produced will be of higher quality because they will be more uniform. You are convinced these outweigh the additional costs: there is a cost to purchase the machine, and it will consume some electricity. Any other costs will be insignificant.

You calculate the selling price of the 100 additional units per hour multiplied by the number of production hours per month. Two percent is added for the units that are not rejected because of the quality of the machine-produced output. You also add the monthly salaries of the three workers. Then you calculate the monthly cost of the machine, by dividing the purchase price by 12 months per year, and divide that by the 10 years the machine should last. The manufacturer tells you what the power consumption of the machine is, and you can get power cost numbers from accounting, so you figure the cost of electricity to run the machine and add the purchase cost to get a total cost figure. You subtract your total cost figure from your total benefit value and your analysis reveals a profit.

All you have to do now is present the calculation to the president, right? Wrong. You have the right idea, but you left out a lot of detail. Look at the benefits first. You used the selling price of the units to calculate the value. But the sales price includes many additional factors (e.g., the profit margin) that will unnecessarily complicate your analysis if you include them. Instead, get the activity-based value of the units from accounting and use that. You rightly added the value of the increased quality by factoring in the average reject rate, but you may want to reduce that a little because even the machine will not be perfect. Last, when calculating the value of replacing three employees, in addition to their salaries, you should add their overhead costs, the costs of their benefits, and so forth, to obtain the workers’ full labor rates. Every future benefit
must be discounted to turn it into a comparable current benefit. In addition to properly quantifying the benefits, it is important to make sure you included all of them.

As for the cost of the machine, in addition to its purchase price and any taxes you will have to pay on it, you must add the cost of interest on the money spent to purchase it. Check the amortization period. Just because the machine may last 10 years, this does not mean the company will keep it on the books that long. It may amortize the purchase over half that period if it is considered capital equipment. If the cost of the machine is not enough to qualify as capital, the full cost will be expensed in one year. Adjust your monthly purchase cost of the machine to reflect these issues. You have the electricity cost figured out, but there are some costs you missed too. The typical failure of a cost–benefit analysis is not including all the costs. In the case of the machine, here are some of the questions you may have to address: Will the machine fit in the existing available space? What will it cost to install the new machine? Will you need someone with special skills to operate the machine? Does this person need special training and what will be the salary, including overheads, cost? Will the new machine require soundproofing to be built around it? Will the new machine increase the insurance premiums for the company?

Once you have collected all the positive and negative factors and have quantified them, you can put them together into an accurate cost–benefit analysis. You will then know whether the machine will save your company a significant amount of money each year.

This is just one simple example of how you can use cost–benefit analysis to determine the advisability of a course of action and then to support your decision once you propose the action. CBA is a relatively simple and widely used technique for deciding whether to undertake a project. As its name suggests, the technique involves simply adding up the discounted value of the benefits of the project, and subtracting the costs associated with it. Costs may be one-off or may be ongoing. Benefits are most often received over time. We build this effect of time into the analysis by calculating a payback period. This is the time it takes for the benefits of the project to repay its costs. Many companies look for payback over a specified period of time (e.g., three years).

25.2.2 Broader Approach

In its simple form, cost–benefit analysis is carried out using only financial costs and financial benefits. For example, a simple cost–benefit analysis of a road scheme would measure the cost of building the road and subtract this from the economic benefit of improving transport links. It would not measure either the cost of environmental damage
or the benefit of quicker and easier travel to work. A more sophisticated approach to
cost–benefit measurement models requires a financial value to be put on intangible costs
and benefits. This can be highly subjective. Take, for example, a historic water meadow.
Is it worth $50,000 to protect the meadow from development or is it worth $500,000
because of its environmental importance? What is the value of stress-free travel to work
in the morning? The example of CBA we explained above is intentionally simple. Where
large sums of money are involved and many persons are concerned (e.g., as would be
the case in a public project), project evaluation can become an extremely complex and
sophisticated art.

To recap, CBA is a powerful, widely used and relatively easy tool for deciding
whether to make a change. To use the tool, first work out how much the change will
cost to make. Then calculate the benefit you will get from it. Where costs or benefits
are paid or received over time, work out the time it will take for the benefits to repay the
costs. CBA can be carried out using only financial costs and financial benefits. You may,
however, decide to include intangible items within the analysis. As you must estimate
a value for these, this inevitably brings an element of subjectivity into the process.

25.3 The Process of CBA

At the heart of any public project appraisal decision is this basic question: Does a
planned project lead to a net increase in social welfare? Conducting a CBA requires a
sequence of steps to be followed, which we now describe.

The first step in any CBA is to clarify the purpose of the analysis, or to define the
“project.” At this stage, it is important to delineate the limits of the project as well
as its standing. The standing of a project is the identification of which benefits and
costs should count as well as the time horizon over which costs and benefits should
count. As a basic rule, all stakeholders should be included insofar as they are affected
by the project. The definition of the project involves the identification of its many costs
and benefits with measures of risk when the project involves probabilistic outcomes and
an assessment of uncertainty when it is impossible to assign probabilities to different
outcomes. The distributional impact of the project must also be identified. The limits
of the project represent the scope and the reach of the various impacts of the project.

The second step is the calculation of social costs and social benefits. This would
include calculation of (1) tangible benefits and costs (i.e., direct costs and benefits),
and (2) intangible benefits and costs (i.e., indirect costs and benefits—externalities).
For example, if a child is effectively immunized against a transmittable disease, the
direct benefit is that the child will stay healthy. Positive externalities arise because that child will not pass on the disease to others, with favorable impacts on productivity levels and health sector budgets.

This process is very important. It involves trying to identify all the significant costs and benefits. The CBA, however, always compares the situation with and the situation without the project. This is needed because in CBA only the effects that may be attributed exclusively to the project should be considered. In the CBA way of thinking effects are defined as changes that take the form of additional costs and additional benefits. It is worth noting that the distinction between the “situation with project” and “the situation without project” is not the same as the “before-project” and “after-project” situations. This basic principle can be illustrated by the following example of a road project. A proposed project involves the reconstruction of a road linking a regional town to the capital. At present, 1,000 cars use this road daily. Two consultants were asked to estimate the benefits of road improvement. After project completion they both agreed that 1,400 cars would use the road daily. However, their estimates of the project benefits (i.e., additional daily traffic) were different, namely 400 and 100 cars. The consultants give the following justifications for their estimates: Consultant 1: the benefit is 400 because this is the increase in daily traffic. Consultant 2: the benefit is 100 because even if the road is not reconstructed, traffic will increase to 1,300, because of autonomous economic growth in the region and drivers will still use the road, even when it is in a bad condition.

Consultant 1 does not apply the CBA principle soundly. To see which changes in traffic may be attributed to the project, consultant 1 should have investigated the difference between the expected situation if the project is implemented and what would have happened without the project. In fact the focus was incorrectly on the difference between the traffic before (present, pre-project) and after the project. This approach is only correct if the project is the only factor influencing traffic. And this is not the case, so consultant 1 is wrong. That does not mean consultant 2 is right either. Suppose a third consultant comes with the following analysis. Consultant 3: the benefit is 800 because, without the project, traffic would decline to 600 cars daily due to fewer drivers being willing to use the road when it continues to deteriorate further. Whereas both consultant 2 and consultant 3 apply the CBA principle soundly, they differ on traffic forecasting and consequently arrive at widely different estimates on the without-case and thus the project benefits. One of them, or both of them, may be wrong depending on the quality of their traffic forecasts. The costs and benefits of the project may be hard to estimate if they include changes in human health and mortality. They may also have long-lasting effects and involve uncertainty.
25.3.1 Discounting Future Values

Another basic requirement of CBA is discounting. This is important as costs and benefits do not occur at the same time but have the form of flows over time. It makes a significant difference when, after having made an investment (incurred a cost), one has to wait a long time before benefits are generated. Generally, people would rather have money right away than wait for it. They are only willing to wait for something, if a larger reward is promised. This is the consumption side of time preference.

There is also an investment side of time preference. Consider holding $100 that could be invested and in one year’s time would realize a profit or earn interest. The rate of time preference, or discount rate, is often equated to the interest rate. Hence individuals normally prefer to enjoy the benefits now rather than later—so the value of future benefits has to be discounted.

Costs occur mostly in the short-term (investment) and to a lesser extent in the long term (maintenance and operational costs). In the case of nuclear power station, there are significant operational costs in the long term. Benefits can arise in the long term. So the choice of the discount rate will impact particularly the benefits far away in the future (i.e., back-end vs. front-end loaded benefits). All costs and benefits are “brought back” to the starting time. With a rate of 8 percent, a time line of 30 years is usually applied. After this, the discounted costs and benefits become so small as to be insignificant. For example, a sum of $1,000 discounted at 8 percent over 30 years shrinks to less than $100. Lower rates imply a longer timeline.

These comments can be now be incorporated into a formal calculation. Let the benefits of a project at time $t$ be $B(t)$ and the costs of the project be $C(t)$. The basic rule of cost–benefit analysis is that the project should be adopted if the associated net present value (NPV) is positive,

$$
NPV = \sum_{t=0}^{T} (1 + d)^{-t} [B(t) - C(t)].
$$

(25.1)

where $T$ is the time line of the project.

The benefit–cost ratio $BCR$ is similar to the NPV. Whereas the $NPV$ is the difference between all costs and benefits, the $BCR$ is the ratio of (discounted) costs and benefits. It is given by

$$
BCR = \frac{\sum_{t=0}^{T} (1 + d)^{-t} B(t)}{\sum_{t=0}^{T} (1 + d)^{-t} C(t)}.
$$

(25.2)
It is assumed initially that benefits are gross benefits, and costs are gross costs, which means that all costs (investments and operational recurrent costs) are added together. For a project to be selected, the BCR should exceed 1. The BCR is used least often, as there are no strict guidelines on how to treat recurrent operational costs: some analysts prefer to include them in the cost denominator, whereas others deduct them from the benefits. So the outcome depends on the approach followed. Although when systematically applied in all cases the problem is small, the NPV is preferred for its unequivocal outcome.

Another way of analyzing the flow of discounted costs and benefits of a project is calculating its internal rate of return (IRR). The IRR is the rate with which the discounted costs equal the discounted benefits. That is, the rate \( d = d^0 \) such that the NPV is zero at that particular rate:

\[
NPV = \sum_{t=0}^{\infty} (1 + d^0)^{-t} [B(t) - C(t)] = 0. \tag{25.3}
\]

The IRR can then be compared with a baseline or standard rate, for example, the current interest rate or a certain minimum rate, and if the IRR is higher, the project would be profitable.

These values can be used to compare different projects. The government may have limited funds at its disposal and therefore faces a choice about which projects should be given the go-ahead. The IRR is the most popular profitability indicator because it is a relative measure that allows a direct comparison between investments and market interest rates (yield). Nevertheless, there are three constraints on its use.

First, the use of the IRR in the case of mutually exclusive projects may lead to incorrect recommendations. Mutually exclusive projects occur if implementation of one project makes the other project impossible. For instance, if a site is used for a factory, it cannot be used for agriculture. The IRR may then suggest that the factory project be chosen, whereas the agriculture project would be preferable at a given discount rate. Consider two different size projects with immediate cost and future benefit: project A (big) costs 100 for a benefit of 110, and project B (small) costs 10 for a benefit of 12. The IRR is higher for project B with a return of 20 percent against 10 percent for project A. However, project A displays a higher NPV than project B for a given discount rate. For example, with discounting at 5 percent, the NPV of project A is 5, whereas it is only 1.5 for project B. So the IRR would recommend the (small) project B whereas the NPV would recommend the (big) project A.
Second, the IRR may not be unique when the flow of net benefits is fluctuating between positive and negative values. Take a project C with a flow of net benefits over three periods of $-1, +5, -6$. Then the IRR can either be 100 or 200 percent. The reason is that when computing the NPV as a function of discount rate, we find that the NPV curve is a bell—shaped with roots at 100 percent and 200 percent. So, for any discount rates outside this range, the NPV is negative.

Third, if the IRR of project A is 15 percent and 20 percent for project B, the IRR does not prove that alternative B is better and should be selected. If the discount rate is 10 percent, both should be implemented. The NPV is an absolute profitability indicator and, like the IRR, should not be used to rank project alternatives. A small, very profitable project might have a lower NPV than a large, marginally profitable project, as illustrated by the previous example. The NPV simply shows whether or not a project should be selected. The advantage of the NPV is that it is also applicable in the case of mutually exclusive projects, when the IRR should not be applied.

Finally, the IRR is a purely positive concept—it is the solution that makes the present discounted value zero. In contrast, the NPV can be a normative concept that involves welfare judgments about the relative social weights assigned to current and future generations. This point is developed in detail in sections 25.7.2 and 25.6.

Before a decision is reached, a final step has to be taken in the CBA that concerns an important question: If you estimate that the possible benefit (or cost) is $1 million, how likely is that outcome? In some cases, the occurrence of benefits or costs may be uncertain. Even if it is reasonably clear that a benefit or cost will “occur,” it is still important to determine the degree of uncertainty about the actual values of the costs and benefits. In CBA it is unwise to assume that because costs of equipment and capital infrastructure can be estimated, the costs are more certain than the benefits. History provides many examples where the costs of major projects were seriously understated or overstated (e.g., the dismantling of a nuclear power plant and decontamination of the site). In other words, there may be cost pessimism or cost optimism.

In view of this problem it is important to conduct a sensitivity analysis, namely to learn how the final net benefit figure changes if costs are increased or decreased by some percentage. Because of the uncertainty often involved in CBA, and opinions that differ on the “correct” costs and benefits or other assumptions regarding effects, many CBA studies include a sensitivity analysis. Two approaches can be followed in a sensitivity analysis. The first is to determine NPVs for different assumptions regarding costs and benefits (including prices, time frame, and discounting). The second, and alternative approach, is to analyze switching values. These are the specific values of costs, benefits, and timing that cause a project to switch from being beneficial to being
nonbeneficial. Switching values provide information on the cost and benefit items for which the overall result is most sensitive. For instance, it may be found that—starting from initial assumptions—a 10 percent increase in investment costs would be needed to decrease the $NPV$ to 0 but that a 1 percent increase in the (average) operational and maintenance cost has the same effect. Or that increasing the discount rate at 3 to 3.5 percent means that the project no longer yields a profit.

25.4 Principles of CBA

One of the problems of CBA is that the method for the computation of many components of benefits and costs is intuitively obvious, but there are others for which intuition fails to suggest methods of measurement. Therefore some basic principles are needed as a guide. We now briefly review a number of these and then look in more detail at two of them in the following sections.

25.4.1 Common Unit of Measurement

In order to reach a conclusion as to the desirability of a project all aspects of the project, positive and negative, must be expressed in terms of a common unit; that is, there must be a “bottom line.” The most convenient common unit is money. This means that all benefits and costs of a project should be measured in terms of their equivalent monetary value. A program may provide benefits that are not directly expressed in terms of dollars, but there is some amount of money the recipients of the benefits would consider to be just as good as the project’s benefits. For example, a project may provide for the elderly in an area a free monthly visit to a doctor. The value of that benefit to an elderly recipient is the minimum amount of money that recipient would take instead of the medical care. This could be less than the market value of the medical care provided. It is assumed that more intangible benefits such as from preserving open space or historic sites have a finite equivalent monetary value to the public.

Not only do the benefits and costs of a project have to be expressed in terms of equivalent monetary value, but they have to be expressed in terms of dollars of a particular time. A dollar available five years from now is not as good as a dollar available now. This is due to the differences in the value of dollars at different times because of inflation and because of the effect of interest payments. A dollar available now can be invested and earn interest for five years and would be worth more than a dollar in five years. If the interest rate is $r$, then a dollar invested for $t$ years will grow to be $(1 + r)^t$. Therefore the amount of money that would have to be deposited now so
that it would grow to be one dollar \( t \) years in the future is \((1 + r)^{-t}\). This is called the discounted value or present value of a dollar available \( t \) years in the future.

When the dollar value of benefits at some time in the future is multiplied by the discounted value of one dollar at that time in the future, the result is the present discounted value of the benefit of the project. The same thing applies to costs. The net present value of the project is just the present discounted value of the benefits less the present discounted value of the costs. The choice of the appropriate interest rate to use for the discounting is a separate issue that will be treated later.

25.4.2 Revealed Preferences

According to some economists, the valuation of benefits and costs used in CBA should reflect the preferences revealed by choices that have been made by individuals in the different markets in which they act. Information contained in the demand curve tells us much about what people are willing and able to pay for something. This is important in revealed preference theory. When consumers make purchases at market prices, they reveal that the things they buy are at least as beneficial to them as the money they relinquish.

Hard choices made in markets are the best guide to private benefits. For example, improvements in transportation frequently involve saving time. The question is how to measure the monetary value of that time saved. The value should not be merely what transportation planners think time should be worth or even what people say their time is worth (the stated preference for time saving). The value of time should be that which the public reveals their time is worth through choices involving trade-offs between time and money (revealed preference for time saving). The value of time is revealed in many choice situations. If people have a choice of parking close to their destination for a fee of 50 cents or parking farther away for free and spending 5 minutes more walking, and they always choose to spend the money and save the time and effort, then they have revealed that their time is more valuable to them than 10 cents per minute. If they were indifferent between the two choices they would have revealed that the value of their time to them was exactly 10 cents per minute.

The most challenging part of CBA is finding past choices that reveal the trade-offs and equivalencies in preferences. For example, the valuation of the benefit of cleaner air could be established by finding how much less people paid for housing in a more polluted area that is otherwise identical in characteristics and location to housing in a less polluted area. Generally, the value of cleaner air to people as revealed by the hard market choices seems to be less than the stated valuation of clean air.
25.4.3 Valuing Market Goods: Marshallian Surplus

Consumers will increase their consumption of any good up to the point where the benefit of an additional unit (marginal benefit) is equal to the market price of that unit (the marginal cost). Therefore, for any consumer buying some of a commodity, the marginal benefit is at least equal to the market price. The marginal benefit will decline with the amount consumed just as the market price has to decline to get consumers to purchase a greater quantity of the commodity. The relationship between the market price and the quantity consumed is summarized in the demand curve. Thus the demand curve provides the information about marginal benefit that is needed to place a money value on an increase in consumption.

The increase in gross benefits resulting from an increase in consumption is the sum of the marginal benefit times each incremental increase in consumption. As the incremental increases considered are made smaller and smaller, the sum goes to the area under the marginal benefit curve. But the marginal benefit curve is the same as the demand curve, so the increase in gross benefits is the area under the demand curve. The area is taken over the range from the lower limit of consumption before the increase to consumption after the increase. In figure 25.1 the lower limit of consumption is $x_0$ and the increase is the amount $\Delta$. The benefit of this increase is the shaded area. When the increase in consumption is small compared to the total consumption, the gross benefit is adequately

![Figure 25.1](image_url)

Increase in gross benefit
approximated, as is shown in a welfare analysis, by the market value of the increased consumption; that is, market price times the increase in consumption. The cost of this additional consumption is also equal to price times the increase in consumption, so the net benefit is zero.

Formally, consider a quasi-linear utility function given by \( U(x, y) = v(x) + y \), with \( v(x) \) increasing \( (v'(x) > 0) \) and concave \( (v''(x) < 0) \). Let good \( y \) be the numéraire and \( p \) the (relative) price of good \( x \). The budget constraint is \( px + y \leq M \). The inverse demand function is obtained by equating price to marginal utility \( p = v'(x) \). By inversion, the demand function is \( x = v^{-1}(p) \equiv x(p) \) (assuming an interior solution, \( y > 0 \), for otherwise, \( x(p) = \frac{M}{p} \)). The gross benefit of this level of consumption is

\[
\int_0^{x(p)} v'(x) dx = v(x(p)) - v(0). \tag{25.4}
\]

The Marshallian surplus (or net benefit) is obtained from the gross benefit by subtracting the cost of consumption at the market price:

\[
\int_0^{x(p)} v'(x) dx - px(p) = v(x(p)) - v(0) - px(p). \tag{25.5}
\]

This is the welfare gain from consuming \( x(p) \) at price \( p \). An allocation is efficient if it maximizes the sum of individual surpluses. This is not a complete measure of welfare because it does not take into account the income level \( M \). The Marshallian surplus is only valid as a measure of efficiency in a quasi-linear economy. This is because with quasi-linear preferences utility is transferable, so efficiency and the distribution can be treated as separate issues.

### 25.4.4 Valuing Nonmarket Goods: Hedonic Prices

It is sometimes necessary in CBA to evaluate the benefit of saving human lives. There is considerable antipathy in the general public to the idea of placing a dollar value on human life. Economists recognize that it is impossible to fund every project that promises to save a human life and that some rational basis is needed to select which projects are approved and which are turned down. The controversy is defused when it is recognized that the benefit of such projects is in reducing the risk of death. There are many cases where people voluntarily accept increased risks in return for higher pay, such as in the oil fields or mining, or for time savings in higher speed in automobile
travel. These choices can be used to estimate the personal cost people place on increased risk and thus the value to them of reduced risk. This computation is equivalent to placing an economic value on the expected number of lives saved.

Methods based on hedonic pricing use the fact that some market goods are in fact bundles of characteristics, some of which are intangible goods. By trading these market goods, consumers are thereby able to express their values for the intangible goods, and these values can be uncovered through the use of statistical techniques. This process can be hindered, however, by the fact that a market good can have several intangible characteristics, and that these can be collinear. It can also be difficult to measure the intangible characteristics in a meaningful way.

Similarly defensive expenditures can be used to estimate the value of intangible impacts. This refers to individual behavior to avoid negative intangible impacts. For example, people might buy goods such as safety helmets to reduce accident risk, and double-glazed windows to reduce traffic noise, thereby revealing their valuation of such intangible goods as safety and quietness. Travel cost methods use the fact that market and intangible goods can be complements, to the extent that a purchase of market goods and services is required to access an intangible good. Specifically, people have to spend time and money traveling to recreational sites, and these costs reveal something of the value of the recreational experience to those people incurring them.

25.4.5 Impact Assessment

The impact of a project is the difference between what the situation in the study area would be with and without the project. That is, when a project is being evaluated the analysis must estimate not only what the situation would be with the project but also what it would be without the project. For example, in determining the impact of a rapid transit system (RTS) in a metropolitan area, the number of rides that would have been taken on an expansion of the bus system should be deducted from the rides provided by RTS, and likewise the additional costs of such an expanded bus system would be deducted from the costs of RTS. In other words, the alternative to the project must be explicitly specified and considered in the evaluation of the project. As already noted the with-and-without comparison is not the same as a before-and-after comparison.

Another example shows the importance of considering the impacts of a project and a with-and-without comparison. Suppose that an irrigation project is proposed to increase cotton production in Arizona. If the US Department of Agriculture limits the cotton
production by a system of quotas, then expanded cotton production in Arizona might be offset by a reduction in the cotton production quota for Mississippi. Thus the impact of the project on cotton production in the United States might be zero despite the increase in the amount of cotton produced by the project.

Double counting of benefits or costs must be avoided. Sometimes an impact of a project can be measured in two or more ways. For example, when an improved highway reduces travel time and the risk of injury, the value of property in areas served by the highway will be enhanced. The increase in property values due to the project is a very good way, at least in principle, to measure the benefits of a project. But, if the increased property values are included, then it is not necessary to include the value of the time and lives saved by the improvement in the highway. The property value went up because of the benefits of the time saving and the reduced risks. To include both the increase in property values and the time saving and risk reduction would amount to double counting.

25.4.6 Decision Criteria

If the discounted present value of the benefits exceeds the discounted present value of the costs, then the project is worthwhile. This is equivalent to the condition that the net benefit must be positive. Another equivalent condition is that the ratio of the present value of the benefits to the present value of the costs must be greater than one. If there are more than one mutually exclusive projects that have positive net present value then there has to be further analysis. From the set of mutually exclusive projects the one that should be selected is the one with the highest net present value. If the funds required to carry out all of the projects with positive net present value are less than the funds available, then the discount rate used in computing the present value is too low and does not reflect the true cost of capital. The present values must be recomputed using a higher discount rate. It may take some trial and error to find a discount rate such that the funds required for the projects with a positive net present value are no more than the funds available. Sometimes as an alternative to this procedure, people try to select the best projects on the basis of some measure of goodness such as the internal rate of return or the benefit–cost ratio. This is not valid for several reasons.

The magnitude of the ratio of benefits to costs is to a degree arbitrary because some costs such as operating costs may be deducted from benefits and thus not be included in the cost figure. This is called “netting out of operating costs.” Netting out may be done for some projects and not for others. The manipulation of the benefits and
costs will not affect the net benefits but may change the benefit–cost ratio. However, it will not raise a benefit–cost ratio that is less than one to above one. More important, the ranking of different projects can be reversed when choosing the benefit–cost ratio instead of the \( NPV \). To better see this, consider three projects \( i = 1, 2, 3 \) with the following (discounted) benefits and costs \( (B_i, C_i) \): \((200, 100), (110, 50), \) and \((120, 50)\). Based on the benefit–cost ratio, project 3 \( \succ \) project 2 \( \succ \) project 1 with respect to the benefit–cost ratios of 2.4, 2.2, and 2.0. Conversely, on the basis of the \( NPV \), project 1 \( \succ \) project 3 \( \succ \) project 2 with respect to the \( NPVs \) of 100, 70, and 60.

25.4.7 Difficulties for CBA

CBA should be seen as a highly practical method for making decisions. The various approaches that we have described were developed to refine the method. There are still difficulties that CBA must confront in use. These are now briefly described.

Attaching Valuations to Costs and Benefits
Some costs are easy to value such as the running costs (e.g., staff costs) and capital costs (e.g., purchases of new equipment) of a project. Other costs are more difficult to evaluate—not least when a project has a significant impact on the environment. The value attached to the destruction of a natural habitat is to some people priceless but to others entirely worthless. Costs are also subject to change over time. This is frequently illustrated by the cost overruns in large construction projects or innovative IT schemes (e.g., the introduction of electronic road pricing).

Covering Everyone Affected
The ideal application of CBA would include all parties that are directly or indirectly affected by a project. Inevitably, with any major project, such as the construction of a new airport or a new road, there are a large number of potential “stakeholders” who stand to be affected (positively or negatively) by the project. For example, people living many miles from an airport could be affected by noise. CBA cannot hope to include all stakeholders, so there is a risk that some groups might be left out of the decision process. The effect of discounting is to make costs and benefits that occur in the future insignificant relative to current costs and benefits. This can be interpreted as effectively excluding the effects on future generations from the evaluation of the project. This is an issue that we return to in chapter 26. The position of “nonhuman” stakeholders raises a further set of issues.
Distributional Consequences
Costs and benefits mean different things to different income groups—benefits to the poor are usually worth more (or are they?). Those receiving benefits and those burdened with the costs of a project may not be the same. Are the losers to be compensated? To many economists, the equity issue is as important as the efficiency argument.

Social Welfare Is Not the Same as Individual Welfare  What we want individually may not be what we want collectively. Do we attach a different value to those who feel “passionately” about something (e.g., building of new housing on greenfield sites) contrasted with those who are more ambivalent? Valuing the environment: How are we to place a value on public goods such as the environment where there is no market established for the valuation of “property rights” over environmental resources? How does one value “nuisance” and “aesthetic values”?

Valuing Human Life  Some measurements of benefits require the valuation of human life—many people are intrinsically opposed to any attempt to do this. This objection can be partly overcome if we focus instead on the probability of a project “reducing the risk of death.” There are indeed insurance markets in existence that tell us something about how much people value their health and life when they take out insurance policies.

Discounting the Future  Discounting reduces all future costs and benefits to express them as today’s values. The key question is: How do you choose an “interest rate” for reducing future costs to give them a present value today? Setting a general discount rate for new projects has important implications for the environment.

A low discount rate is often favored by economists, since they argue that investing a high proportion of current income is a good way of providing for the future. A high discount rate may also be favored because it discourages investment (and by implication environmental damage) in the present. Most projects have lifetimes of 20 to 30 years—with many of the big costs arising early in a project, such as from construction, whereas the stream of benefits from a project occur over a much longer period of time. But for many huge construction projects, some of the costs only become apparent in the long run. Consider the building of a new nuclear power station. Environmentalists would argue that there is a long list of costs from waste management and decommissioning that stretch over 100 years into the future whereas no social benefits exist to offset these costs beyond year 30 or 40 (when the nuclear power station might reasonably be expected to be ready for closure). The value of decommissioning costs over 100 years
away is almost negligible no matter what discount rate we use. This makes discounting difficult to justify.

**Accommodating Risk**

Consider a cost–benefit analysis of the effects of genetically modified foods. Two principles can be applied to the risk that is inherent in these foods. Under the *precautionary principle*, it is assumed that the foods are toxic until they are proved to be safe. From this principle follows the policy conclusion: if in doubt, regulate. The contrasting alternative is the *free-market principle*: assume that the foods are safe until a hazard is identified. The consequent policy maxim is: if in doubt, do not regulate. These alternative principles have also been applied to climate change. Some economists apply the precautionary principle to argue that action should be taken now to mitigate greenhouse gas emissions even if there is no compelling evidence that climate is changing. Other economists follow the free-market principle and wait for incontrovertible evidence before the cost of mitigation is incurred. Despite these conflicts, most economists argue that CBA is better than other ways of including the environment in project appraisals.

Many projects to which CBA is applied involve risk. Cost and benefits are risky because of future contingencies and the problems of forecasting. The issue is how to take account of risk. It is not correct to use the expected value of the forecast outcomes. This does not work because it does not take account of *risk aversion*. Individuals who are risk averse will not accept a fair gamble. The only gambles they will accept will have a strictly positive expected payoff. The amount required to accept the gamble is the risk premium. With risk aversion the value of a set of risky payoffs is always less than the expected value. There is compelling evidence that in aggregate individuals are risk averse. This is reflected by the fact that in financial markets riskier assets (e.g., shares in small companies) must offer a higher expected return that safer assets (e.g., government bonds). This risk aversion should be reflected in social preferences and embedded within CBA. Last, there is a potential tension between subjective and objective measurement of risks. To illustrate this problem, Pollack (1998) told a story about Happyville where the citizens became afraid of a possible contaminant in their drinking water. They requested a chemical analysis of the water in order to have an objective evaluation of the risk. It turned out that there was no risk to the quality of the water and that there was no need for a purification plant because it would cost money without creating a benefit. However, the citizens of Happyville did not trust the chemical analysis and they insisted on having the purification plant. Should we consider this demand based on subjective
risk evaluation in the CBA, or should we favor the objective risk evaluation and not go ahead with a purification plant?

25.5 Valuing Life

There are many projects that may involve mortality. The construction of a new hospital may help save lives through more readily accessible treatment. This reduction in mortality should be included as a benefit in a CBA of the hospital. However, the construction of the hospital may prove to be hazardous and potentially expose the construction workers to serious injury or death. The cost of this increased risk of mortality should also be factored into the CBA calculation.

The logic of CBA requires that all mortality risks be given a monetary valuation. Many noneconomists may have moral objections to this idea. But there are two counter-arguments. First, if life is assumed to be beyond valuation, then no project that involves even the slightest risk of death could ever be undertaken, or else unlimited resources would be spent on any project that involves the slightest reduction in death risk. Both are obviously ridiculous positions, and no society could function under these circumstances. Second, individuals in their own daily choices frequently place a valuation on life. Consider a worker choosing between two jobs with different risks of death and injury. How much of a wage premium do they require to accept the riskier job? Consider an individual choosing how much to spend on safety equipment (e.g., a motorcycle helmet). Is it worth purchasing the higher quality product? Consider the choice of a car. Is it worth paying extra for additional safety bags or for stability control? In every case the decision will weigh the value of life and the probability of death against the additional cost.

A further justification for valuing life follows from the consequences of not doing so. In the absence of an explicit value, decisions will be made that imply different valuations. Policy choices can be inconsistent, and limited funds will not be used efficiently. For example, assume that it is decided to fund a road safety project that costs $2 million and saves four lives but a medical treatment that costs $3 million and saves twelve lives is not funded (the projects are mutually exclusive due to a budget limit). The choice to fund the first project implies that the value of life is at least $0.5 million, the choice not to fund the second project implies a value of life of less than $0.25 million. Clearly, these choices are inconsistent, and their bearing on the choice of project is incorrect. Such decision-making can be costly if it leads to a misallocation of resources that results in lives being needlessly lost.
The value of life is normally addressed in a setting where the increase or decrease in mortality can only be expressed as a possibility. This is done to avoid the moral questions that would arise if it were known exactly whose life would be lost. The underlying assumption is that the consequences of implementing a project are uncertain: it is not known who will die or be saved, nor how many people will die or be saved. What is known is the reduction (or increase) in the probability of mortality due to the project. For example, a construction project may increase the probability of death for all workers, while an irrigation project may reduce the probability of death through crop failure. These considerations give rise to the concept of the value of statistical life (VSL), which is defined as the benefit arising from the expected saving of one life. We will now consider the theory behind the VSL, and then proceed to methods of evaluation. Finally, we will review some current estimates of the VSL.

The first step to obtain a VSL is to consider the willingness to pay (WTP) for a risk reduction that will extend the length of life. The WTP is defined as the maximum amount that a person is willing to pay to avoid a risk. Let $U(M, a)$ denote the level of utility with income $M$ when the individual is alive, and $U(M, d)$ the utility when dead. The latter can either be zero or be interpreted as the utility derived from leaving a bequest to descendents. If $p$ is the probability of dying in the current period the level of expected utility is

$$EU = [1 - p]U(M, a) + pU(M, d).$$

(25.6)

The willingness to pay to avoid the risk of death is defined by the value of $\rho$ that ensures the utility of life with income $M - \rho$ is equal to the expected utility with income $M$ but with a probability $p$ of death

$$U(M - \rho, a) = [1 - p]U(M, a) + pU(M, d).$$

(25.7)

The VSL is the marginal value of a reduction in the risk of death, so is defined as the rate at which the WTP increases with $p$

$$VSL \equiv \frac{\partial WTP}{\partial p} = \frac{\partial \rho}{\partial p}.$$  

(25.8)

An equivalent definition of the VSL is the total WTP of a group of $N$ people who experience a uniform increase of $\frac{1}{N}$ in their risk of dying. For example, consider 1,000 individuals who each have a WTP of $500 to avoid an increase in the risk of dying of 1 in 1,000. The VSL is then the total WTP, which is $500,000.

The concept of value of a statistical life year (VOLY) is related to the VSL. Specifically, assume that a VOLY is constant over the rest of a person’s remaining lifetime,
and assume the person lives for $T + 1$ years, including the present year. The $VOLY$ and the $VSL$ are related by

$$VSL = \sum_{t=0}^{T} (1 + d)^{-t} VOLY,$$

where $d$ is an appropriate discount rate. The notion of $VOLY$ is used in policy analyses in addition to (or instead of) that of $VSL$, but depending on the age of the people whose lives are saved by the policy, $VOLY$ can lead to recommendations in conflict with those obtained by using $VSL$. Consider, for example, two alternative public programs that both save 100 lives. With the first program the lives that are saved are of young adults, whereas with the second the lives saved are of elderly people. If the $VOLY$ is constant with respect to age, the longer life expectancy of young adults gives the program saving young lives greater benefits when the $VOLY$ is used. In contrast, if there is a single value of $VSL$ for all people, then the two programs would give the same benefit.

Medical treatments can have different outcomes in terms of the level of future health. Some treatments can return a patient to full health, whereas others can prolong life but with a reduced degree of functionality. It is accepted that a life lived with incapacity or in pain is not as rewarding as a life lived in good health. The value assigned to a treatment in a CBA should reflect this fact. The concept that has been adopted in medical applications is the *quality-adjusted life year* ($QALY$). Every treatment is assigned a value $q$, $0 \leq q \leq 1$, that measures the quality of life for the patient after the treatment. The variable $q$ will equal 1 if the treatment restores the patient to full health and a value close to 0 if it does little more than just sustain life. The $VSL$ for that patient is then given by $qVSL$, and this value should be used as the benefit of treatment in CBA.

The $VSL$ has clear conceptual foundations, but this does not make identifying a value for it straightforward. One approach is to apply the ideas of section 25.4.4 to markets in which actions reveal individual valuations of life. The labor market has proved particularly useful since there are many different types of job available and these differ in the probability of mortality at work. Many forms of employment have negligible mortality risk but some, such as mining, fishing, and construction, have significant risks. The theory of compensating differentials asserts that wages for different jobs should reflect the characteristics of the jobs. Hence jobs with mortality risk should have a wage premium to reflect that risk. The wage premium can then be used to infer the $VSL$. Alternatively, it is possible to infer the $VSL$ from the expenditure that individuals make to reduce the risk of mortality. The amount spent on the purchase of protective equipment, such as a motorcycle helmet or an optional safety aid on a car,
coupled with the amount of risk reduction it secures provides evidence on the private assessment of VSL. The fundamental problems faced by both these methods are the behavioral anomalies identified in chapter 3. Individuals can suffer from optimism bias and underestimate the probability of a poor outcome or have excessive belief in their own skill levels. They can assess the outcomes using preferences that do not conform with the expected utility assumption on which the VSL was constructed. There will also be self-selection of the least risk-averse into risky occupations so that the wage premium reflects the valuation of an extreme, not an average.

Table 25.1 provides some examples of the values estimated using the methods described. The key point is that there is considerable variation in these estimates. It is also interesting that some of the numbers are above the typical lifetime income. The mean income in the United States in 2010, according the Census Bureau, is just short of $70,000. For a working life of 45 years this gives total mean lifetime income of $3.15 million. It seems inconsistent that the VSL can be significantly above this. The US Department of Transport is explicit about the value it uses. In January 1993 the Department adopted a guidance memorandum, “Treatment of Value of Life and Injuries in Preparing Economic Evaluations,” that stated the recommended value to be used in Departmental regulatory and investment analyses. The initial value was set at $2.5 million and the last adjustment to $3.0 million was made on January 29, 2002.

One of the justifications for valuing life was to avoid the policy inconsistencies that could arise if life was not explicitly valued. Inconsistencies can still arise even with explicit values. As an example, different branches of the UK government apply different values. The cost effectiveness of drugs supplied by the National Health Service is determined in England, Wales, and Northern Ireland by the National Institute for Clinical Excellence (NICE). NICE usually determines a treatment not to be value for

<table>
<thead>
<tr>
<th>Study</th>
<th>Method</th>
<th>VSL ($ millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thaler and Rosen (1976)</td>
<td>US labor market</td>
<td>1.0</td>
</tr>
<tr>
<td>Viscusi (1981)</td>
<td>US labor market</td>
<td>7.9</td>
</tr>
<tr>
<td>Portney (1981)</td>
<td>Property values</td>
<td>1.0</td>
</tr>
<tr>
<td>Atkinson and Halvorsen</td>
<td>Automobile prices</td>
<td>4.8</td>
</tr>
<tr>
<td>Kniessner and Leeth (1991)</td>
<td>Australian labor market</td>
<td>4.0</td>
</tr>
</tbody>
</table>
money if it costs more than £20,000 to £30,000 per quality-adjusted life year (QALY) saved. This implies that the value of one year of healthy life is £30,000 in decisions on new pharmaceutical products. If a 70-year life span is assumed, then the total value of life from this perspective is, at most, £2.1 million. In contrast, the value of one human life saved in the assessment of traffic fatalities reduction policies is £12 million. The VSL can also differ widely across similar countries. Figure 25.2 summarizes the values used in a number of countries around the world and clearly illustrates the wide degree of variation.

### 25.6 Valuing the Future

Many projects will have effects that last for many years. The construction of a major infrastructural project might take 5 to 10 years, and once completed, it can be expected to deliver benefits for decades afterward. Some projects can have exceptionally long
lives: the half-life of plutonium-239 in spent nuclear fuel rods is 24,110 years, so any evaluation of a nuclear power plant must incorporate the very long-run costs of dealing with this material. The value assigned to costs and benefits occurring in the future relative to those occurring in the present will be critical in calculating the NPV of a long-lived project.

A basic economic fact is that when people are offered a choice between receiving $1 today or receiving a $1 in a year’s time then the $1 today will be chosen. There are three reasons proposed to explain this fact:

1. *Pure time preference* reflects a basic desire to consume in the present rather than the future. This is closest to the general notion of impatience, which leads to a unit of money today being preferred to a unit of money tomorrow.
2. *Expectation of income growth* builds on the belief that the individual will be wealthier in the future. This implies that the marginal utility of consumption is expected to be lower in the future, so that it is optimal to consume the additional $1 today when the marginal utility is higher.
3. *Uncertainty* captures the possibility that if the receipt of the $1 is postponed for a year, there may be some chance that it will not be received. Accepting it today ensures that it is received and therefore is preferable.

These three reasons are not mutually exclusive so that all can operate simultaneously. In total they imply that $1 in one year is valued by the individual as worth $\frac{1}{1+d}$ today, for some discount rate $d > 0$. The higher the value of $d$ is, the lower the relative worth of future dollars compared to current dollars. The issue in moving from this observation about intertemporal preferences to a process that can be used in CBA is to determine the correct value of the discount rate, $d$.

At this point it is worth stressing that different considerations apply to determining the discount factor of an individual and determining the discount factor of a society. An individual discount factor has to be compared to what “is,” whereas a social discount factor is more an expression of what “ought to be.” The force of this distinction will become clear as the argument proceeds.

### 25.6.1 Intertemporal Arbitrage

Consider an individual confronting a financial market where there are opportunities to borrow or invest at a rate of interest $r$ per year. With this rate of interest, a $1
invested for \( t \) years (or borrowed and paid back with interest after \( t \) years) becomes $\{1 + r\}^t$. This implies that $1$ today is worth $\{1 + r\}$ next year. Whenever \( r \) is positive, any individual offered a choice between $1$ today or $1$ in a year will take the money today: the dollar today can be spent immediately or invested for a year to give $\{1 + r\}$. Since at least one of these alternatives is strictly better than waiting a year to receive a dollar, the dollar today is strictly preferable. This argument can also be run in reverse. If the individual wishes to have an extra $1$ of consumption next year, the most they need to give up is $\frac{1}{1+r}$ since investing this amount will deliver the extra dollar. Hence, if the individual operates with any discount rate other that \( r \), there will be an arbitrage opportunity—meaning a transaction (borrowing or lending) that they would find beneficial.

What we have deduced is that exponential discounting should apply so that the value today of $1$ at time \( t \) should be $\{1 + r\}^t$. Consequently the interest rate should determine the individual discount rate, \( d \), so that \( d = r \). The natural question is to wonder how this can incorporate individual differences in preference. The answer is that the functioning of the market will ensure that the choices made according to individual preferences will ensure that, in equilibrium, all consumers will agree on the relative values of consumption in different time periods. The theoretical underpinning for this statement are now demonstrated by reviewing the economics of intertemporal choice.

Consider an economy that last for two periods. Time 0 is interpreted as the present and time 1 as the future. The production possibilities for the economy are determined by the production possibility frontier

\[ G(X_0, X_1) = 0, \quad (25.10) \]

where \( X_t \) is output at time \( t \). Now assume initially that there is a single consumer with preferences given by

\[ U = U(X_0, X_1). \quad (25.11) \]

The optimal allocation of consumption for the economy is where the production possibility frontier is tangential to the highest attainable indifference curve. This is illustrated in figure 25.3. Let the gradient of the tangent to the indifference curve at the optimum be given by $- [1 + r]$. Then, at the optimum,

\[ \frac{\partial U}{\partial X_1} = 1 + r, \quad (25.12) \]
Figure 25.3
Intertemporal equilibrium

\[ \frac{\partial U}{\partial X_0} = \frac{1}{1 + r} \frac{\partial U}{\partial X_1}. \]  

(25.13)

Hence \( r \) has the property of a discount rate that relates the marginal utility of consumption next period to the marginal utility this period. It is traditional to call this the \textit{social time preference rate}, since it governs the trade-off between present and future consumption.

The optimum can be decentralized. Assume that there is a perfect capital market so that borrowing and lending can be undertaken at the same rate of interest, \( r \). Assume also that there is no uncertainty so that all plans will be realized. Let production be controlled by competitive firms. The consumer will allocate consumption across time guided by the fact that reducing consumption today by 1 unit permits an increase of \( 1 + r \) units next year. Similarly firms will base their productions plans on the fact that each unit borrowed today will require repayment of \( 1 + r \) units. The interest rate will adjust until the plans of the consumers and the firms are in equilibrium, and the resulting value \( r \) is the equilibrium rate of interest for the economy. This is just an intertemporal version of argument concerning the Robinson Crusoe economy of chapter 2. The decentralization argument provides a direct link between the market rate of interest and the social time preference rate: the nature of equilibrium guarantees that the two are equal.
25.6.2 The UK Green Book

A version of this reasoning is put forward by the UK Green Book as an explanation of the proposed choice of a discount rate. The argument is presented in the following way:

For individuals, time preference can be measured by the real interest rate on money lent or borrowed. Amongst other investments, people invest at fixed, low risk rates, hoping to receive more in the future (net of tax) to compensate for the deferral of consumption now. These real rates of return give some indication of their individual pure time preference rate. Society, as a whole, also prefers to receive goods and services sooner rather than later, and to defer costs to future generations. This is known as “social time preference”; the “social time preference rate” (STPR) is the rate at which society values the present competed to the future. (The Green Book 2003, p. 26, HM Treasury)

It may be suspected that there is an additional step involved in moving from the analysis of a single-consumer economy to an economy with many consumers. With a single consumer it is possible to talk unambiguously about the rate of time preference. Is this still possible when there are many consumers with different preferences? The answer is that it is because the interest rate is essentially a price: it measures the rate at which present consumption can be transformed into future consumption. Under the assumption of perfect capital markets each consumer faces the same market rate of interest. The consumers adjust their consumption plans (the allocation across time) to match this market rate. In equilibrium some consumers will be net lenders in the first period (which transfers consumption from time 0 to time 1) while others will be borrowers (transferring consumption from 1 to 0). The interest rate \( r \) adjusts to achieve an equilibrium where consumption plans match the production plan. At this equilibrium the situation shown in figure 25.3 is satisfied for all consumers, so every consumer agrees on the social time preference rate.

This logic justifies the approach to calculating an NPV used in (25.1) with the market rate of interest used as the discount rate. The basic cost–benefit logic is then applied so that a project is accepted if \( NPV > 0 \), since this ensures that implementation will raise welfare. It is interesting to note the advice in the UK Green Book (Annex 6). The Green Book (Section 5.49) adopts the value of \( r = 0.035 \) (3.5 percent). There is an issue with the value of \( r = 0.035 \). Observe that at 3.5 percent $1,000 to be received 10 years into the future is worth $708.92 today, the same amount in 20 years is worth $502.57 and in 50 years $179.05. The problem that this causes is that projects with benefits that arrive long in the future will not be accepted, while projects that incur costs in the distant future will be accepted.

The nature of discounting favours projects delivering benefits in the near future with costs in the distant future. This means that we will choose not to deliver benefits to future
generations but will impose costs on them. The response to this is somewhat ad hoc in practice. The UK *Green Book* recommends that the discount rate should gradually be reduced the further into the future are the benefits and costs of the project. The logic given for this is increased uncertainty in the future, though this runs converse to the normal treatment of risk that we discuss shortly. Even at the lower rates recommended future flows are heavily discounted (e.g., $100 in 100 years is worth $2.50 today when discounted at 2.5 percent).

**LONG-TERM DISCOUNT RATES.** Where the appraisal of a proposal depends materially upon the discounting of effects in the very long term, the received view is that a lower discount rate for the longer term (beyond 30 years) should be used. The main rationale for declining long-term discount rates results from uncertainty about the future. This uncertainty can be shown to cause declining discount rates over time. In light of this evidence, it is recommended that for costs and benefits accruing more than 30 years into the future, appraisers use the schedule of discount rates provided in Table 6.1 below.

<table>
<thead>
<tr>
<th>Period of years</th>
<th>0–30</th>
<th>31–75</th>
<th>76–125</th>
<th>126–200</th>
<th>201–300</th>
<th>301+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discount rate</td>
<td>3.5</td>
<td>3.0</td>
<td>2.5</td>
<td>2.0</td>
<td>1.5</td>
<td>1.0</td>
</tr>
</tbody>
</table>


There are three possible responses to these implications of discounting. The first is to accept what the method is delivering. The fact that it reduces future values so significantly relative to current values reflects how individuals behave. Society, as the aggregate of individuals, should behave in the same way and be resigned to the consequences. The second approach is to adopt the perspective of intergenerational equity. The fact that individuals discount future flows in their own decisions does not imply that society should do the same. This is because society consists of different generations of people, and the individual logic does not extend to discounting across generations. The extreme version of this argument—which we explore further in chapter 26—uses a pure rate of time preference of zero, ignores the effect of decreasing marginal utility of income, and discounts only for risk. The third approach is to appeal to the implications of behavioral economics as described in chapter 3. There is evidence that people behave as if they apply quasi-hyperbolic discounting. That is, choices in the near term display significant present bias (so there is heavy initial discounting), but choices in the longer term seem to imply lower relative rates of discount. People are likely to
prefer $100 today to $101 tomorrow, but they are also likely to prefer $101 in 1 year + 1 day compared to $100 in one year. Such a pattern of discounting is approximately what is captured in the table reproduced from the UK Green Book. Despite what the notes to the table in the Green Book claim, the numbers are justified by quasi-hyperbolic discounting rather than risk.

25.6.3 Adjustment for Risk

The discounting rules that we have been discussing need to be modified when the project being assessed is risky. We can think about the discount rate \( r \) that we have used so far in the \( \text{NPV} \) formula as the interest rate paid on a safe bank account. Risky assets always have higher rates of return to provide the inducement required for risk-averse investors to hold them. The same argument leads to the conclusion that risky projects should also have a higher discount rate, and so a smaller discount factor. A convenient approach is to write the discount rate for a risky project as \( r + r_p \), where \( r_p \) is the risk premium. The \( \text{NPV} \) of the project is

\[
\text{NPV} = \sum_{t=0}^{T} \frac{1}{[1 + r + r_p]^t} [B_t - C_t]. \tag{25.14}
\]

The issue now is to determining what the value of \( r_p \) should be.

The answer to finding \( r_p \) is to look at the return either required by the market for projects with similar risk or offered on assets with similar risk. This works provided that something similar can be found. Generally, we can use a model that predicts the risk premium. The most popular model in finance is the capital asset pricing model (CAPM). The key variable in CAPM is the beta of an asset, which is defined by

\[
\beta_i \equiv \frac{\text{cov}(r_i, r_m)}{\text{var}(r_m)}, \tag{25.15}
\]

where \( \text{cov} (r_i, r_m) \) is the covariance of the return on the asset to the market, and \( \text{var}(r_m) \) is variance of the market return. At the equilibrium of the CAPM the expected return on asset \( i \) is is given by

\[
\bar{r}_i = r + \beta_i[\bar{r}_m - r], \tag{25.16}
\]

so that the risk premium is

\[
r_p = \beta_i[\bar{r}_m - r]. \tag{25.17}
\]
The CAPM is a useful model because it provides a precise formula for the risk premium. But it does require data to calculate $\beta_i$, and it might not be clear how the return on a project can be specified in these terms. The CAPM is also only one among many models of financial markets that will generate a value of the risk premium.

The most challenging critique to the CAPM model is the risk premium puzzle. In a nutshell, the puzzle is this: why do stocks have a much higher long-term rate of return than bonds? In theory, investors should demand a higher return on stocks, but only enough to make up for the fact that stocks are riskier than bonds. In real life, though, it turns out that even when you take risk into account, stocks still have higher returns with, on average, an extra return about twice as high as what risk aversion and the CAPM model would predict. Why? Robert Barro’s answer, building on a solution proposed nearly two decades earlier by Thomas Rietz, is that investors are fundamentally irrational: they overestimate the probability of unlikely but catastrophic losses (i.e., the black swan), and this fear makes the demand for risk-free investments larger than it rationally ought to be and thus drives down the return on bonds. It is an explanation akin to prospect theory, which is based on the fact that people are not so much risk-averse as loss-averse (see chapter 3). Consistent with prospect theory, most people feel much more strongly about the probability of a loss than they do about the probability of an equivalent gain, and it seems like this is partly what is going on here. As a result a small increase in this kind of risk, such as due to the 2008 financial crisis, leads to a noticeable response of real interest rates. When this probability goes up, the risk-free rate goes down because people put more of a premium on holding a relatively safe asset. Correspondingly the risk premium gets augmented to reflect this small-probability–big-loss event.

25.7 Theoretical Foundations of CBA

The fundamental principle guiding CBA is to let individual preferences be the (main) determining factor in guiding the social decision on the implementation of a project. In accord with this “democratic presumption” in CBA, only individual preferences should count in any collective decision-making. The essential theoretical foundations of CBA are that benefits are defined as increases in welfare and costs are defined as reductions in welfare. For a project or policy to qualify on cost–benefit grounds, the social benefits it produces must exceed the accompanying social costs. We now make the link between projects and social welfare explicit, and then consider methods of calculating whether a project increases welfare.
25.7.1 Social Welfare Evaluation

Cost–benefit analysis seeks to determine which projects make society better off. These projects can be identified by the fact that social welfare will increase if they are implemented. This section identifies what this means in the abstract and then proceeds to examine how the direction of welfare change can be determined.

The general form of Bergson–Samuelson social welfare function can be written as

\[ W(x) = W(U^1(x^1), U^2(x^2), \ldots, U^H(x^H)), \tag{25.18} \]

where \( x = \{x^1, x^2, \ldots, x^n\} \) denote the consumption levels of both tangible and intangible goods of individuals \( h = 1, 2, \ldots, H \), and \( U^h(x^h) \) the resulting utility level of individual \( h \). Social welfare is assumed to be increasing and concave in individual utility levels. Consider a project that, if it is implemented, will change the individual consumption levels to \( x + \Delta = \{x^1 + \Delta^1, x^2 + \Delta^2, \ldots, x^H + \Delta^H\} \). The new levels of consumption produce social welfare

\[ W(x + \Delta) = W(U^1(x^1 + \Delta^1), U^2(x^2 + \Delta^2), \ldots, U^H(x^H + \Delta^H)). \tag{25.19} \]

For a given social welfare function, we can conclude that the project increases social welfare if \( W(x + \Delta) > W(x) \). If it does, then the project should be implemented. The interesting issue is how to proceed if we do not know the social welfare function. More specifically, when will a project increase welfare for every social welfare function? This is an important question to ask, since CBA can be a considerably more powerful tool when its recommendations are independent of the form of social welfare function.

A first step is to apply the Pareto criterion to project evaluation. The criterion states that social welfare will increase if nobody is worse off after the project is implemented and at least some people are strictly better off. This occurs when \( U^h(x^h + \Delta^h) \geq U^h(x^h) \) for all \( h \) and \( U^h(x^h + \Delta^h) > U^h(x^h) \) for at least one \( h \). Clearly, if a project benefits everyone, then the Pareto criterion can be applied to confirm that the project will increase any social welfare function. The fundamental difficulty with this social welfare evaluation occurs when there are losers \( (U^h(x^h + \Delta^h) < U^h(x^h)) \) and gainers \( (U^h(x^h + \Delta^h) > U^h(x^h)) \) from the project. Whether or not the gains offset the losses will then depend on the social welfare function. In fact, if there are some gainers and some losers, it will always be possible to find one social welfare function for which the project increases welfare and an alternative social welfare function for which it reduces welfare. This makes it clear that CBA can only provide an evaluation of projects in such cases if it is adopts a method to trade off utility gains against utility losses.
A long-established approach in welfare economics to accommodate these trade-offs is to employ the compensation tests proposed independently by Kaldor (1939) and Hicks (1939). These compensation tests are based on the idea that a project raises welfare if the gainers are able to compensate the losers and still remain better off than without the project. Consider the project introduced previously that changed the consumption levels in the economy from \([x^1, x^2, \ldots, x^H]\) to \([x^1 + \Delta^1, x^2 + \Delta^2, \ldots, x^H + \Delta^H]\).

According to the Kaldor compensation principle, the project is desirable if it is hypothetically possible to redistribute income (and hence consumption) so that everyone is made better off with the project. This requires the gainers to compensate the losers. Alternatively, according to Hicks’s compensation principle, the project is desirable if it is not possible for the losers from the project to bribe the gainers to forgo the project.

Formally, consider the individual consumption bundle \([x^h, y^h]\), where \(y^h\) distinguishes a numéraire good that is to be used for compensation. Suppose that the individual utility functions are quasi-linear with respect to the numéraire good, so that \(U^h(x^h, y^h) = v(x^h) + y^h\). Define the set of compensation transfers as \(T = \{(t^1, \ldots, t^H) \mid \sum_{h=1}^H t^h \leq 0\}\). The two tests can then be stated:

- **Kaldor’s test** Can the gainers compensate the losers after the project is implemented?

  This test implies the project is desirable if there exist transfers \((t^1, \ldots, t^H) \in T\) such that \(U^h(x^h + \Delta^h, y^h + t^h) > U^h(x^h, y^h)\) for all \(h = 1, \ldots, H\).

- **Hicks’s test** Can the losers compensate the gainers to forgo the project?

  Under this test, the project is desirable if there do not exist transfers \((t^1, \ldots, t^H) \in T\) such that \(U^h(x^h + \Delta^h, y^h + t^h) > U^h(x^h + \Delta^h, y^h)\) for all \(h = 1, \ldots, H\).

  It is possible to combine these two tests. Doing so gives the Scitovsky test:

- **Scitovsky’s test** The project is desirable if both the Kaldor and Hicks tests are satisfied.

  The compensation tests demand those who gain to sufficiently compensate those who lose so that they will accept the project or, alternatively, those who lose to bribe those who gain so that they forgo the project. Both compensation tests are based on the hypothesis that efficiency can be treated separately from the question of distribution. This is because the tests treat all gains and losses identically, regardless of whom they accrue to.
The problem with these tests is the possibility of reversal: it is possible for a project to pass the test, and then the opposite of the project (a move back to the original point) to also pass the test. To illustrate this point, compare an existing state $A$ with a new state $B$ that can be achieved after implementation of a project. The existing state $A$ generates the distribution of welfare between consumers 1 and 2 as represented in figure 25.4 by point $A_0$ on the associated utility possibility frontier. Now consider state $B$ with a new distribution of welfare between the consumers represented by $B_0$ on the new utility possibility frontier. Clearly, implementing the project and moving from $A_0$ to $B_0$ makes consumer 2 better off and consumer 1 worse off, so it is not a Pareto improvement. However, it is possible to compensate consumer 1 so as to reach point $B_1$ with the new project, at which point both consumers are better off and the project will pass the compensation test. Note that we only need to verify that consumer 2 can compensate consumer 1 to pass the test. Consumer 2 does not actually have to make the transfer. The problem of reversal follows from observing that a move back to the original state $A$ will also pass the compensation test. To see this, consider moving back to state $A$ with compensation to move along the utility possibility curve from $A_0$ to $A_1$. Again, both consumers will then prefer $A_1$ to $B_0$. So the reverse of the project will also pass the test after the project is implemented. The fact that there can be cycles is also called the Scitovsky reversal paradox. This reversal arises because the test is based on the hypothetical allocations $A_1$ and $B_1$ achieved after compensation whereas only allocations $A_0$ and $B_0$ are actually achievable.

![Figure 25.4](image-url)  
Scitovsky reversal
An alternative to the compensation tests is to obtain a value of the project from each person that it affects. The value will be positive if the person gains from the project but negative if they lose. Such values can be related to the concepts of *compensating variation* and *equivalent variation*.

Using the same notation as above, let the consumption bundle of individual $h$ be $\{x^h, y^h\}$, where $y^h$ is the numéraire good, and assume utility is quasi-linear $U^h(x^h, y^h) = v(x^h) + y^h$. Consider a project that changes the consumption levels from $\{x^h_0, y^h_0\}$ to $\{x^h_1, y^h_1\}$ with respective utility levels $U^h_0$ and $U^h_1$. The compensating variation measures how much must be given to a person to compensate them for the project being implemented. The compensating variation is defined as

$$CV^h((x^h_0, y^h_0), (x^h_1, y^h_1)) = y^h_1 - y^h_0, \quad U^h(x^h_0, y^h_0) = U^h_0.$$ (25.20)

Hence a project that raises utility must be compensated by a reduction in the consumption of the numéraire good to maintain utility unchanged, so that $CV^h = y^h_1 - y^h_0 > 0$. That is, a good project ($x^h_1 > x^h_0$) has a positive compensating variation. The equivalent variation measures how much must be given to a person to forgo the project. The equivalent variation is defined as

$$EV^h((x^h_0, y^h_0), (x^h_1, y^h_1)) = y^h_1 - y^h_0, \quad U^h(x^h_0, y^h_0) = U^h_1.$$ (25.21)

Hence a project that raises utility is equivalent to increasing the consumption of numéraire good, so that $EV^h = y^h_1 - y^h_0 > 0$. That is, a good project ($x^h_1 > x^h_0$) has a positive equivalent variation. The equivalent variation is the willingness to accept (WTA) the project.

Note the difference between our definition of compensating variation over consumption changes, and the more standard definition of compensating variation over price changes. Indeed the compensating variation of a price change is defined as $CV = E(p_1, U_0) - E(p_0, U_0)$, where the expenditure function $E(p, U)$ is the minimal expenditure at given prices $p$ to reach the utility level $U_0$. Thus the CV of a price change from $p_0$ to $p_1$ is positive if this price change raises the cost of obtaining utility $U_0$. So a bad outcome ($p_1 > p_0$) has a positive CV. Similarly the equivalent variation of a price change is defined as $EV = E(p_1, U_1) - E(p_0, U_1)$, which is again positive for a bad outcome, $p_1 > p_0$. It is the income that can be taken off the person to give $U_1$ at initial prices $p_0$. 
The notions of \textit{WTP} and \textit{WTA} are firmly grounded in the theory of welfare economics and correspond to notions of compensating and equivalent variations. \textit{WTP} and \textit{WTA} should not, according to theory, diverge very much. In practice, they appear to diverge, often substantially, with \textit{WTA} > \textit{WTP}. Hence the choice of \textit{WTP} or \textit{WTA} may be of importance when conducting a CBA. The difference between the \textit{WTP} and \textit{WTA} can be highlighted by using these definitions to express the compensation tests as follows:

\textbf{Compensation tests} \quad The Kaldor test is positive if and only if \(\sum_{i=1}^{n} CV_i > 0\), and the Hicks test is positive if and only if \(\sum_{i=1}^{n} EV_i > 0\).

This is as far as we can proceed using this direct approach of welfare valuation unless we specify a particular social welfare function. To do so, it is necessary to be explicit about how society forms social preferences. “Society” is simply the sum of individuals, so social welfare is obtained by aggregation of individual welfare. There are two alternative philosophies for undertaking the aggregation of benefits across different social groups or nations. These philosophies imply very different approaches to the calculation of welfare change.

The first philosophy is that all benefits and costs should be treated the same way. Aggregation then involves simply summing the willingness to pay for benefits, or willingness to accept compensation for losses, regardless of the circumstances of the beneficiaries or losers. This is precisely what we did with the \textit{WTP} and \textit{WTA} above. The implication is that the evaluation of the project has been turned into a comparison that involves a numéraire good. This numéraire can be thought of as money. That money is given the same value for all people is to claim that “a dollar is a dollar” whoever receives it or pays it. The validity of this claim has been widely debated. Using a welfare function assumes that individuals have different marginal utilities of income and different social welfare weights (unless the optimal distribution of welfare is reached). Then, by the assumption of quasi-linearity, everyone is assigned the same marginal utility, and the differences in social welfare weights are ignored in the compensation tests.

The second philosophy for aggregation requires that higher weights be given to benefits and costs accruing to disadvantaged or low-income groups. One rationale for this second form is that marginal utilities of income will vary, by being higher for the low-income groups. If this is accepted, the approach we have outlined in this section cannot be applied. The next sections investigate what can be done.
25.7.2 Distributive Weights

We have previously noted that social welfare should be based on individual preferences. It is a basic requirement of democracy that individual preferences should count for social decisions. A second key requirement in social decision rules is that those individual preferences should be weighted to reflect the existing distribution of income when the distribution is not socially optimal. If a society cares about equity, then the social decision rule will give more weight to the preferences of those with low incomes. When undertaking CBA, the important issue of equity, or the distributional incidence of costs and benefits, should not be overlooked. Incorporating distributional concerns implies initially identifying, and then possibly weighting, the costs and benefits of individuals and groups on the basis of differences in some characteristic of interest (e.g., income or wealth).

Consider the following simple setting. Denote the (ordinary) demand function of individual \( h \) by \( x^h = x^h(p, m^h) \), where \( p \) are the market prices of goods (identical for all individuals) and \( m^h \) is the income of \( h \). This demand function can be substituted into the utility function to obtain the indirect utility function (maximum utility level attainable given income and price levels). Denote the indirect utility function for individual \( h \) as \( V^h(p, m^h) \). Consider a project as equivalent to a policy change with marginal impact on incomes and prices \( (dp, dm^h) \). Then the welfare change of the project for individual \( h \) is

\[
dV^h = \frac{\partial V^h}{\partial m^h} dm^h + \frac{\partial V^h}{\partial p} dp,
\]

(25.22)

where \( \frac{\partial V^h}{\partial m^h} \) is the marginal utility of income of individual \( h \). Dividing the utility change by the marginal utility of income gives the willingness to pay for the project for individual \( h \):

\[
WTP^h = dm^h + \frac{\partial V^h}{\partial V^h/\partial m^h}.
\]

(25.23)

The willingness to pay is the compensating variation, and the willingness to accept is the equivalent variation. Both welfare evaluations of the policy change are the same when the project is marginal (so \( CV = EV = WTP = WTA \)).

Now, to evaluate the social welfare change resulting from the project, consider a general social welfare function \( W(V^1, \ldots, V^n) \). Then the social welfare change of the project is
\[ dW = \sum_{h=1}^{n} \beta_h WTP^h, \quad (25.24) \]

where \( \beta_h = \frac{\partial W}{\partial V} \frac{\partial V}{\partial m_i} \) is the social marginal utility of income for individual \( h \). Hence the social welfare change of the project is the weighted sum of the individual willingness to pay for the project where the weights are the social marginal values of income for different individuals. This is Meade’s formula for CBA. It follows that projects with a change of allocation that favors individuals with greater social marginal value may improve social welfare even if the total willingness to pay for the project is negative!

How is it possible to compute those social weights in practice? The common practice is to adopt the function \( \beta_h = K \left[ m^h \right]^{-\varepsilon} \) used in chapter 15 (see equation 15.28) where the parameter \( \varepsilon > 1 \) represents the social aversion to inequality.

### 25.7.3 Market Prices and Shadow Prices

Cost–benefit analysis is built on the assumption that the monetary valuation of a project reflects the welfare consequences. In more formal terms, what we require is that the implementation of a project with a positive \( NPV \) will increase social welfare, and that any project that would raise social welfare has a positive \( NPV \).

To explain what is involved, it is necessary to distinguish between two different aspects of a project. The first aspect is the real effects that the project will have. These real effects are measured in terms of a change in quantities. For example, for a flood defense scheme the change in quantities would include the labor and materials that are used, and the property that goes undamaged because of a successful defense. In formal terms the change in quantities are the values \( \Delta_1 \) that enter (25.19). The second aspect is the social value of these changes. We have so far considered the process of aggregating individual values to find social values. We now consider an alternative approach. This approach uses prices to obtain a value for the quantity changes. If both quantity changes and the prices are known, then multiplying the two will give the cost and benefit values to be inserted into the \( NPV \) equation.

As a starting point for our formal development of cost–benefit analysis, we will assume that we know the change in quantities that will arise due to the project. This is helpful but not always realistic. The next chapter will discuss climate change, which is the primary example of where we do not know with any degree of certainty what the effects of alternative projects will be. Accepting this assumption for the present, what remains to be determined are the prices that should be used to value the quantity changes.
The natural solution is to use observed market prices. There are two problems that arise immediately with this proposal. First, we have already observed that there may be nonmarket goods, such as environmental quality, that do not have observable market prices. We acknowledge this problem but assume that it can be resolved using the methods we have discussed earlier. The second problem is based on the role that we wish prices to play. For the NPV to correspond to the change in welfare requires that the prices used in the valuation capture social valuations. There are circumstances where market prices have this property, as we now demonstrate, but this is not always the case.

The relationship between market prices and social values can be explored by returning to the competitive economy of chapter 2. The Two Theorems of Welfare Economics can be summarized as saying that in the absence of market failure a competitive economy is Pareto-efficient, and that any Pareto-efficient allocation can be decentralized as a competitive equilibrium. Furthermore it was shown in chapter 13 that the application of the decentralization invoked by the Second Theorem allowed the social optimum to be obtained. The implications of these statements for CBA are now investigated in a two-consumer, two-good economy.

Denote the utility function of consumer $h$ by $U^h(x^h_1, x^h_2)$ and the social welfare function by $W(U^1, U^2)$. With a fixed stock, $X_i$, of good $i$ the social optimum solves

$$\max_{\{x^1_1, x^1_2\}} W\left[U^1(x^1_1, x^1_2), U^2(X_1 - x^1_1, X_2 - x^1_2)\right].$$  \hfill (25.25)

The optimization in (25.25) has necessary conditions

$$\frac{\partial W}{\partial U^1} \frac{\partial U^1}{\partial x^1_i} = \frac{\partial W}{\partial U^2} \frac{\partial U^2}{\partial x^2_i}, \quad i = 1, 2. \hfill (25.26)$$

If the social optimum is decentralized as a competitive equilibrium, then each individual is maximizing utility subject to a budget constraint. This implies that marginal utility must be proportional to market prices, or

$$\frac{\partial U^h}{\partial x^h_i} = \alpha^h p_i, \hfill (25.27)$$

where $\alpha^h$ denotes the private marginal utility of income for consumer $h$. Combining (25.26) and (25.27) gives

$$\frac{\partial W}{\partial U^1} \alpha^1 p_i = \frac{\partial W}{\partial U^2} \alpha^2 p_i, \quad i = 1, 2. \hfill (25.28)$$
The equality in (25.28) is key to demonstrating that CBA can use market prices in this setting.

Cancel $p_i$ from both sides of (25.28), and recall the definition of the marginal social utility of income for $h$, $\beta^h \equiv \frac{\partial W}{\partial U}a^h$; see (15.26). The equality in (25.28) shows that

$$\beta^1 = \beta^2 = \beta.$$  

(25.29)

Hence, the marginal social utilities of income are equalized at the social optimum. This equality implies the following observation: assume that the economy suddenly discovers an extra amount $dx_i$ of good $i$ (where this is meant to be interpreted as a very small increase in the total quantity). The increase in social welfare if this is allocated to consumer $h$ is

$$\frac{\partial W}{\partial U^h} \frac{\partial U^h}{\partial x^h_i} dx_i = \beta p_i dx_i,$$  

(25.30)

which is the same for both consumers at the social optimum. The two features of the social optimum (Pareto-efficiency and optimal distribution) combine to ensure that the increase in social welfare will be the same regardless of how this extra unit is allocated among the consumers.

The formal concept of a project is now introduced. For the purpose of the analysis it is assumed that a project is "small," so that it is appropriate to work with derivatives. In practical applications of CBA many projects really are small when compared to the economy as a whole. Some are clearly not, and the arguments have to be modified accordingly. A project is a change in the supply of the two commodities $dx = (dx_1, dx_2)$. For example, the project $dx = (2, -3)$ would represent 3 units of good 2 being used as inputs to produce an additional 2 units of good 1. Let the change in consumption levels for $h$ because of the project be $(dx^h_1, dx^h_2)$. The project $dx$ will raise social welfare if

$$dW = \sum_{h=1}^{2} \left[ \frac{\partial W}{\partial U^h} \frac{\partial U^h}{\partial x^h_1} dx^h_1 + \frac{\partial W}{\partial U^h} \frac{\partial U^h}{\partial x^h_2} dx^h_2 \right] > 0.$$  

(25.31)

We now use (25.30) and the fact that $dx_i = dx^h_1 + dx^h_2$ to write

$$dW = \sum_{h=1}^{2} \beta^h \left[ p_1 dx^h_1 + p_2 dx^h_2 \right] = \beta \left[ p_1 dx_1 + p_2 dx_2 \right].$$  

(25.32)

The project will raise welfare if
Part IX: Applications

\[ dW = \beta [p_1dx_1 + p_2dx_2] > 0. \]  \hspace{1cm} (25.33)

Since \( \beta > 0 \), (25.33) shows that \( dW > 0 \) if

\[ NPV = p_1dx_1 + p_2dx_2 > 0 \]  \hspace{1cm} (25.34)

So the project has positive value at market prices. This is the central conclusion of the analysis: at the social optimum the implementation of a project that has a positive \( NPV \) computed using market prices will increase welfare.

Two points need to be made about this claim. First, the correspondence between \( NPV \) and social welfare arises only because the correct prices have been used in the valuation. In this case market prices could be used because we assumed a social optimum with the optimal distribution of income and with no market failure. The correspondence between \( NPV \) and social welfare will apply in other cases but only if the valuation is conducted using the appropriate prices. We shortly explore this question in more detail.

Second, the fact that the distribution of the costs and benefits of the project across people does not matter needs closer inspection. The correct conclusion is that the allocation does not matter when the appropriate prices are used in the valuation. However, the process of allocation will determine what those price are.

To explore the role of the allocation process in more detail, it is helpful to look again at the analysis. The argument above required both the Pareto-efficiency of equilibrium and the socially optimal distribution of income. So Pareto-efficiency alone is not enough to ensure that market prices are used directly in a cost–benefit analysis. If there is not a socially optimal distribution of income, then the effect on social welfare of an extra unit of the commodity will depend on who it is allocated to (the division of the costs and benefits of the project between people). When these assumptions do not hold—and given their severity they are unlikely to do so—it is necessary to use what are called \textit{shadow prices} to calculate the \( NPV \). Shadow prices have the property that they represent the social value of a change in quantities. They can sometimes be equal to market prices, but in most situations they are different. The key point is that when shadow prices are used in a CBA calculation, the implementation of a project with a positive \( NPV \) will necessarily raise welfare. As we will show, this correspondence is achieved by directly linking the definition of shadow prices to an increase in welfare.

\textbf{Definition} A set of shadow prices are such that any project with a positive \( NPV \) will raise social welfare if it is implemented. Conversely, any project that will raise welfare if implemented will have a positive \( NPV \) when evaluated using shadow prices.
The process for obtaining the shadow prices can be explained by studying the welfare effects of implementing a project. The setting for this is an economy with production. The government chooses policies to maximize welfare. The demands for the private sector are represented by an excess demand function; see (2.10). The level of excess demand depends on the policy of the government, the distribution of the initial endowment, and the resource allocation process. It is now demonstrated how the shadow prices emerge naturally from the optimal choice of policy by the government.

Let \( s = (s_1, \ldots, s_K) \) be the policy variables of government, \( z = (z_1, \ldots, z_n) \) be government net supplies of the \( n \) commodities, and \( \omega = (\omega_1, \ldots, \omega_n) \) be the initial endowment of commodities. The level of welfare is a function \( W(s; \omega) \) that is dependent on policy and endowment. There are \( H \) households and \( m \) firms. Let the excess demand from the private sector for good \( i \) be \( Z_i(s; \omega) \), so that

\[
Z_i(s; \omega) = H \sum_{h=1}^{H} x_i^h (s, \omega_h) - \sum_{j=1}^{m} y_i^j (s) - \omega_i, \quad (25.35)
\]

where \( x_i \) denotes the demand and \( y_i \) the supply of good \( i \). Equilibrium requires that private excess demand minus government net supply be zero:

\[
Z_i(s; \omega) - z_i = 0. \quad (25.36)
\]

The Lagrangian for the optimal choice of policy is

\[
\mathcal{L} (s; \omega, z) \equiv W(s; \omega) + \sum_{i=1}^{n} \nu_i [z_i - Z_i(s; \omega)]. \quad (25.37)
\]

where the equilibrium conditions act as constraints on policy and \( \nu_i \) are the Lagrange multipliers associated to market equilibrium for each good \( i \). The necessary conditions for choice of policies are

\[
\frac{\partial W(s; \omega)}{\partial s_k} - \sum_{i=1}^{n} \nu_i \frac{\partial Z_i(s; \omega)}{\partial s_k} = 0, \quad k = 1, \ldots, K. \quad (25.38)
\]

Denote the chosen policy by \( s^* \) and the maximized level of welfare by \( W^*(\omega, z) \). Then differentiation of the Lagrangian evaluated at the optimal policy gives

\[
\frac{\partial \mathcal{L} (s^*; \omega, z)}{\partial z_i} = \nu_i. \quad (25.39)
\]
The envelope theorem states that the effect on the Lagrangian of a change in \( z \) is the same as the effect on maximum value, so that
\[
\frac{\partial W^*(s^*: \omega, z)}{\partial z_i} = \upsilon_i. \tag{25.40}
\]
This produces the conclusion we require. If the shadow price is taken to be the Lagrange multiplier on the resource constraint, then it follows that the price captures the welfare effect of a marginal increase in government supply of good \( i \).

This can be seen by observing that the NPV of a project \( dz \) at these shadow prices is given by
\[
V = \sum_{i=1}^{n} \upsilon_i dz_i = \sum_{i=1}^{n} \frac{\partial W^*(s^*: \omega, z)}{\partial z_i} dz_i. \tag{25.41}
\]
The NPV is therefore just the total derivative of the welfare function, so there is a direct correspondence between the two. By contrasting (25.33) and (25.41), we can see that the shadow prices for the social optimum case are actually \( \upsilon_i = \beta p_i \).

The practical application of cost–benefit analysis develops the insights of this general theory into detailed conclusions for particular cases. The analysis of each of these cases follows the logic that has been applied in this example. A project for the government is a change in its net supply vector. The policy variables, \( s \), are changed optimally to match the new net supply. These two effects determine the new equilibrium allocation via the operation of the market. The shadow prices, and the ultimate effect on welfare, therefore depends on both the project and the policy variables that the government has under its control. For example, a project implemented by a government that is able to target the output of the project directly at some consumers (through a benefit-in-kind which is means tested) will have a different welfare effect than a project where the output is placed on the open market. So policy is interlinked with shadow prices. The key point is that available policies are interlinked with shadow prices.

There is one final point to observe. The model that has just been described can be given an intertemporal interpretation. This can be achieved by dating the commodities (as discussed in chapter 2) and treating the individual utility functions as describing intertemporal preferences. The shadow prices will then capture social preferences over consumption at different points in time. Conceptually this is straightforward. In practice it raises a number of issues as were discussed in section 25.6.
25.8 Conclusions

CBA is an evaluation procedure firmly grounded on solid theoretical foundations of welfare economics. However, actual decisions may in practice be made on a very different basis from this analytical approach. The reasons lie in the role played by political forces rather than the social welfare functions of economics. First, there is widespread distrust of, or disbelief in, measurement in monetary terms concerning social costs and benefits of different projects and policies. Second, there is the capture of political processes by individuals not trained in economics. Third, there is the belief that economics is equivalent to common sense and easily understood. Fourth, there is a genuine mistrust of CBA and its theoretical foundations based on the debates that continue within the CBA community and outside it. But explaining the gap between actual and theoretical design is not an excuse to justify the gap. Economists need to develop a far better understanding of the pressures that affect actual decisions and the distributional impacts of the project for the many different stakeholders. But those who make actual decisions perhaps also need a far better understanding of economics, and it was the purpose of this chapter (as well as the rest of the textbook) to provide the basic theory underlying public policies and their impact on social welfare.

Further Reading

The early manuals in the late 1960s on project evaluation are:

A useful short guide for cost–benefit analysis is:

The classic texts still worth reading are:

A persuasive defense of the use of CBA in social decisions is given by:
Part IX: Applications


Some more recent references are:


On discounting the distant future, see:


Exercises

25.1 A project has an initial cost of $300, and yields benefit of $80 for 5 years. It then has a decommissioning cost of $60 in year 6.

a. Should the project be implemented if the discount rate is 5 percent?

b. What is the internal rate of return for the project?

c. Is the project sensitive to the decommissioning cost?

25.2 Three investment projects have the cash flows detailed in the table. Is it possible to rank these projects without specifying the discount rate?

<table>
<thead>
<tr>
<th>Period</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project A</td>
<td>$-2,000</td>
<td>400</td>
<td>400</td>
<td>800</td>
<td>1,200</td>
</tr>
<tr>
<td>Project B</td>
<td>$-1,200</td>
<td>0</td>
<td>0</td>
<td>800</td>
<td>1,200</td>
</tr>
<tr>
<td>Project C</td>
<td>$-2,000</td>
<td>0</td>
<td>0</td>
<td>1,200</td>
<td>1,600</td>
</tr>
</tbody>
</table>

25.3 Spent nuclear fuel rods must be stored in a protective environment until radiation emissions no longer pose a problem. Assume that this period is 1,000 years and each pound of nuclear fuel costs $10 a year to store. If fuel rods can be used in a reactor for 5 years, what must be the value of the annual energy produced to make the process worthwhile when the discount rate is 2.5 percent?

25.4 Public-investment decisions share three important characteristics. First, to some extent, these decisions may be irreversible; once committed, the resources cannot be recovered. Second, the outcome of the investment may be uncertain because the input data are uncertain. Third, there may be some leeway with respect to the timing of the investment. Discuss briefly.
25.5 Determine the internal rate of return (IRR) for each of the projects described in the table. Would either project be accepted if the required IRR was 1.75 percent?

<table>
<thead>
<tr>
<th>Period</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project 1</td>
<td>−10,000</td>
<td>0</td>
<td>7,000</td>
<td>0</td>
<td>3,500</td>
</tr>
<tr>
<td>Project 2</td>
<td>10,000</td>
<td>0</td>
<td>−3,500</td>
<td>0</td>
<td>−7,000</td>
</tr>
</tbody>
</table>

25.6 Consider a project that costs $5 million and involves building a dam and canal to provide water to irrigate desert land that has no other use. If irrigated, the land will be used to grow cucumbers. Suppose that the demand function for cucumbers is known to be

\[ q = 18 - 0.1p, \]

where \( q \) is in units of millions of pounds per year and \( p \) is the price of cucumbers in cents per pound. Suppose that the project produces 0.5 million pounds of cucumbers per year. Without the project, the quantity of cucumbers available is 6 million pounds per year. With the project the quantity is 6.5 million pounds per year.

Suppose that the only input required to produce cucumbers, besides the land and water, which have no opportunity cost, is labor, and each 10,000 pounds of cucumbers requires 1 person-year of labor. This means the project’s 0.5 million pounds of cucumbers require 50 person-years of labor. Suppose also that the supply function for labor in the area of the project is

\[ L = -20,000 + 4w, \]

where \( L \) is person-years of labor supplied and \( w \) is the wage rate in dollars per person-years. Without the project, the amount of labor utilized is 2,000 person-years. With the project, the labor demand is 2,500 person-years.

a. What is the gross benefit of the project?

b. What is the net benefit of the project?

c. What is the impact of the project on consumer and producer surplus?

25.7 Queen Elizabeth I is reported to have said on her deathbed, “All my possessions for an instant of time.” What are the implications for the question of the value of a life? What are the three methods used by economists to value life?

25.8 Assume utility when alive is \( U(M, a) = \ln(M) \) and that utility in the case of death is \( U(M, d) = \ln\left(\frac{M}{c}\right) \), where \( c > 1 \). Find the WTP to avoid death, and show that the VSL is

\[ \text{VSL} \equiv \frac{dp}{dp} = \ln(c) \left[M - \rho\right]. \]

Assuming that \( c = 4 \) and \( \rho = 0.001 \), evaluate the VSL for \( M = 10,000, 100,000, \) and \( 1,000,000 \). Hence discuss whether it is annual income or lifetime income that should enter this calculation.

25.9 In “Choosing Priorities” (Journal of Medical Ethics, 1979) Muir Gray writes: “The strength of cost–benefit analysis, or any other concept is a function of its weakest point, which is that
it attempts to place a monetary value on human life,” which he claims “is not like the value of sheet steel, ball bearings, or any of the other commodities for which cost–benefit analysis is usually employed. It cannot be expressed in monetary terms.” Is this correct?

25.10 All workers have preferences with constant relative risk aversion, \( U = \frac{M'}{\gamma} \), but differ in the value of \( \gamma \). \( \gamma \) is uniformly distributed on the interval \([0, g]\) (\( \gamma = 0 \) being risk neutrality). There are two occupations. Occupation 1 is safe and pays a (normalized) wage of 1. Occupation 2 has a risk of death of \( p \) and pays a (relative) wage of \( w \). Labor demand requires that one-half of all workers be employed in each occupation. Assume that the utility in the event of death is 0.

a. Find the equilibrium wage, \( w \), that clears the labor market.
b. Find the WTP to avoid death of the median worker (ranked by risk aversion) in occupation 2.
c. Use the median WTP to define the VSL. How does the VSL depend on \( g \)? Discuss this result.

25.11 The Environmental Protection Agency (EPA) is conducting a study on the efficiency of its expenditures on environmental protection programs. You have been hired to work on the analysis with the aim of deciding if money should be reallocated among the programs. You are presented with the following information:

i. $20 million is spent to reduce airborne particulates. This is estimated to save 2 people from death and 1,000 from illness during periods with very low wind speeds; once wind speeds increase, the particulates are dispersed with no lasting effects. Those affected are primarily those with preexisting breathing problems and can be of any age.

ii. $40 million is spent to reduce the exposure of children to radiation in schools. This is estimated to prevent 20 children per year from developing cancers when they are middle-aged. Then 40 percent of the cancer cases will be fatal; the remainder survive as normal after treatment.

iii. $10 million is spent to reduce hormones levels in beef. This is estimated to save 20 people from death and 1,000 from other complications. 60 percent of those affected are children.

iv. $4 million is spent to discourage people from jaywalking. This is expected to save 100 people from death and 200 from other serious injuries each year. There is no particular relation between jaywalking and age.

a. If the value of a statistical life (VSL) is $3 million, and the value of avoiding a serious injury or illness is $50,000, which of the programs has positive net benefits?
b. Which program maximizes the saving of lives per dollar of expenditure?
c. Assume that the value assigned to serious injury or illness reflects quality-adjusted life years (QALY). Which program saves the most QALY’s?
d. What recommendation would you make on the reallocation of funding among these programs?

25.12 In cost–benefit analysis, we normally consider market prices as being good measures of the costs and benefits of an investment. (When market prices do not exist in usable form, the analyst has to construct them). Frequently, however, the market price (if any) is only an approximate measure of a cost or benefit. Explain why. Under what condition is such approximation acceptable or not?
Chapter 25: Cost–Benefit Analysis

25.13 The Venice of the Midlands visitor center located on the canal system in Rummidge does not attract many visitors. In the past year no more than 5,000 visitors viewed the exhibition at the center. When the Venice of the Midlands center was purchased by the government from a private theme park company, it cost $200 million.

a. What is the value of benefit for each visitor if the $200 million correctly reflects the discounted value of the annual flow of benefits from the center and the discount rate is 5 percent? Is this a realistic value for benefits?

b. Assume that each visitor obtains benefits of $50 from the center. Is it socially beneficial for the center to lay on an entertainment program that raises visitor numbers to 150,000 per year but has an annual cost of $125,000?

c. How would these answers change if the center also played a key role in preserving the historical legacy of canals in Rummidge?

25.14 "Kelsey asks, ‘Given that a cost–benefit analysis is going to be performed anyway, what could be cheaper than changing the weights used in the calculations?’ However, the use of unequal weights would sanction projects of lower aggregate net benefits. These are the efficiency (or distortion) costs involved.” (Y. K. Ng, Oxford Economics Papers, 1988.) Comment on these observations.
26.1 Introduction

Climate change caused by human activity is possibly the most critical threat that the world economy has ever faced. The continued accumulation of carbon dioxide (CO₂) and other greenhouse gases in the atmosphere is predicted to cause a significant increase in global temperatures and disruption to weather patterns worldwide. Rising temperatures will melt polar ice caps and low-lying areas will be flooded due to rising sea levels. Changing rainfall and snow patterns will cause widespread drought, particularly in important agricultural regions. The shortage of food and water will lead to major displacements of populations and civil unrest. Extreme weather events will become more common and lead to frequent extensive damage to property and significant loss of life. The enormity of these effects has lead to calls for decisive and immediate action to reduce emissions of CO₂.

Not everyone agrees that all of these effects will occur. The scientific consensus is that temperatures will rise and lead to damaging changes in weather patterns. Within the consensus there is a range of opinion on precisely how strong the effects will be and what level of policy response is required. The potential scale of the effects of climate change does not necessarily imply that major sacrifices should be made now to combat climate change. An economic analysis of climate policy should be based on the same methods applied to other policy questions. The appropriate analytical tool is the cost–benefit analysis (CBA) we developed in chapter 25 and, in particular, the trade-off between the cost of changing behavior now against the benefits that will accrue in the future. The issues involved in climate policy—notably the degree of uncertainty and the timescale—raise important questions about the application of CBA. Any evaluation process must assess the extent to which the climate will change and the damage that this will cause. The assessment has to be guided by scientific input from other disciplines, and utilize economists and political scientists to determine the economic and social costs. The science is important since the appropriate policy response to anthropogenic (human-made) climate change is very different to the response to climate change as part of a natural cycle.

It needs to be acknowledged that the consensus opinion on climate change is not universally accepted. Four separate arguments have been advanced through which the consensus is disputed. First, the accumulation of CO₂ may not be anthropogenic but instead be part of a natural cycle. The high levels of atmospheric CO₂ in the
distant past are cited as support for this claim: for example, in the late Ordovician period, approximately 444 million years ago, CO₂ levels were around 5,000 parts per million (in contrast, current CO₂ levels are 400 parts per million). Second, it is possible that the increase in atmospheric CO₂ will not affect climate in any significant way. If temperatures do not rise, then weather patterns will not change, and none of the adverse effects of climate change will occur. This argument revolves around the level of climate sensitivity, which is a measure of how much temperature responds to accumulation of greenhouse gases (we discuss this in more detail below). Third, even if CO₂ does affect climate it is perfectly possible that technological advancement will rapidly mitigate the effects. There is already significant development of “clean” energy sources and the beginnings of a move toward alternative power sources for road vehicles. Fourth, the evidence for warming is directly disputed by most recent measurements that show, contrary to the warming hypothesis, average temperature is not increasing and that reconstructions are open to doubt.

We make no effort here to provide an opinion on any of these debates. Instead, we emphasize only that good science requires an open mind and the willingness to continual questioning of the evidence, no matter how much support there currently is for a particular position. No scientist should forget that the Ptolemaic system agreed with observations—until those observations were refined (and that Galileo was sentenced to a lifetime of house arrest by the Inquisition for supporting the alternative heliocentric system). Similarly Newtonian mechanics fitted every known observation until relativity and quantum mechanics revealed its limitations for the very fast and the very small. With this in mind, the point we want to stress is that the central issue is about policy design when there is uncertainty about causes, effects, and consequences. There is no doubt that CO₂ levels are rising. The points at issue are what is causing this rise and what the consequences of the rise will be. The current understanding of climate science points to the answer but is by no means definitive. For an economist the limitation of the science is compounded with the difficulties of linking climate change to economic consequences.

This is not a chapter about the science of climate change. We cover the science only to the extent necessary to reveal the issues that economic policy formulation must confront. It is a chapter about how economics should be applied given the current state of the science and, especially, the limitations of the economic applications. The disputes about the science are not the issue but instead are the backdrop to the economics. One thing is clear: the doubts about the science ensure that there is little basis for making economic predictions with any degree of confidence. Moreover it is hard to see how to even begin to correctly model what the economic effects will be. This has not prevented people from trying and, in some cases, even making confident announcements about
Chapter 26: Economics of Climate Change

the economic costs. We will review the method of modeling the attempts to quantify the economic costs as an illustration of how economics can be applied. If this does no more than reveal the current limitations of the economic analysis, it will at least have revealed the true situation behind the public facade.

This chapter is an application of economic ideas from earlier chapters to climate change. It is intended to show how the economics can be used to address significant questions, even if the answers are not clear-cut. We begin with a discussion of what is special about climate change compared to other economic problems. We then progress to a very brief review of the science and note some of the disputes about this science. This is followed by a review of how the interaction between climate and economics has been modeled. An insight into the form of the models is essential for assessing the predictions of economic damage from climate change and the gains from policy intervention. Attention then turns to the application of cost–benefit analysis to climate change and, in particular, the choice of discount rate. The chapter is completed by a review of current policies.

26.2 Special Features

Anthropogenic climate change can be categorized as the consequence of an economic externality, albeit a very major one. Individual consumers and firms burn fossil fuels and consume products that have fossil fuels as inputs in their production without taking into account the external effects of these activities. Each individual contribution may be small and inconsequential, but in aggregate they have a substantial effect. The discussion of chapter 8 demonstrated that externality problems can be solved by using either Pigouvian taxes or quantity controls. This section explores what is special about climate change that justifies an analysis beyond the application of these familiar policy solutions.

Four features of climate change can be identified that distinguish it from other externalities:

1. The problem is literally global in nature, so that an individual government can do little to affect the process.
2. There is considerable uncertainty about the cause, the degree, the effects, and the consequences of climate change.
3. The uncertainty about climate change feeds into even greater uncertainty about the economic damage it will cause and the potential benefits of policy intervention.
4. The process of climate change and the time scale of effects may be very long.
Each of these points is now discussed in more detail.

The emission of greenhouse gases anywhere in the world adds to the stock that is accumulated in the atmosphere, and dissipation of the gases throughout the atmosphere ensures that local emissions have a global effect. Unilateral action by the European Union, United States, and China could have a noticeable impact on the reduction accumulation, but a reduction of emissions by most other countries will have a negligible effect. This makes international coordination necessary for any effective response. The difficulties involved in achieving international agreements are readily illustrated. The 1997 Kyoto Protocol was signed and ratified by 191 countries, of which 37 agreed to reduce their collective emissions of the Kyoto gases by 5 percent from the 1990 level. Significantly, the largest single emitter of greenhouse gases—the United States—signed the Kyoto Protocol but did not ratify it, so was not bound by the commitment to reduce emissions. The Copenhagen Accord of 2009 was expected to produce a more rigorous successor to the Kyoto Protocol. The draft accord produced by the European Union promised significant reductions in emissions, but this was not accepted. The Copenhagen Accord was ultimately drafted by the United States in agreement with the BASIC countries (Brazil, South Africa, India, and China). It is not legally binding, nor does it commit countries to agree to a binding successor to the Kyoto Protocol. Effectively, the European Union was politically sidelined by China and the United States in negotiations.

It is helpful to think of the emission reduction process as a prisoner’s dilemma. If a country undertakes unilateral action to reduce emissions, it may benefit a little from the reduction in emissions but is likely to lose more due to the responses of firms and consumers to the policy. Consider, for example, a country that is securing a reduction in emissions by imposing a carbon tax on its domestic industries. This will cause a loss of international competitiveness and give an incentive to firms to relocate in other countries. These reactions impose costs that will almost certainly outweigh the benefit of a reduction in emissions. As a consequence unilateral action to cut emissions is not privately rational. In contrast, collective action by all countries can be socially rational. If a carbon tax is introduced by every country, then there is no incentive for the relocation of production. The countries can then receive the benefits of cutting emissions without the additional costs. The problem in a prisoner’s dilemma is sustaining the collective action: it will always be privately rational for a country to renege on the collective policy and to abandon the carbon tax. A country will gain directly when other countries cut emissions, and make a further indirect gain from an increase in international competitiveness. The reason for the United States not ratifying Kyoto, and the differences between Kyoto (which was a binding commitment)
and Copenhagen (which was not), are obvious when set within the framework of a prisoners’ dilemma.

There is inarguably considerable uncertainty in predictions of future climate change even if the level of greenhouse gases is taken as known. In section 26.3.3 we report on attempts to quantify this uncertainty, but it should be stressed that all such estimates should be treated cautiously. Climate is determined by a complex system, including a range of components such as rainfall patterns, snow patterns, winds, solar energy, and cloud patterns. All these elements interact to determine climate. Furthermore, climate predictions are forecasts for events that are outside previous experience. In economics this is called “out-of-sample” prediction. No matter how well a forecasting model may perform within the sample of observations, there is no guarantee that it will remain successful out of sample. It should also be noted that numerical weather prediction has developed only since 1950 which makes it a recent science. One contribution of numerical weather prediction has been the contention that weather is chaotic (in the formal sense of being sensitive to initial conditions). If so, weather necessarily becomes less predictable further into the future. Climate prediction may not suffer from precisely the same issues, but nonetheless, it remains a largely unsubstantiated science. Hence there is considerable uncertainty in translating temperature changes into climate effects.

This uncertainty is important for the economic analysis of climate policy. Section 8.6 analyzed the difference between prices and quantities as policy instruments when the outcome is uncertain. To apply that analysis, it is necessary to know how the marginal cost and marginal benefit curves are related to the uncertain variables. For example, if the cost of abatement is uncertain, then the curves need to be known for each possible value of abatement cost. It is also necessary to know the probability of different realizations of the uncertain variable. These are demanding requirements. Normally, we would accommodate the uncertainty by basing policy analysis on expected utility theory, but this requires knowledge of probabilities. With climate change the situation is more difficult because policy has to be chosen when even this information is not known: it is not even obvious that it is currently possible to say what future events might occur. This negates any possibility that we can assign probabilities to the events. From this perspective, the policy design is being done without benefit of applicable theory.

If there is to be an economic policy toward climate change, then it must be justified on the basis of benefits and costs. The uncertainties surrounding the effects of climate change make it exceptionally difficult to construct convincing estimates of the costs of not making any attempt at mitigation (and equally of the benefits of mitigation). In fact, even if there were a perfect way to forecast climate change, there would still be
exceptional difficulty in translating the effect on economic activity. It has been argued that the potential costs of climate change are so large that serious action is necessary even if the estimates later prove to have been exaggerated. This is a bad basis for a policy. Overreaching policy is not applied in any other area and would lead to absurd outcomes if it were. This does not mean we cannot or should not try to formulate policy, but it does mean that the uncertainties have to be part of the policy formulation process. We will describe below how these costs have been represented in economic models involving climate change. It will become clear that climate change is an area in which further development of technique is critically needed.

The effects of climate change are long term, so many future generations may be affected. This raises the question of how decisions made in the present can accommodate the interests of people who are not yet alive. Expressed differently, it is necessary to consider what right those currently alive have to inflict costs upon those who follow by not acting on climate change. The standard approach of cost–benefit analysis is to define a set of intertemporal preferences in which future utility flows are discounted. If this process is applied to long-term climate change the discounting effectively disenfranchises future generations. This has lead to a re-assessment of what is appropriate discounting in the context of climate change.

26.3 The Science of Global Warming

It is not possible to adequately address climate change, or to understand the issues involved, without having some knowledge of the science involved. This section briefly reviews the science and presents a summary of the evidence that is used by the International Panel on Climate Change (IPCC) to inform its conclusions. Some of the issues concerning this evidence are then addressed. The section is completed by a review of some predictions of temperature increases, but with a stress on their uncertainty as is central to policy design questions.

26.3.1 Greenhouse Gases and Temperatures

The atmosphere is more permeable to incoming solar radiation than the outgoing infrared, which traps heat within the atmosphere. This process is illustrated in figure 26.1. The amount of heat that is trapped depends on the composition of the atmosphere. There are some atmospheric molecules that trap heat—these are called the greenhouse gases. The basis of concerns about global warming is that human activity is increasing the concentration of greenhouse gases and so will lead to increased global temperatures.
Seven classes of greenhouse gases that accumulate in the atmosphere have been identified. These are carbon dioxide (CO₂), methane (CH₄), nitrous oxide (NO), perfluorocarbons (PFCs), hydrofluorocarbons (HFCs), sulphur hexafluoride (SF₆), and chlorofluorocarbon (CFCs). The first six of these are regulated under the Kyoto Protocol so are known as the Kyoto gases. The emission of CFCs is regulated separately under the 1989 Montreal Protocol. To allow easy comparisons, the quantities of the gases are converted into a CO₂ equivalent (CO₂e). The equivalence is defined by the mixture of gases having the same degree of radiative forcing—the extent to which they affect the balance between incoming and outgoing solar radiation—as the equivalent amount of CO₂. The quantity is then measured as the proportion of CO₂e in the air and expressed in parts per million (ppm) where

\[
ppm = \frac{\text{Mass of CO}_2e}{\text{Mass of air}} \times 1,000,000
\]

Figure 26.2 displays the quantity of greenhouse gases in CO₂e from years 1750 to 2008. A sustained increase can be observed in the amount of CO₂ and other greenhouse gases. The pre-industrial level (taken as that in 1750) was 278 ppm. Since the midnineteenth century atmospheric concentrations have grown from 290 ppm to about 470 ppm. This increase is attributed to human activities such as burning fossil fuels and deforestation. Over the past decade concentrations have been growing at about 2.5 ppm a year. Forecasts prepared for the IPCC suggest that the concentration will rise to over 600 ppm by the middle of this century and up to 900 ppm by the end of the century.
Figure 26.3 shows data on measured global average temperatures for the period 1850 to 2011. The figure plots the annual difference of average near-surface temperatures from the average level in the period 1961 to 1990. The temperature data are obtained from measurements of air temperature at a global network of land stations and sea-surface temperatures measured from ships and buoys. The figure displays two periods of warming: 1920 to 1940 and from the mid-1970s through to 2000. It is not possible to determine whether temperatures are still increasing because short-term fluctuations conceal the long-term trend in the data from 2000 onward. However, on the basis of this evidence it has been claimed that temperatures are rising at 0.2°C per decade.

Figure 26.4 shows a plot of the difference from the normal level in the period 1961 to 1990 from 800 through to 2000 based on a reconstruction of temperature data. This plot is reproduced from the Intergovernmental Panel on Climate Change Third Assessment Report. It was very prominent in that report, since it shows relatively constant temperatures in the pre-industrial period but a rapid, unprecedented rise since 1900 to current levels. The figure subsequently generated considerable controversy. The basis for the controversy is the fact that the figure was reconstructed from proxies (tree rings, corals, ice cores). There are serious scientific questions over whether the method of reconstruction is statistically valid. Credible arguments have been given that it is not, and that the figure is an artifact of the method used to construct it.
Figure 26.3
Global average temperature (annual average difference from normal)

Figure 26.4
Hockey stick plot
26.3.2 Controversies

There have been many objections to the science and the data that were just described. We choose to focus primarily on the temperature data since this has attracted the attention of economists. This is not surprising given the importance of the interpretation and analysis of data in economics.

The one nonstatistical controversy we note concerns the list of greenhouse gases. It is frequently argued that water vapor should be included because it has a substantial greenhouse effect. Clouds are particularly responsible for trapping heat (a cloudy night can be much warmer than a clear night) and so could have a large role in global heating. One reason why water vapor is not included is because the human contribution cannot be measured or distinguished from natural processes, and is probably negligible. But, even if this is the case, any modeling of the warming process needs to take into account cloud formation and the interaction of water vapor with the other greenhouse gases (GHGs). This is acknowledged to be a difficult task in climate modelling.

The temperature data in figure 26.3 show a process of warming that seems to stabilize after 2000. This occurrence can be better understood by plotting the series in more detail for later years. Figure 26.5 plots the monthly difference from normal for the period 1998/01 to 2011/07. Inspection of the figure shows no obvious trend, and this can be confirmed statistically. Hence the data since 1998/01 shows no increase or decrease over the period. What can be concluded from this observation? The response favorable to the climate change hypothesis is that these data series are too short to see what is happening in the longrun. This explanation might be acceptable if it were not for the fact that the climate change hypothesis is built on the claim of severe warming in the 50 years from 1950 to 2000 (see figure 26.3). The 13 years in figure 26.5 are not a small sample relative to the period in which warming is observed. Furthermore the concentration of GHGs reported in figure 26.2 show an ever increasing rate of accumulation. If there is a direct link between temperature and GHGs, then recent temperature increases should be even greater than in the past. The alternative response is that global warming has ceased, and that there is not actually any link between GHGs and temperature. In summary, the absence of a temperature increase in the recent past raises serious questions about the climate change hypothesis, and no adequate explanation has been proposed so far.

The additional controversies we wish to discuss involve arguments that are more formal and technical. It might seem surprising that academics from outside the field of climate change can make valid criticisms. However, outsiders have access to different tools and a different way of looking at problems, whereas those within become wedded to their own methods. It should also not be forgotten that economics has developed its
own set of advanced statistical methods that can be applied to the data. Doing so has revealed some of the issues involved in interpreting measurement figures.

The first set of questions focus around whether the reported average measurements are comparable across time. To be comparable, the measurements must be obtained under the same conditions and the sample should remain the same. There are two related issues. First, over time the data have been obtained from a decreasing number of measurement stations. This introduces a measurement bias if there is any consistent process behind the reduction in number. Second, the measurement stations that remain in the sample may be subject to changing conditions. It has been argued that many of the measurement stations are located in areas that have been encroached upon by urbanization. Since urban areas tend to be warmer than rural areas in otherwise similar locations, this will lead to an increase in measured temperature. Tests using spatial econometrics have shown that indexes of economic development are correlated with the temperature increases even after GHG accumulation has been taken into account. This evidence supports the urbanization argument.

The next criticism requires the introduction of some definitions from econometrics. Time series data are said to be stationary when the mean, variance, and other distributional characteristics do not change. The time series for a variable that is growing over time cannot be stationary because the mean is increasing. If the first difference (the value at time \( t \) minus that at \( t - 1 \)) of a nonstationary variable is stationary, then the
variable is said to be integrated of order 1. It is integrated of order 2 if the first difference is also nonstationary but the second difference is stationary. Because of the different growth characteristics, there cannot be a long-term relationship between two variables with different orders of integration. Econometric analysis has shown that temperature is integrated of order 1 but that the accumulation of GHGs is integrated of order 2. This prevents any stable relationship from being established between the levels of the two variables. If there is any causal effect from GHG accumulation to temperature, these orders of integration imply instead that the temperature will fall if the rate of increase of GHG accumulation slows.

The major controversy has been around the hockey stick in figure 26.4. To understand the issues, it is important to know how it was constructed. The hockey stick is actually composed of two separate parts. The early part of the data set is reconstructed from observations of natural phenomena. The later part of the data is taken from measurements of temperatures. The issues raised about the latter have already been discussed, so the focus is now placed on the process of temperature reconstruction.

The process of reconstruction employs numerous separate data series on natural phenomena that are believed to be related to temperature. For example, if trees grow faster when the temperature is higher then the width of tree rings provides a record of temperatures. Similarly ice cores, which reveal the depth of annual snow fall, indicate temperatures if it snows less when temperature is higher. The data series are then subjected to a principal components analysis in which a common “factor” is extracted from different data series that provide a common explanation for the variation in the data. This factor is interpreted as temperature. The factor extracted is calibrated to temperature over a period for which measurements exist, and then the remainder of the temperature series is inferred from the calibration.

This method has the obvious problem that it is not based on direct measurement so incorrect inference may take place. At the very least, it requires that the extracted factor is related to temperature (it need not be) and that the physical processes do have a monotonic relationship with temperature. Since the growth of trees is stunted both by low temperatures and by very high temperatures the latter is particularly doubtful. There are other problems related to the actual implementation of the method. This aspect has been the subject of considerable public dispute since the exact data and programs used to produce the figure have not been made public. Whether the hockey stick is right or wrong, it is clearly a poor practice in science to not make data available to allow thorough academic scrutiny and permit reproducibility to be checked. What little has been revealed appears to contain errors in method and in the actual data that undermine the validity of the exercise. A re-estimation that corrects these errors has
produced a very different figure. Furthermore the method used will produce a hockey-stick shaped figure out of any data series that increases in the final few years. It is probably fair to say that there is a significant body of evidence against the hockey stick.

### 26.3.3 Predictions

The data show what has happened to the level of GHGs and, subject to the criticisms noted, to temperatures. This is helpful, but what is necessary for policy-making is information on the future path of temperatures and the effect that this will have on climate. This is the role of climate prediction models. The results from such models will now be reviewed. This also requires the introduction of two further important concepts.

The standard analysis for judging the effect of GHG accumulation is to predict the effect of a doubling of the level of CO$_2$e from the pre-industrial level, that is, from 278 ppm to 550 ppm CO$_2$e. The increase in temperature that is predicted is called climate sensitivity. The degree of climate sensitivity is central to forecasting the effects of GHGs. If climate sensitivity is very low, then the accumulation will have little effect on climate and little need for policy. Conversely, if climate sensitivity is high, then GHGs will pose a serious threat and immediate policy response is critical.

The estimates of climate sensitivity from a number of different studies are shown in figure 26.6. The curves are the probability density functions (pdfs) for the predicted distribution of climate sensitivity. Hence the area under the curve up to a given temperature is the probability of an increase equal to, or less than, that temperature. The bars show the 5 to 95 percent confidence intervals for each estimate, and the dots are the median estimates. It can be seen that the median estimates of climate sensitivity typically fall in the range of 2°C to 4°C. The presentation of the estimates using pdfs is a reminder that there is uncertainty about the value of climate sensitivity and that these are estimates from models. The predictions are only as good as the models that generated them.

The second concept that is used to describe the consequences of GHG accumulation is that of stabilization. Stabilization is achieved when the level of GHGs in the atmosphere is constant. The stabilization level of GHGs if the world continues on its current course is unknown; this is apparent from the graphs in figure 26.2, which show no sign of reaching a peak. Equally, it is not possible to predict the extent to which policies to reduce the emission of GHGs will lower the stabilization level. This is a source of uncertainty for policy-making.
Figure 26.6
Estimates of climate sensitivity

Despite these limitations on knowledge, the concept of stabilization can still provide useful information. Table 26.1 reports estimates of the predicted level of warming for a range of stabilization levels. It is clear that as the stabilization level rises so does the increase in temperature. Furthermore the table shows that to achieve a stabilization level below 550 ppm, it is necessary to make significant reductions in emissions.

The temperature that will be achieved at any stabilization level is uncertain. One way to express this uncertainty is to consider the probability of exceeding a given increase in temperature. Table 26.2 presents a set of such estimates. It has to be stressed that these probabilities are generated from a climate modeling exercise, so they represent model predictions rather than facts. The estimates describe the probability of exceeding a particular global temperature increases at different concentration levels of CO2e.

The temperature figures are expressed as global averages. This obscures the prediction that climate change will result in very different local effects. Predictions for some of these local effects are shown in figure 26.7. The figure shows the mean changes in surface air temperature (°C, left), precipitation (mm per day, middle) and sea level pressure (hPa, right) for northern winter (top) and northern summer (bottom). The changes are given for the IPCC A1B scenario (very rapid economic growth, a
### Table 26.1
Stabilization and predicted temperature increases above the pre-industrial level

<table>
<thead>
<tr>
<th>CO₂-equivalent concentration at stabilization (ppm; 2005 = 375 ppm)</th>
<th>Peak years for CO₂ emissions</th>
<th>Global average CO₂ emissions (% of 2000 emissions)</th>
<th>Change in temperature increase (°C, using best-estimate of climate sensitivity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>445–490</td>
<td>2000–2015</td>
<td>−85 to −50</td>
<td>2.0–2.4</td>
</tr>
<tr>
<td>490–535</td>
<td>2000–2020</td>
<td>−60 to −30</td>
<td>2.4–2.8</td>
</tr>
<tr>
<td>535–590</td>
<td>2010–2030</td>
<td>−30 to 5</td>
<td>2.8–3.2</td>
</tr>
<tr>
<td>590–710</td>
<td>2020–2060</td>
<td>10 to 60</td>
<td>3.2–4.0</td>
</tr>
<tr>
<td>710–855</td>
<td>2050–2080</td>
<td>25 to 85</td>
<td>4.0–4.9</td>
</tr>
<tr>
<td>855–1130</td>
<td>2060–2090</td>
<td>90 to 140</td>
<td>4.9–6.1</td>
</tr>
</tbody>
</table>


### Table 26.2
Probabilities of exceeding temperature increases, relative to 1850, at various stabilization levels of CO₂e

<table>
<thead>
<tr>
<th>Stabilization level (ppm CO₂e)</th>
<th>Temperature increase (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2°C</td>
</tr>
<tr>
<td>450</td>
<td>78</td>
</tr>
<tr>
<td>500</td>
<td>96</td>
</tr>
<tr>
<td>550</td>
<td>99</td>
</tr>
<tr>
<td>650</td>
<td>100</td>
</tr>
<tr>
<td>750</td>
<td>100</td>
</tr>
</tbody>
</table>


Midcentury peak in global population followed by decline, and the rapid introduction of new and more efficient technologies) for the period 2080 to 2099 relative to 1980 to 1999. The stippling denotes areas where the predictions have the greatest uncertainty.

The figure shows that higher latitudes and continental regions are predicted to experience greater temperature increases. The greatest increase is in the Arctic, which will lead to a melting of the polar ice cap and a rise in sea levels. Rainfall patterns are also predicted to change with the level of rainfall decreasing in the bands north and south of the equator. These changes could lead to significant water shortages across much of the globe, including some important agricultural areas since much of Africa, southern Europe, the United States, and Australia will see reduced rainfall.
These figures summarize the predictions of the effects of GHG accumulation on climate. It should be stressed again that they are predictions from models of climate. The predictions could be modified if the change in climate causes feedback effects, such as damage to carbon-absorbing organisms in the oceans and the release of methane from warming wetlands and permafrost, that add to the accumulation of GHGs. There is considerable uncertainty, and the policy analysis has to work with this uncertainty.

26.4 Integrating Economics and Climate

The emission of GHGs is at least partly a result of economic activity. This provides a causal link from the economy to climate. Climate change from GHG accumulation affects the level of economic activity through the changing climate patterns. This gives a causal link from climate to the economy. Both of these processes operate over time—quite possibly very long periods of time. To model this interaction, it is necessary to employ an intertemporal model with feedbacks from economic activity to climate, and from climate to economic activity. The purpose of this section is to review a leading example of such a model. Before proceeding to the model, a brief summary will be given of some of the predicted effects of climate change. By contrasting the predictions
with the model, it is possible to assess the extent to which the modeling captures the damage scenarios.

### 26.4.1 Sources of Damage

There are many potential sources of damage from climate change. Three of the most important are the effects on water supply, food production, and energy use. In addition there are the effects of extreme weather events. The consequences will differ across countries: latitude and current climate are important determinants of the effects. It should be stressed once again that uncertainty surrounds all current predictions.

The water supply at any location is affected by altered rainfall and runoff from surrounding land. It is predicted that the supply may rise in higher latitudes but fall in lower latitudes due to lower rainfall and less snow meltwater. More precisely, summer water availability in southern Europe could fall by 20 to 30 percent, and the West Coast of the United States would have erratic water supply as the snow pack decreases by 25 to 40 percent, if average temperatures rise by $2^\circ C$. Australia will become drier as storms track farther south. Agriculture is directly affected by the supply of water and by temperature change. At higher latitudes small temperature increases may increase output due to more favorable conditions until the tolerance thresholds of the crops are reached. At lower latitudes (southern Europe, western United States, western Australia) water shortages will lead to declines in crop yields. There are two offsetting effects of temperature increases on energy use: less heating in winter but more air conditioning in summer. Climate change can also affect energy production, since warmer water reduces the cooling capacity available for power production and reduced rainfall affects production of hydroelectric power.

The location of developing countries in the lower latitudes and their dependence on agriculture means that they will face the most serious costs. Many developing economies, such as Bangladesh and the Maldives, also have low-lying coastal areas that are vulnerable to flooding if melting ice caps cause sea levels to rise. Developed economies will suffer less because of their lower dependence on agriculture and location in higher latitudes. Furthermore the developed economies have the resources to ensure greater adaptive capacity. Small increases in temperature may actually be beneficial for some developed economies if it increases agricultural output. However, it is predicted that an increase in temperature of $4^\circ C$ to $5^\circ C$ will have severe effects for all countries.

The effects of climate change will be unevenly distributed in the United States due to its geographical diversity. There may be short-term benefits in the north with better conditions for agriculture and lower energy requirements for heating. Extensive
damage is predicted in the south through reduced snowfall and shorter winters. This will affect the water supply of the Pacific coast, California, and the Mississippi basin. The temperature changes will cause farm production to shift northward, with total output falling once the temperature increase exceeds 3°C. In addition the south will face increased energy use for cooling purposes.

The prediction for the United Kingdom is that infrastructural damage from flooding and storms will increase. Water availability will be constrained with more frequent droughts. Milder winters will reduce energy use. Agricultural productivity may increase, but this is dependent on the changes in water supply. These predictions are all conditional on changes to the Gulf Stream, which is a flow of warm water up the west coast of the United Kingdom. This flow is responsible for the United Kingdom having much warmer winters than the Canadian east coast at the same latitude. If the Gulf Stream were to slacken or cease, then the average winter temperature in the United Kingdom would fall significantly.

Australia is vulnerable to the impact of rising sea temperatures. More storms and damage will occur in the northwest, whereas the east coast will suffer from longer droughts and declining rainfall. The water supply available for cities will fall. Drier and hotter summers will damage the rain forest, and sea temperatures will damage coral reefs.

In addition to these effects there is predicted to be an increase in extreme weather events and consequent damage to property and loss of life. Most areas of the world will be at threat from these events. The consequence of the range of changes outlined could be political and social unrest driven by hunger and thirst resulting from reductions in agricultural output and water supply. An ensuing consequence could be significant displacement and migration of populations.

26.4.2 Modeling Economic Damage

The preceding comments have described some of what may happen. To make predictions of economic damage, the purported changes need to be incorporated within an economic model.

An Integrated Assessment Model (IAM) combines economic activity and climate change in a single framework with feedback between the two: the economy affects climate, and climate affects the economy. An IAM is a set of equations drawn from economic analysis and climate modeling. The key equations for an IAM are those that link economy and climate. The Dynamic Integrated model of Climate and Economy
Chapter 26: Economics of Climate Change

(DICE) produced by William Nordhaus at Yale is now used to illustrate a typical IAM. We focus on this model because the equations are publicly and openly available.

The economic side of the DICE model adopts a standard form of growth through capital accumulation (see chapter 24). It assumes that industrial production generates emissions that affect temperature through accumulation of GHGs. The temperature then determines the damage caused by climate change. The model incorporates twelve regions: the United States, European Union, Other High Income, Russia, Eastern Europe, non-Russian former Soviet Union, Japan, China, India, Middle East, Sub-Saharan Africa, Latin America, and Other Asia. The countries that are modeled represent 97 percent of emissions and 94 percent of world output.

The economic side of the model describes production and welfare. The level of industrial output is determined by a Cobb–Douglas production function

\[ Y(t) = K(t)^\gamma L(t)^{1-\gamma}, \]  

(26.2)

where \( K(t) \) is the capital stock at time \( t \) and \( L(t) \) is the quantity of labor used. It is assumed that all consumers supply one unit of labor, so \( L(t) \) is also the population size at time \( t \). Some of the output produced at \( t \) is used in the abatement of emission and some is lost to damage through climate change. The net level of output after damage (\( \Omega(t) \)) and abatement cost (\( \Lambda(t) \)) are taken into account is

\[ Q(t) = \Omega(t)[1 - \Lambda(t)]Y(t). \]  

(26.3)

Per capita consumption, \( c(t) \), and gross investment, \( I(t) \), must satisfy

\[ Q(t) = L(t)c(t) + I(t). \]  

(26.4)

The level of investment determines the evolution of the capital stock

\[ K(t) = I(t) + [1 - \delta_K]K(t - 1). \]  

(26.5)

The time path of consumption (and hence of investment) is chosen to maximize the social welfare function

\[ W = \sum_{t=1}^{T_{\text{max}}} \frac{1}{[1 + \rho]^t} \left[ L(t) \frac{c(t)^{1-\alpha}}{1-\alpha} \right]. \]  

(26.6)

The pure rate of social time preference, \( \rho \), provides the welfare weights on the utilities of different generations. With \( \rho > 0 \), this form of social welfare function implies a discount on the economic well-being of future generations. Here \( \alpha \) is the elasticity with respect to consumption of the marginal utility of consumption. The parameters
\( \alpha \) and \( \rho \) are ultimately calibrated to match market data. This is explored further in section 26.5.

The next set of equations describe the climate side of the model. The first three represent the carbon cycle and the transfer of carbon between three different locations (or "sinks"). Carbon can be trapped in the atmosphere (AT), the upper ocean (UP), and the lower ocean (LO). Carbon is transferred slowly from the upper ocean to the lower ocean, with the lower ocean modeled as a large sink that can store vast quantities of carbon. The level of carbon in the atmosphere reservoir (AT) is

\[
M_{AT}(t) = E(t) + \phi_{11}M_{AT}(t - 1) + \phi_{21}M_{UP}(t - 1).
\]  
(26.7)

In the upper oceans (UP),

\[
M_{UP}(t) = \phi_{12}M_{AT}(t - 1) + \phi_{22}M_{UP}(t - 1) + \phi_{32}M_{LO}(t - 1),
\]  
(26.8)

and in the lower oceans (LO),

\[
M_{LO}(t) = \phi_{23}M_{UP}(t - 1) + \phi_{33}M_{LO}(t - 1).
\]  
(26.9)

These three first-order difference equations form a system that is driven by the quantity of emissions, \( E(t) \). These accumulation equations are next linked to the degree of radiative forcing that determines warming. The climate model for temperatures in the atmosphere and in the lower oceans is a two-equation system. This system involves the past values of temperature and an outside force, \( F(t) \), that drives change in temperature. The outside force is determined by the radiative forcing equation

\[
F(t) = \eta \log_2 \left( \frac{M_{AT}(t)}{M_{AT}(1750)} \right) + F_{EX}(t),
\]  
(26.10)

where the first term is the consequence of carbon accumulation and \( F_{EX}(t) \) is exogenous forcing. The temperature system is

\[
T_{AT}(t) = T_{AT}(t - 1) + \xi_1 \left[ F(t) - \xi_2 T_{AT}(t - 1) \right. \\
- \xi_3 (T_{AT}(t - 1) - T_{LO}(t - 1)) \\
(26.11)
\]

\[
T_{LO}(t) = T_{LO}(t - 1) + \xi_4 [T_{AT}(t - 1) - T_{LO}(t - 1)].
\]  
(26.12)

If \( F(t) \) is constant, these equations have a steady state with constant temperatures. Only if \( F(t) \) increases over time will the steady-state temperature levels increase.

The final equations link the two systems. Production generates industrial emissions of amount

\[
E_{Ind}(t) = \sigma(t) [1 - \mu(t)] Y(t).
\]  
(26.13)
where $\sigma(t)$ is the ratio of uncontrolled emissions to output and $\mu(t)$ is the emissions control rate (the abatement). Total emissions are sum of industrial and land

$$E(t) = E_{\text{Ind}}(t) + E_{\text{Land}}(t). \quad (26.14)$$

The abatement cost in (26.3) is related to the emissions reductions rate by

$$\Lambda(t) = \chi(\mu(t)). \quad (26.15)$$

The climate damage function $\Omega(t)$ in (26.3) is assumed to depend on global mean temperature change:

$$\Omega(t) = \frac{1}{1 + \psi_1 T_A(t) + \psi_2 T_A(t)^2}. \quad (26.16)$$

This equation is constructed from damage estimates.

It can be seen that (26.16) is a macro-level representation of damage. In contrast to the range of effects described in section 26.4.1, damage in the model is just a reduction in the net level of output. This is a consequence of using an aggregate economic model that has a single consumption good. Such models are adequate for addressing the fundamentals of growth theory but limit the level of sophistication in modeling climate damage. This is a criticism that applies not only to the DICE model—which represents the frontier of economic modeling in this area—but to economic models of growth in general. The need for developments to produce compelling models that can be used for more detailed economic forecasting is clear.

### 26.4.3 Effects of Policy

Economic forecasting is about the prediction of future levels of economic activity. The production of a forecast requires assumptions to be made on the values of exogenous variables and a model that determines the values of endogenous variables. A standard technique is to compare a baseline case—a particular assignment of values to exogenous variables—to alternative cases conducted under different assumption. The baseline case is often chosen to represent what would happen to the economy if there were no policy intervention. This is often called “business as usual.” A policy intervention is a change in the exogenous variables and the effect of policy is the change in the endogenous variables. The predictions made by the DICE model are now reviewed. These give an idea of the magnitude of the damage from climate change and the extent to which one particular form of policy intervention can reduce these.
Figure 26.8 shows the time path of CO₂ emissions for the baseline run of the DICE model compared to the predictions from two Intergovernmental Panel on Climate Change (IPCC) scenarios. The baseline run for DICE assumes that no climate-change policies are implemented. The A1B scenario involves very rapid economic growth, a midcentury peak in global population followed by decline, and the rapid introduction of new and more efficient technologies. The A1F1 scenario has low population growth. The next two sections will look in detail at the issues behind the choice of the key economic parameters: the pure rate of social time preference, \( \rho \), and the elasticity of the marginal utility of consumption, \( \alpha \). It is therefore worth noting for later comparison that the baseline run assumes the values \( \rho = 0.015 \) and \( \alpha = 2 \). When contrasted to the two IPCC scenarios, the DICE projection is the lowest of the three projections until the mid–twenty-first century but, unlike the A1B scenario, continues to rise. At the end of this century the DICE projection is between the two IPCC scenarios.

The policy studied by the DICE model is a tax levied on the output of carbon. Table 26.3 shows the benefit from the optimal policy and contrasts this to the alternatives of limiting the increase in CO₂ and limiting the increase in temperature. The key feature of all the policies is that the carbon tax rises over time, and quite dramatically so, when policy is designed to limit the increase in temperature to 2°C. The increase in the carbon tax over time is a consequence of emissions in the baseline run rising over time. These results suggest that limiting the temperature increase to 2°C may be costly. Compared
Table 26.3
Damage estimates

<table>
<thead>
<tr>
<th>Run</th>
<th>Present value of abatement costs plus damage (trillions of US$)</th>
<th>Carbon tax (2005 US$ per ton C)</th>
<th>Global temperature change from 1900 (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2010</td>
<td>2100</td>
<td>2100</td>
</tr>
<tr>
<td>Optimal</td>
<td>19.52</td>
<td>33.8</td>
<td>202.4</td>
</tr>
<tr>
<td>Limit to $2 \times CO_2$</td>
<td>19.92</td>
<td>39.6</td>
<td>445.5</td>
</tr>
<tr>
<td>Limit to 2 degrees</td>
<td>24.39</td>
<td>60.2</td>
<td>863.4</td>
</tr>
<tr>
<td>Limit to 3 degree</td>
<td>19.57</td>
<td>37.9</td>
<td>256.7</td>
</tr>
</tbody>
</table>

To the optimal policy, the abatement cost is high. The increase in temperature under the optimal policy is $2.76^\circ C$, which shows that abatement costs rise quickly for small reductions in temperature increase.

26.5 Competing Generations

Climate models predict that the current accumulation of GHGs will have consequences that could last for hundreds of years into the future. At the very least any policy response has to account for effects over the next two centuries. This section addresses the implications of the very long run aspect of climate change. There are costs today if we wish to mitigate climate change and there are potential costs for future generations if we do not. The problem in the policy design area is how these two are balanced.

The standard approach to policy evaluation is to apply cost–benefit analysis. Let the benefits of an emissions reduction policy at time $t$ be $B(t)$ and the costs of the policy be $C(t)$. The basic rule of cost–benefit analysis is that the policy should be adopted if the associated net present value (NPV) is positive, where

$$NPV = \sum_{t=0}^{\infty} \frac{[1 + d]^{-t}}{[B(t) - C(t)]}.$$  (26.17)

Chapter 25 discussed the issues involved in calculating the benefits and the costs. Those issues are now assumed to be resolved, and we focus instead on the discount rate, $d$. The effect of discounting when applied to long-run decisions is shown in figure 26.9. This displays the present discounted value ($pdv$) of receiving $1,000 in the future for four different discount rates. For example, $1,000 received in 20 years is worth...
$500.66 today when the discount factor is 3.5 percent (the UK government discount rate). Expressed differently, the society would be willing to pay $500.66 today to avoid the loss of $1,000 in 20 years. If the loss occurs in 100 years, then the willingness to pay is $31.46, and if in 200 years, $0.99. This last number summarizes the point of these calculations: for a reasonable discount factor the willingness to pay today for losses that occur in the distant future is close to zero. Applying cost–benefit techniques to climate policy, for this magnitude of discount factor, will require future generations to suffer very major losses (relative to current costs) before a policy is approved by the NPV criterion.

This raises the question of what discount rate should be used in climate policy. For the purpose of the discussion that follows, two different discount rates need to be identified:

1. Discount rate in financial markets, $r$

This discount rate can also be called the real return on capital, the real interest rate, the real return, or the discount rate on goods. It is observable in the marketplace as the real interest rate offered on risk-free securities. The closest assets to being risk-free are the treasury bills (short-term government borrowing) of the major economies. Many applications in finance use US Treasury bills as the practical equivalent of a risk-free asset. The annual real return on these bills has averaged approximately 3 percent. Assets that have greater degrees of risk have to offer higher expected returns, so corporate stock have an average real return of 6.6 percent, and human capital an average real return in the range 6 to 20 percent. As we observed in chapter 25, this discount rate is determined by the equilibrium of supply and demand for assets in financial markets.
2. Relative weight of different generations in welfare, $\rho$

The relative weight assigned to different generations in the social welfare function is the pure rate of social time preference or time discount rate of the social planner. A value of zero for this discount rate means that the utility of different generations is treated equally. A positive value means that the utility of future generations is discounted relative to that of the current generation.

If we treat (26.17) as a welfare calculation, then $\rho$ should be used as the discount rate. This raises the question of what should be the value of $\rho$. This question has led to considerable debate in the literature on climate policy. Partly, this debate has been about whether $\rho$ can be freely chosen in a policy analysis or whether it has to be consistent with the market interest rate, $r$. To a lesser extent there has also been a debate about the correct value of $\rho$ if it can be freely chosen. The chosen value has significant implications for the trade-off between present costs and future benefits.

An understanding of how preferences are related to observable variables can be obtained by using the Ramsey growth model. In this model a single individual has to choose consumption and saving to maximize lifetime utility. The basic trade-off in the choice is that increasing consumption in one period raises utility but reduces saving. This leads to a lower capital stock in the next period, which reduces output and therefore consumption and utility. The time path of consumption is chosen optimally to accommodate this trade-off.

We now employ the derivation used in section 24.3.2 of chapter 24 to characterize the optimal consumption path. The intertemporal preferences of the social planner are described by the utility function

$$W = \sum_{t=1}^{\infty} \frac{C_t^{1-\gamma}}{(1 + \rho)^t},$$  \hspace{1cm} (26.18)

where $\rho$ is the inter-generational discount rate. The social planner chooses the path $\{C_t\}$ over time to maximize utility. The standard condition for intertemporal choice must hold for the optimization, so the ratio of the marginal utilities of consuming at $t$ and at $t + 1$ must equal the gross interest rate. Hence

$$\frac{\partial W}{\partial C_t} = \frac{1 + \rho}{C_{t+1}} = 1 + r_{t+1}. \hspace{1cm} (26.19)$$

By solving for $\frac{C_{t+1}}{C_t}$ and then subtracting $\frac{C_t}{C_t}$ from both sides of the resulting equation, this optimality condition can be rewritten in terms of the growth rate of consumption.
Thus
\[
\frac{C_{t+1} - C_t}{C_t} = \left[ \frac{1 + r_{t+1}}{1 + \rho} \right]^{1/\gamma} - 1. \tag{26.20}
\]

Now define the growth rate of consumption \( g_{t+1} \equiv \frac{C_{t+1} - C_t}{C_t} \). This gives
\[
1 + g_{t+1} = \left[ \frac{1 + r_{t+1}}{1 + \rho} \right]^{1/\gamma}, \tag{26.21}
\]

which is the relation between the chosen growth rate and the interest rate. For the usual order of magnitude of these variables, a reasonable approximation is given by
\[
r_{t+1} = \rho + \gamma g_{t+1}. \tag{26.22}
\]

This result is the basic identity that links the equilibrium return on capital to the intertemporal discount rate on the optimal path. It is usually called the Ramsey rule for intertemporal consumption, since it describes the condition that must be met by the optimal time path of consumption.

The two parameters in the social welfare function can be given a useful interpretation. The discount rate \( \rho \) can be interpreted as aversion to inequality across generations, and the elasticity \( \gamma \) as aversion to inequality within a generation (see the construction of the Atkinson index in chapter 14). The implication of (26.22) is that the observed market interest rate can only be used to infer the aversion to inequality across generations if the aversion to inequality within generations and the optimal growth rates are known. An alternative expression of the same statement is that it is not possible to choose \( \rho \) and \( \gamma \) independently without reaching a contradiction to the observed data: if we observe \( r \) and \( g \), then either \( \rho \) or \( \gamma \) can be freely chosen, but not both.

It should be noted that the arguments just made were based on the assumption that the social preferences in (26.18) align with the underlying individual preferences. If they do not, the market and the social planner will choose different growth paths. The observed market return on capital will be different from the return along the optimal growth path and so cannot be used to relate \( \rho \) and \( \gamma \). This is important to keep in mind before taking the outcome of (26.22) literally.

Alternative analyses of climate policy have taken different approaches to selecting values for \( \rho \) and \( \gamma \). The approach followed by Nordhaus, when implementing the DICE model, was to take observed values in the market and deduce the parameters. As already noted, this assumes that the market outcome is optimal relative to social preferences. The approach of the Stern Review (2007) was to adopt a welfarist perspective.
and to select values that the planner sees as ethically correct. This has the drawback that the chosen social preferences may be inconsistent with individual preferences. Furthermore there may be no consensus as to what the values should be because it is perfectly acceptable for different people to have different ethical views. In the sense of Arrow’s impossibility theorem of chapter 13 the welfare function must be dictatorial. The implications of these observations are now explored in some detail.

The Stern Review applies the ethical philosophy of utilitarianism across generations and so chooses a discount rate that is very close to zero. An appeal is made to underlying uncertainty—in the example of the Review that the world may end by being struck by an asteroid—to adopt a small but positive value of $\rho$. The more pragmatic reason for not using a value that is exactly zero is to obtain a finite value for social welfare. The Review uses a discounted utilitarian social welfare function of the form

$$W = \sum_{t=0}^{\infty} \frac{1}{(1 + \rho)^t} U(t)$$

(formally, the Review uses the continuous time analogue of equation 26.23). Since $U(t) > 0$ for all $t$, this summation will be infinite for $\rho = 0$. Hence the utilitarian philosophy of $\rho = 0$ cannot be applied unmodified; otherwise, the level of social welfare will be infinite and comparison of policies will not be meaningful (every policy will lead to infinitely large welfare). The appeal to uncertainty is a device used to make the philosophy of no intergenerational discounting consistent with a finite value of social welfare over an infinite horizon.

The Stern Review assumes that $\gamma = 1$ (so that utility is logarithmic) and $g = 0.013$ (1.3 percent). The implication of this choice of growth rate is that if per capita income is $10,000$ today, it will increase to $135,000$ in two centuries. The value of the intergenerational discount rate is taken as $\rho = 0.001$ (0.1 percent). The choices of $\gamma$ and $\rho$ can be put into perspective by observing that

$$\ln(10,000) - \ln(9,389) = \frac{1}{(1.001)^{200}} \ln(135,000)$$

$$- \frac{1}{(1.001)^{200}} \ln(125,000),$$

meaning, the society will give up $611$ of consumption today (reducing consumption from $10,000$ to $9,389$) to avoid a $10,000$ reduction in consumption (from $135,000$ to $125,000$) that would take place in 200 years. It is hard to believe that most, if any, people alive today would agree that this is an acceptable trade-off. But an even more extreme trade-off for a long-run future reduction in consumption may be considered:
per capita consumption of $10,000 today would be reduced to effectively $23 for one year to prevent a reduction of consumption from $135,000 to $134,000 starting in 200 years and continuing onward. That is,

\begin{equation}
\sum_{t=200}^{\infty} \frac{1}{(1.001)^t} \ln(135,000) - \sum_{t=200}^{\infty} \frac{1}{(1.001)^t} \ln(134,000) = \ln(10,000) - \ln(22.565).
\end{equation}

This is obviously not an acceptable trade-off. The assumed values do not correspond to any existing experimental or market setting. This is why the choices of the Stern Review have to be treated as explicitly ethical judgments that do not represent any values ever observed a market outcomes.

Using the Stern Review values for \( \rho \) and \( g \) in (26.22), we have a return on capital of \( r = 0.014 \) (1.4 percent). This rate is below the market rate, making the values inconsistent with market equilibrium. Interestingly, if a return on capital of \( r = 0.04 \) is assumed, then \( g = 0.013 \) and \( \rho = 0 \) imply \( \gamma = 3 \), which is a much higher rate of within generation inequality aversion than the Review assumes. Alternatively, a value of \( \gamma = 1 \) implies \( \rho = 0.027 \), which is close to the discount rate recommended in the UK government cost–benefit analysis.

In contrast, the baseline for the DICE model assumes \( \gamma = 2 \), \( g = 0.02 \), and \( r = 0.055 \). The motivation for the value of \( r \) is that it is close to the long-term return on government bonds. The value of \( g \) is consistent with what most major economies have achieved over the past thirty years. Substituting these into (26.22) gives a growth rate of 0.015. Moving from \( \rho = 0.001 \) to \( \rho = 0.015 \) causes a major shift in the valuation of future consumption relative to present consumption. This is one of the reasons why the optimal policy of the DICE model is a carbon tax that increases over time: this shifts the cost of policy more into the future.

The discussion so far has not properly modeled or accommodated uncertainty. One of the features of climate policy that we have continually stressed is the degree of uncertainty involved. The consequences of uncertainty must be considered because they are significant for the value of the discount rate.

Uncertainty can be introduced by returning to the determination of the optimal path of consumption and observing that uncertainty about future levels of consumption makes the marginal social welfare of future consumption a random variable. It is therefore the expected value of the marginal social welfare of future consumption that enters into the optimality condition (26.19). If the interest rate is assumed to be constant, (26.19) can be rewritten as
\[ \frac{\partial W}{\partial C_0} = (1 + r)' . \] (26.26)

Social welfare is the discounted value of utility so \( \frac{\partial W}{\partial C} = \frac{1}{(1 + \delta)^t} \frac{\partial U}{\partial C} \). This allows (26.26) to be expressed as

\[ 1 = \frac{1}{(1 + \delta)^t} \frac{E[\partial U/\partial C_t]}{\partial U/\partial C_0} (1 + r)' . \] (26.27)

Taking the logarithm of both sides, rearranging, and using the approximations \( \ln(1 + r) \approx r \) and \( \ln(1 + \delta) \approx \delta \) gives

\[ r = \delta - \frac{1}{t} \ln \left( \frac{E[\partial U/\partial C_t]}{\partial U/\partial C_0} \right) . \] (26.28)

Now suppose that the growth rate of consumption is random with

\[ C_{t+1} = C_t e^{x_t} , \] (26.29)

where \( x_t \) is drawn from a normal distribution with mean \( \mu \) and variance \( \sigma^2 \). The expected growth rate of consumption is

\[ g = \ln \left( \frac{E[C_{t+1}]}{C_t} \right) = \ln \left( E[e^{x_t}] \right) = \ln \left( e^{\mu + 0.5\sigma^2} \right) = \mu + 0.5\sigma^2 . \] (26.30)

For the utility function \( U(C_t) = \frac{C_t^{1-\gamma}}{1-\gamma} \) and the random growth rate of consumption,

\[ \ln \left( \frac{E[\partial U/\partial C_t]}{\partial U/\partial C_0} \right) = \ln \left( E[e^{-\gamma x_t}] \right) = \gamma \ln(\mu - 0.5\sigma^2) . \] (26.31)

By combining (26.28), (26.30), and (26.31), we obtain the value of the discount rate incorporating the uncertainty:

\[ r = \rho + \gamma g - 0.5\gamma (\gamma + 1) \sigma^2 . \] (26.32)

Here \( \rho \) represents impatience, \( \gamma g \) the wealth effect arising from the growth in wealth between generations, and \( -0.5\gamma (\gamma + 1) \sigma^2 \) the precautionary effect due to the uncertain future growth rate. Clearly, uncertainty over the growth rate of consumption reduces the discount rate that should be applied in the CBA evaluation, and higher uncertainty (higher \( \sigma^2 \)) reduces it still more. It should be noted that it is possible for \( g \) to be negative as consumption falls over time. This could be a relevant outcome if
climate change has dramatic and damaging effects or the cost of combating change is very high. The result in (26.32) shows that for a given value of $r$, a negative value of $g$ reduces $\rho$.

These observations reveal the issues involved in selecting the basic parameters of an IAM. The standard discounting method implies that long-run climate change should not be confronted if there is a significant present cost in doing so. To get around this fact, it is necessary to use, and justify, a much lower discount rate. One way to do so is to treat the choice of discount rate as an ethical issue. However, though acceptable, it is hard to reconcile market equilibrium with the implications of ethical principles. Either approach to discounting can be justified, since there is no certain way to settle these debates. Nevertheless, it is possible to combine the two approaches in a more general framework, as we show next.

### 26.6 Ecological Discounting

Nordhaus (2007) and Stern (2007) offer contrasting approaches to the determination of a discount rate. There has been considerable debate in the literature about the relative merits of the two positions. While both appear to be based on very distinct sets of arguments, it is still possible to achieve a reconciliation by accommodating both within an extended framework. In doing so, we can gain an important insight into the that allows a different perspective to be formed as to their relative merits.

The extended framework treats environmental quality as a distinct good that enters the utility function. It then becomes possible to isolate a discount rate that can evaluate changes in environmental quality from the discount rate for produced goods. The value of the discount rate for environmental quality is based on society’s preferences for environmental quality and for produced goods. Nordhaus and Stern apply as special cases of this general result.

Let $C_t$ denote the consumption of produced goods at time $t$, and $Q_t$ be environmental quality in the same time period. Generation $t$, who are alive during period $t$ only, have preferences described by the constant elasticity of substitution utility function

$$u(C_t, Q_t) = \left[ C_t^{(\mu-1)/\mu} + Q_t^{(\mu-1)/\mu} \right]^{\mu/(\mu-1)}.$$  

(26.33)

Here $\mu$ is the elasticity of substitution between goods and environment. This parameter represents a measure of how easy it is to compensate for decreased environmental quality by increasing the consumption of produced goods. When $\mu \to \infty$, there is
perfect substitutability, and when \( \mu \to 0 \), there is no substitutability. The elasticity has a key role in the analysis. The social planner cares about inequality across generations values and applies a discount to future welfare. The social welfare function is

\[
W = \sum_{t=0}^{\infty} \frac{1}{[1 + \rho]^t} \frac{u(C_t, Q_t)^{1-\gamma}}{1 - \gamma}.
\]  

(26.34)

Within this framework two different discount rates can be applied. Let \( U(C_t, Q_t) \equiv \frac{u(C_t, Q_t)^{1-\gamma}}{1 - \gamma} \). The discount rate for produced goods between periods \( t \) and \( t+1 \) is defined by the value of \( r_t \), which satisfies

\[
\frac{1}{1 + r_t} = \frac{1}{1 + \rho} \frac{U_C(C_{t+1}, Q_{t+1})}{U_C(C_t, Q_t)}. 
\]  

(26.35)

The value of \( r_t \) measures the rate at which \( C_t \) has to be increased to compensate for a reductions in \( C_{t+1} \). Similarly the ecological discount rate, \( \beta_t \), between periods \( t \) and \( t+1 \) is defined by

\[
\frac{1}{1 + \beta_t} = \frac{1}{1 + \rho} \frac{U_Q(C_{t+1}, Q_{t+1})}{U_Q(C_t, Q_t)}. 
\]  

(26.36)

where \( \beta_t \) determines the rate at which environmental quality has to be increased in \( t \) to compensate for a reduction of quality in \( t + 1 \).

Two simplifying assumptions can now made. First, the consumption of produced goods grows at the constant rate \( g \), so that

\[
C_{t+1} = [1 + g] C_t. 
\]  

(26.37)

Second, the level of environment is fixed at \( \bar{Q} \). Damage to the environment can be modeled as an exogenous reduction in \( Q \). Under these assumptions

\[
\frac{1}{1 + r_t} = \frac{1}{1 + \rho} \left[ \frac{C_{t+1}}{C_t} \right]^{-1/\mu} \left[ \frac{C_t^{(\mu-1)/\mu} + \bar{Q}^{(\mu-1)/\mu}}{C_t^{(\mu-1)/\mu} + \bar{Q}^{(\mu-1)/\mu}} \right]^{(1-\mu\gamma)/(\mu-1)} . 
\]  

(26.38)

Taking the log of both sides of this equation and using the approximation that \( \ln(1 + x) \approx x \) gives

\[
r_t = \rho + \gamma g + \frac{1 - \mu \gamma}{\mu - 1} \ln \left( \frac{1 + \eta_{t+1}}{1 + \eta_t} \right). 
\]  

(26.39)
or equivalently,

\[ r_t = \rho + \frac{1}{\mu} g + \frac{1 - \mu \gamma}{\mu - 1} \ln \left( \frac{1 + \eta_t^{-1}}{1 + \eta_{t+1}^{-1}} \right), \quad (26.40) \]

where

\[ \eta_t = \left[ \frac{C_t}{\bar{Q}} \right]^{(1-\gamma)/\gamma}. \quad (26.41) \]

The formula in (26.39) shows how the standard result on discount rates is affected by environmental quality entering the utility function; see (26.22). The additional term depends on the ratio of the consumption of produced goods to environmental quality. From (26.36) and (26.38) it follows that

\[ \beta_t = r_t - \frac{g}{\mu}. \quad (26.42) \]

This result emphasizes the effect of the growth of private consumption on the ecological discount rate: a higher rate of growth reduces the discount rate of environmental quality. Hence more attention must be given to maintaining the future environment when consumption grows faster. Climate policy can be distinguished from many other economic issues by the potentially long time period that is involved in valuing policy interventions. This motivates the analysis of the discount rates for produced goods and for environmental quality as \( t \) increases without limit. The degree of substitutability, \( \mu \), is critical in the determination of the discount factors in the long run.

First, consider \( \mu > 1 \). In this case the utility function can be written as

\[
\begin{align*}
 u(C_t, y_t) &= \left[ \frac{C_t^{(\mu-1)/\mu}}{\bar{Q}^{(\mu-1)/\mu}} + \frac{\bar{Q}^{(\mu-1)/\mu}}{C_t^{(\mu-1)/\mu}} \right]^{\mu/(\mu-1)} \\
 &= C_t \left[ 1 + \left( \frac{\bar{Q}}{C_t} \right)^{\gamma/(\mu-1)} \right]^{\mu/(\mu-1)}. \\
\end{align*}
\quad (26.43)
\]

Because \( C_t \) is growing at a constant rate and \( \frac{\mu-1}{\mu} > 0 \), it follows that \( \left( \frac{\bar{Q}}{C_t} \right)^{\gamma/(\mu-1)} \) tends to zero as \( t \) increases. Hence the level of utility will grow at the same rate as consumption of the produced good. Now use the expression \( r = \rho + \gamma g + \frac{1-\mu \gamma}{\mu-1} \ln \left( \frac{1+\rho_t}{1+\rho_{t+1}} \right) \). Since
\[ \rho_t = \left[ \frac{C_t}{Q} \right]^{(1-\mu)/\mu} \] and \[ \frac{1-\mu}{\mu} < 0, \] then \( \rho_t \) tends to 0 as \( t \) increases. Hence \( \ln \left( \frac{1+\rho_{t+1}}{1+\rho_t} \right) \) tends to \( \ln(1) = 0 \). This gives

\[ r = \rho + \gamma g. \] (26.44)

The value of \( \beta \) is obtained by substituting (26.44) into (26.42) to obtain

\[ \beta = \rho + g \left( \gamma - \frac{1}{\mu} \right). \] (26.45)

In this case the the produced good and environment are good substitutes, so increased production can keep raising utility even for a fixed environment. So the standard considerations still apply in obtaining the long-run discount rate for goods. The limit value of the ecological discount rate is significantly below the value of the standard discount rate (the difference is given by \( \frac{\gamma}{\mu} \)) but is above the intertemporal discount rate.

The consequences of this formula can be appreciated by working through an example. Let the loss of environmental benefit have loss of value \( L \) in every future period. The present discounted value at time 0, if the loss occurs from time 1 with discount rate \( d \), is

\[ pdv = \sum_{t=1}^{\infty} \frac{1}{(1+d)^t} L = \frac{1}{d} L. \]

With an interest rate of \( r = 0.03 \), the present discounted value from a standard cost–benefit calculation is

\[ pdv = \frac{1}{r} L = 33.3L. \]

To evaluate the ecological discounting formula, assume that \( \sigma = 1.5 \) and \( \rho = 0.001 \). Figure 26.10 plots the value of \( \frac{1}{\beta} \) as a function of \( \mu \) for \( \mu > 1 \). Note that the value of \( \frac{1}{\beta} \) is considerably higher than that of \( \frac{1}{r} \) for values of \( \mu \) close to 1 but tends to \( \frac{1}{r} \) as \( \mu \) increases.

Now consider the case where \( \mu < 1 \). Write the utility function as

\[ u(C_t, y_t) = \left[ C_t^{(\mu-1)/\mu} + \frac{\bar{Q}^{(\mu-1)/\mu}}{C_t} \right]^{\mu/(\mu-1)} \]

\[ = \bar{y} \left[ 1 + \left( \frac{\bar{Q}}{C_t} \right)^{1-\mu} \right]^{\mu/(\mu-1)}. \] (26.46)
Since $\frac{1-\mu}{\mu} > 0$, \[\bar{Q}\left(\frac{1-\mu}{\mu}\right)\] will tend to 0, and utility will tend to the constant value of $\bar{Q}$. So there is a limit on the attainable utility level. Take $r = \delta + \frac{g}{\mu} + \frac{1-\mu}{\mu-1}\ln\left(\frac{1+r_{i+1}^{-1}}{1+\rho_{i+1}}\right)$ and observe that since $\rho_{i+1}^{-1} = \left[\frac{C_t}{\bar{Q}}\right]^{(\mu-1)/\mu}$ and $\frac{\mu-1}{\mu} < 0$, then $\rho_i$ tends to 0 as $t$ increases.

Hence $\ln\left(\frac{1+r_{i+1}^{-1}}{1+\rho_{i+1}}\right)$ tends to $\ln(1) = 0$. This gives

$$r = \rho + \frac{g}{\mu}. \quad (26.47)$$

Because of the low degree of substitutability between the private good and the environmental good, the standard discount rate for the produced good is adjusted. Then $\beta$ is obtained by substitution as

$$\beta = \rho. \quad (26.48)$$

In this case the value of the ecological discount rate depends only on the intertemporal discount rate and does not depend on the rate of growth. The present discounted value of a permanent loss, $L$, in environmental quality is

$$pdv = \frac{1}{\rho}L.$$  

The value used in the example above ($\rho = 0.001$) gives $pdv = 1,000L$. The loss of environmental quality is very costly in this case.
Chapter 26: Economics of Climate Change

The main point of this discussion is that there does not need to be inconsistency between using the Ramsey model to deduce the discount rate applied to the produced goods and the intertemporal discount rate used to deduce the social welfare consequences of change in environmental quality. Because environmental quality has a limit, its increasingly high value will constrain the level of welfare that can be achieved at its expense. The analysis, however, does not define what value should be used for \( \rho \), so the issues of debate in the previous section are still relevant but enlightened somewhat.

26.7 Climate Policy in Practice

Climate policy is primarily concerned with the reduction of harmful emissions. Reductions can be achieved either through price controls or through direct controls of the quantity of emissions. Price controls can take the form of carbon taxes or other environmental taxes. Quantity controls can range from prohibitions of some emissions or activities to requiring emissions permits. This section discusses the practical implementation of these policy tools.

26.7.1 Pricing and Policies

In recent years policies to reduce emissions have proliferated. Often the policies are inadequately coordinated, and therefore emissions from different sectors are priced very differently and inconsistently. The situation in the United Kingdom is used as an example to illustrate this point, and then we consider what the efficient outcome should be.

Large electricity producers in the United Kingdom need to hold emissions permits under the EU ETS (Emissions Trading Scheme), so the trading price of permits can be treated as a tax on emissions. Electricity from some sources is taxed under the climate change levy (CCL) if used by industry. The CCL does not apply to electricity from renewable sources or that used by domestic consumers. The Renewables Obligation (RO) requires electricity suppliers to source a proportion of their electricity from renewable generation, which is more expensive than conventional generation. Table 26.4 summarizes this information and provides a calculation of the implicit tax per ton of \( \text{CO}_2 \) resulting from the set of policies. The first column in the table is an estimate of \( \text{CO}_2 \) emissions per kilowatt-hour (kWh) of energy produced. For gas and electricity, the second column is an estimate of the impact of the Renewables Obligation (RO) on electricity prices per kWh. The third column is simply the 2009 to 2010 rate of the climate change levy. For the EU ETS, the government’s central estimate of £22 per ton...
Table 26.4
Implicit carbon taxes, 2009 to 2010

<table>
<thead>
<tr>
<th>Fuel</th>
<th>CO₂ emissions (g/kWh)</th>
<th>RO (p/kWh)</th>
<th>CCL (p/kWh)</th>
<th>EU ETS tax (p/kWh)</th>
<th>Implied tax (£/ton CO₂)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Industrial use</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Electricity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coal</td>
<td>910</td>
<td>0.36</td>
<td>0.47</td>
<td>2.00</td>
<td>2.83</td>
</tr>
<tr>
<td>Gas</td>
<td>393</td>
<td>0.36</td>
<td>0.47</td>
<td>0.86</td>
<td>1.70</td>
</tr>
<tr>
<td>Nuclear</td>
<td>0</td>
<td>0.36</td>
<td>0.47</td>
<td>0</td>
<td>0.83</td>
</tr>
<tr>
<td>Renewables</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Gas</strong></td>
<td>184</td>
<td>0</td>
<td>0.16</td>
<td>0</td>
<td>0.16</td>
</tr>
<tr>
<td><strong>Domestic use</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Electricity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coal</td>
<td>910</td>
<td>0.36</td>
<td>0</td>
<td>2.00</td>
<td>2.36</td>
</tr>
<tr>
<td>Gas</td>
<td>393</td>
<td>0.36</td>
<td>0</td>
<td>0.86</td>
<td>1.22</td>
</tr>
<tr>
<td>Nuclear</td>
<td>0</td>
<td>0.36</td>
<td>0</td>
<td>0</td>
<td>0.36</td>
</tr>
<tr>
<td>Renewables</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Gas</strong></td>
<td>184</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Source: Johnson, Leicester, and Levell (2010).
Note: g = grams, p = pence = £1/100.

of CO₂ is used to estimate the cost of the EU ETS per kWh of electricity. The table reveals the extent to which the implicit tax varies between gas and electricity, between electricity from different sources, and between industrial and domestic users.

A similar situation occurs with the taxation of fuel used in transport. Fuel for road transport is taxed through excise duties and the Road Transport Fuel Obligation (RTFO), which requires fuel suppliers to include a proportion of biofuel. Gasoline for aviation is subject to fuel duty, but turbine fuel is not. The rates of these duties and the implied taxes are reported in table 26.5. The combination of the policies again results in implied taxes that vary considerably across sources of emissions.

As the tables demonstrate, the combination of policies in the United Kingdom toward emissions creates different prices for different sources. Similar observations could be made for other countries. The next question is whether this matters. An argument that has been made is that all units of emissions are the same whatever their source, so
Table 26.5
Implicit carbon taxes, 2009 to 2010

<table>
<thead>
<tr>
<th>Fuel</th>
<th>CO₂ emissions (g/liter)</th>
<th>Fuel duty (p/liter)</th>
<th>RTFO (p/liter)</th>
<th>Implied tax (£/ton CO₂)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gasoline</td>
<td>2303</td>
<td>56.19</td>
<td>0.54</td>
<td>246.33</td>
</tr>
<tr>
<td>Diesel</td>
<td>2639</td>
<td>56.19</td>
<td>0.54</td>
<td>214.96</td>
</tr>
<tr>
<td>Aviation gasoline</td>
<td>2226</td>
<td>34.57</td>
<td>0.15</td>
<td>155.30</td>
</tr>
<tr>
<td>Aviation turbine fuel</td>
<td>2528</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Source: Johnson, Leicester, and Levell (2010).
Note: g = grams, p = pence = £1/100.

the price should be equal across sources. This position cannot be correct. The aim of the taxes is to secure a reduction in emissions. The optimal design of the taxes should be to secure the chosen level of reduction most efficiently. This is what is behind the arguments in section 8.5. The optimal tax should reflect the ease with which emissions can be reduced from each source and the welfare benefit from a reduction. This argument supports differentiation of the taxes, but it does not imply that the UK differentiation is optimal. An alternative argument for uniformity of pricing is that it reduces the incentive for political lobbying to obtain advantageous reductions. This argument cannot be dismissed so lightly.

26.7.2 The EU Emissions Trading Scheme

The EU ETS was established in 2005 as part of the climate policy of the European Union. This is the first major attempt at a widespread scheme to control carbon emissions.

The first phase covered 2005 to 2007, the current second phase finishes in 2012, and a planned third phase will run from 2013 until 2020. The scheme currently covers more than 10,000 installations in the energy and industrial sectors that are collectively responsible for about half of total emissions in the European Union and three quarters of industrial emissions (see figure 26.11). The large emitters of CO₂ within the European Union must monitor emissions, provide an annual report on emissions, and return a quantity of emissions allowances equivalent to the annual CO₂ emissions. In the first two phases of the EU ETS, individual member states proposed national emissions caps based on their assessment of national emissions levels. In the first phase, the European Union accepted the assessments, and there was a shortage of allowances in only a small number of countries. The response of the European Union was to reject the caps initially proposed by most member states in the second phase because they were judged too high.
When the national emissions caps were agreed, the initial allocation of allowances was made to installations by the governments of member states. In the first two phases this was done by “grandfathering” whereby installations were given a quantity of allowances that reflected their past levels of emissions. A more centralized allocation mechanism is proposed for the third phase. The allowances granted to installations can then be traded, and this establishes a price for emissions and provides a financial incentive to reduce emissions.

The price of the tradable emissions allowances (EU allowances or EUAs) has fluctuated over time as shown in figure 26.12. It fell to zero in 2007 when the quantity of allowances exceeded annual emissions. Enhanced price stability would assist investment decisions by energy and other companies. This could be achieved by expanding the market for allowances across a broader range of activities and regions to make the market deeper and more liquid.

The process of allocating national emissions caps on the basis of reports from member states provided an incentive for overstatement of emissions. This was reflected in the first phase by the fact that there was a shortage of allowances in only a small number of countries. The problem was partially resolved in the second phase when the European Union rejected the proposed caps of most member states. For the third phase the proposed centralized allocation mechanism should take away the issue of overstatement, through the system will still be open to political lobbying. The process of grandfathering allowances to firms on the basis of past emissions levels had the advantage of making the EU ETS acceptable to firms. But it also opened a route for windfall profits to be earned through the sale of unwanted permits. It has been estimated
that during the second phase these profits were £1.6 billion annually for the UK power sector alone.

The EU ETS puts a cap on emissions across the European Union, which limits the value of other policy instruments applied to the same emissions. Consider a member state that imposes a second policy, such as an additional tax on a particular sector, that reduces emissions from that sector. The sector will then be left with unused emissions allowances. These can be traded to be used elsewhere in the European Union, so the EU aggregate level of emissions will not fall. The second policy does not achieve any reduction in emissions. This provides an additional factor to be taken into account when member states formulate policy.

When considering innovation, there is a crucial difference between taxation and the use of emissions permits. One advantage of a tax is that firms will not have an incentive to lobby against policy adjustments in response to innovation and technological change, thereby making such adjustments easier. The basic intuition is illustrated in figure 26.13. For a given time period, the optimal policy is to equalize the marginal damage from emissions and the marginal cost of a reduction in emissions. Denoting the marginal cost and damage curves $MC$ and $MD$, respectively, the optimal policy is to emit $Q_0$. 

---

**Figure 26.12**
Allowance price evolution in the EU ETS, 2005 to 2008
tons of carbon at price $P_0$. This can be achieved either by a carbon tax $P_0$, or through a permit trading scheme that sets an emissions cap $Q_0$. In both cases the equilibrium is $(Q_0, P_0)$ before innovation takes place. Innovation would be expected to reduce marginal abatement cost, and so shift the marginal cost curve $MC$ downward to $MC'$. In the absence of revision to the policy, the new equilibrium would be $(Q', P_0)$ under a tax (with quantity adjustment to $Q' < Q_0$) and $(Q_0, P')$ under permit trading (with price adjustment to $P' < P_0$). However, neither of these equilibria would be optimal, since marginal damages and marginal costs would differ. Therefore the policy maker would have an incentive to revise the policy to achieve the new optimum $(Q_1, P_1)$. Under a tax system, this would imply a reduction in the tax ($P_1 < P_0$), from which firms would benefit. Under permit trading, this would imply a tightening of the cap ($Q_1 < Q_0$) and an increase in the emissions price ($P_1 > P'$), from which firms would lose, and which they are therefore likely to oppose. Because the outcome of such lobbying may be uncertain, policy revisions may be less predictable under permit trading than under a tax.

### 26.7.3 International Agreements

The international aspect of GHG emissions has been stressed frequently in this chapter. The global nature of the problem has led to international agreements on emission reductions. We describe the two major agreements and then explore why they have not been entirely successful.
The Kyoto Protocol is an agreement between countries that is linked to the United Nations Framework Convention on Climate Change. The Protocol was adopted in December 1997 and entered into force in February 2005. At the time of writing, 191 states have signed and ratified the Protocol. The only signatory not to have ratified the Protocol is the United States. The Protocol commits 37 countries (known as the “Annex I countries”) to a reduction of carbon dioxide, methane, nitrous oxide, sulphur hexafluoride, hydrofluorocarbons, and perfluorocarbons by 5.2 percent from the 1990 level. International aviation and shipping are excluded from the agreement. The Annex I countries are also required to submit an annual report of all GHGs emissions.

The Protocol introduced several mechanisms through which Annex I countries could meet their GHG emission reduction targets. Joint implementation in Article 6 of the Protocol allows a country with an emission reduction or limitation commitment to earn emission-reduction units from an emission-reduction or emission-removal project in another similar country that can count toward its target. The Clean Development Mechanism is defined in Article 12. This allows a country with an emission-reduction or emission-limitation commitment to implement an emission-reduction project in a developing country to earn salable certified emission-reduction credits. Article 17 of the Kyoto Protocol set out a process for emissions trading that allows countries with unused emission permissions to sell to countries that are over target.

The Copenhagen Accord is a document that delegates at the 15th session of the Conference of Parties to the United Nations Framework Convention on Climate Change agreed to “take note of” at the final plenary meeting in December 2009. An initial draft of the Accord was prepared by the European Union but was not accepted. The accepted version was drafted by the United States and the BASIC countries (Brazil, South Africa, India and China). The limitation of the Accord is that it is not legally binding, nor does it commit countries to agree to a binding successor to the Kyoto Protocol. It sets no clear targets for reductions in emissions. It is fundamentally the outcome of a political compromise that permits some major emitters of GHGs to continue without needing to secure reductions.

An important aspect of the difficulties in achieving an international agreement can be seen in figures 26.14 and 26.15. Figure 26.14 displays the total CO$_2$ emissions of the 20 countries with the highest emissions. The United States and China have the highest emissions, with India coming third. This should be contrasted with figure 26.15 in which emissions per capita are given for the same 20 countries. As can be seen, although the United States is still at top, it is now closely followed by Canada, Australia, and Saudi Arabia. Both China and India are much further down the ranking. Total emissions are
a function of both size and level of per capita GDP. China and India can legitimately claim that focus on total emissions provides a false impression and that expecting them to reduce emissions would be damaging for their economic development. These arguments for fair and equal treatment have made an effective international agreement difficult to secure.

The debates surrounding these agreements illustrate as much as anything that domestic politics and international relations can still prevent the attainment of a beneficial outcome. Even when all the parties recognize what needs to be done, there is no guarantee that an agreement can be secured. This is in part due to the distinction between the private and the social benefits from an agreement. An agreement will only be secured if it is a Pareto improvement, unless some mechanism is put into place to compensate the losers. The effective control of GHG emissions requires international agreement, but a truly global and binding agreement has not yet been secured.

The EU ETS is the primary example of a cap-and-trade system. A limited number of emissions permits are issued (the cap), and these can be sold if not required (the trade). In principle, this ensures an efficient allocation of permits. There have been some initial
26.7.4 Prices or Quantities

Emissions are an externality. Chapter 8 showed that externalities can be controlled by either price instruments (e.g., taxes) or quantity instruments (e.g., output quotas). When the costs and benefits of the emissions and the policies are known with certainty, the instruments are equivalent: for any quantity control there is a price instrument that achieves the same outcome, and vice versa. This equivalence makes the choice of instrument unimportant.

Unfortunately, climate policy is characterized by several dimensions of uncertainty. There is uncertainty about the process of climate change, uncertainty about the consequences of climate change, and uncertainty about the effects of policy intervention. When there is uncertainty, the equivalence between price instruments and quantity controls no longer holds. This was explored in section 8.6. Which policy instrument is...
better depends on the precise set of circumstances. Consider expressing the costs and benefits of policy as functions of the level of abatement. By fixing the marginal cost of reducing GHG emissions, a tax gives some certainty about overall abatement costs, at the price of uncertainty about the resulting level of total emissions. Insofar as abatement costs are unknown to the policy maker or can be subject to shocks (e.g., to economic growth, energy supply or technology) after policy has been set, a tax is a more efficient pollution control device than a quantity instrument, provided that the marginal climate impacts from lower emissions are less sensitive to emissions levels than marginal abatement costs, and vice versa. It should be stressed that only shocks to (marginal) abatement costs typically matter in this context. Shocks to (marginal) damages from emissions will usually entail the same welfare loss under a price as under a quantity instrument, since emissions will remain unchanged in both cases. Thus, following the Weitzman rule, a tax should be preferred to a quantity instrument when the marginal (climate) damage curve is “flatter” than the marginal cost curve. This condition is likely to be met in the case of climate change because the marginal damage from higher than expected emissions is relatively constant in the short run. It is only cumulative—not current—emissions that matter for the climate. This is true at least in the current situation where GHG concentration levels remain significantly below thresholds that might trigger extreme and irreversible events, for instance, the melting of the Greenland and west Antarctic ice sheets. In this context, under a tax the overall economic cost of unexpected shocks to (marginal) abatement costs will be mitigated by (less harmful) increases in short-run emissions, whereas under an inflexible cap-and-trade scheme emissions will remain fixed and overall economic costs will be higher.

The choice between the instruments has to be related to feasibility. National governments currently levy a wide range of different taxes. The adoption of price controls for emissions would just be an addition to the existing range. The same claim cannot be made for an international price policy. There is currently no supra-national body with the authority to levy taxes (even the European Union does not, currently, have the right to levy taxes). An international tax policy would require either the establishment of some new institution or coordinated policy by individual countries. The right to levy taxation is a defining characteristic of a sovereign state. Granting the right to levy taxes to a supra-national body would constitute a partial transfer of sovereignty and would raise significant political questions. Leaving individual countries to set price controls would present the problem that each country has an incentive to be the one not to impose controls.

A policy of quantity control is more feasible from a political perspective, and practical examples already exist. The EU ETS has created a market within Europe for trading
emissions allowances between firms. Article 17 of the Kyoto Protocol was an attempt to apply emissions trading between sovereign states. The Clean Development Mechanism allows affluent countries to implement a project in a developing country that reduces emissions and earns salable certified emission reduction credits toward meeting the Kyoto reduction target. There are two further arguments in favor of quantity controls. The results reported in table 26.3 show an optimal carbon tax that increases over time. If this were to become the expected time path of the carbon tax, an incentive would be created for the early extraction and use of fossil fuels. This problem is avoided if a quantity control is imposed. The final argument relates again to the uncertainty about climate change. The effects of a price control on the level of abatement will be uncertain. In contrast, the outcome of quantity control is known. This observation has been used to argue that if there is a possibility that the consequences of excessive emissions are exceptionally serious, then quantity control is a better option. However, observe that this argument is only valid if the quantity control is set to avoid excessive emissions and excessive emissions are a possible outcome with price controls.

The global nature of the climate change problem requires an international solution. Political issues are likely to rule out an international system of price control, and individual incentives could undermine a decentralized system. These arguments point toward an international version of the EU ETS cap-and-trade systems as the most likely midterm policy solution.

26.7.5 Further Policies

If the government has access to sufficient information (and can use this information correctly), then price and quantity controls will achieve optimal control of emissions even in the presence of uncertainty. Unfortunately, information, such as the responsiveness of firms and consumers to carbon taxes, is limited. This restricts the use of the instruments. Policy choices are also limited by political feasibility and resistance to change. For these reasons governments need to consider a range of policies to supplement the basic interventions.

GHG emissions are generated from numerous sources and a wide variety of activities. The discussion of Pigouvian taxation in section 8.5 emphasized the importance of differentiating taxes to achieve an efficient outcome. Each tax must capture the externality caused by the activity. This could potentially lead to the need for significant differentiation across sectors. In the case of climate change there is too little information available to implement any degree of sophisticated differentiation. In other words,
a uniform price or quantity control scheme may be too blunt for dealing with emissions. This provides support for the introduction of regulations that can be more specific to each activity. Governments currently use numerous regulatory policies to address climate change. These include obligations on energy suppliers to improve the energy efficiency of their customers, requirements on use of biofuels in petrol, and stringent building regulations. These regulations should be seen as supporting the basic price or quantity control.

An important component of any reduction in GHG emissions must be technological advance. Government intervention to encourage technological change is justified only if arguments can be advanced that market incentives are inadequate. There could be several reasons for such market failure, including inability to appropriate gains and asymmetric information. If market-driven change is too slow, then what is required is support offered by governments to increase the rate of technology change. The problem facing the government is knowing how to do this. There is a long history of mistakes in government-sponsored technology projects. One prime example is the Concorde project between the United Kingdom and France that developed a small, fast, expensive aircraft just as the market was looking for large, economical aircraft to deliver cheaper mass transport. It is clearly preferable to provide incentives for technological change, such as tax breaks, and let the private sector find the projects.

The discussion has explored a range of additional policies that can complement price or quantity interventions. Two final points can be made about policy choice. First, the set policies need to be coherent and related to the specific market failures in each case. Second, and more important, it is crucial that the set of policies operate in a consistent way. For example, the European Union has the ETS, a 20 percent emissions reduction target, a 20 percent renewables target for energy supply, and a 20 percent energy efficiency target. These are all intended to be achieved in 2020. There is no evidence that these are consistent targets, or that the policies designed to achieve them will function coherently.

26.8 Conclusions

Writing a balanced chapter on the economics of climate policy is a challenging task. The issue of climate change has polarized opinions with those on each side of the debate taking entrenched positions. For some, climate change has become a dogma that accepts no deviation from an agreed consensus. To these protagonists every unusual weather event is further evidence of climate change and every academic study is behoven to
discover evidence of climate damage. Any evidence that glaciers are not melting and temperatures are not rising are simply inconvenient truths that cannot be allowed to derail the climate bandwagon. Climate science can be likened to an industry engaged in rent-seeking: the members of the industry need the hysteria to be sustained to continue to receive financial support. On the other side, the so-called deniers of climate change are equally entrenched and vociferous with a blanket refusal to accept any evidence whatsoever that supports the hypothesis of anthropogenic climate change.

The existence of these extremes does not assist the conduct of a balanced policy debate. It is clearly correct to test the evidence on climate change; it is an academic duty in fact to question any claim and subject it to the most rigorous scrutiny. This is especially important when proposed policies could impose enormous present costs in exchange for uncertain future returns. It has been suggested that the potential costs of climate change are so large that we must act even though current knowledge is so limited. But this is not the logic that economists would apply in any other area.

There is one point that is clear from our discussion of the economics of climate policy. Economic analysis in this area is in the very early stages of development, and much more work needs to be done to refine predictions of economic costs. This will be a difficult task to perform with limited knowledge of how the climate will change and how this change will affect economic activity. It is a task that needs completing if policy is to be founded on a firm basis. The set of policy tools to secure emissions reductions are known. What the analysis must do is secure international support for their use and determine the level of intervention.

Further Reading

The Stern Review provides a description of the science and an economic assessment of the damage from climate change:


A critical assessment of some aspects of the science is given in:


The problems in the science underlying the hockey stick are explored in:


A discussion of the Stern Review with a focus on the rate of discounting is given in:

The DICE model is described in detail and used to analyze policy in:


The ecological discount rate is explored in:


An extensive discussion of environmental policy is in:


The final reading is not an academic book, but it is a book about academic practice. It lays bare the science behind the Hockey Stick and the practices of those who constructed, promoted it, and support it. The book is not a comfortable read for anyone who places a value on academic integrity. The potential catastrophic consequences of global warming require that it be investigated using the very best scientific methods with rigorous scrutiny of all claims. Mountford shows that in one branch of climate science, quite the opposite was the case.


The references for figure 26.6 are:


---

**Exercises**

**26.1** The commissioning of nuclear power stations to replace existing power stations using gas and coal will reduce carbon emissions and help combat global warming. The downside is that nuclear power leaves a legacy of waste that will be radioactive for many hundreds of years. Analyze the trade-offs in the choice between these alternatives, focusing on the relative value placed on the welfare of present and future generations.
26.2 “My first heresy says that all the fuss about global warming is grossly exaggerated. Here I am opposing the holy brotherhood of climate model experts and the crowd of deluded citizens who believe the numbers predicted by the climate models. Of course, they say, I have no degree in meteorology and I am therefore not qualified to speak. But I have studied the climate models and I know what they can do. The models solve the equations of fluid dynamics, and they do a very good job of describing the fluid motions of the atmosphere and the oceans. They do a very poor job of describing the clouds, the dust, the chemistry and the biology of fields and farms and forests. They do not begin to describe the real world that we live in.” (Freeman Dyson, theoretical physicist, 2009.) Discuss the relevance of these views for planning a climate policy.

26.3 (Gollier) According to the Stern Review, if we don’t act quickly to reduce our emissions of greenhouse gases, global warming will be damaging our economies by as much as 35 percent of GDP by the year 2200. This cost includes reduced agricultural productivity, fatalities due to natural catastrophes, damages to buildings, increased air conditioning, and the loss of environmental assets. The Stern Review is much less optimistic than earlier studies on this question. For example, Nordhaus (2007) estimated the cost of climate change at “only” 3 percent of global GDP, and a group of economists associated with Jorgenson predicted a net benefit of global warming for the US economy that could be as much as 1 percent of GDP in some scenarios (Jorgenson et al. 2004). How do you explain such large differences in the economic evaluations of the cost of global warming?

26.4 A project that will develop a very poor country provides benefits of $5 million over the next year \((t = 0)\). However, the consequences will be severe environmental damage in 200 years \((t = 200)\). The present value of this damage evaluated at \(t = 200\) is estimated to be $1 trillion (1 followed by 12 zeros). There are no costs or benefits between these two dates.

a. Is it worth undertaking the project at a 10 percent discount rate?

b. Is it worth undertaking the project at a 6 percent discount rate?

c. If 10 percent is the interest rate on private capital markets, is this the rate that should be used?

d. What arguments can be made for using a lower discount rate?

e. Are these kinds of calculation the correct way to analyze the problem?

26.5 The socially efficient discount rate in the Stern Review would be around 1.3 percent per year \((0.013)\). Would you find such discount rate acceptable?

26.6 The Stern Review assumes that we are risk-averse, in that facing the risk of a loss is always worse than incurring the mean loss with certainty. The difference between the two is the risk premium (i.e., the willingness to pay to eliminate the risk). In the expected utility framework, risk aversion is represented by a decreasing marginal utility of income, in that the chance of a large positive deviation from the mean loss has a much larger negative effect on welfare than the corresponding chance of a large negative deviation. What objections would you make to the logarithmic utility of income used to determine the risk premium in the Stern Review?

26.7 Compute the discount factor \(\frac{1}{(1+r)^t}\) for values of the discount rate \(r = 1, 3.5,\) and 5 percent interest rates and \(t = 20, 50,\) and 100 years. Do these computations show that the choice of discount factor is important for the decision on whether to implement a (costly) policy now that will only provide benefits in the distant future?
What are the implications of the absence of scientific certainty about the impacts of climate change on the optimal timing of the effort to curb emissions?

If global warming caused a 2 meter rise in worldwide sea level, what do you expect would happen to the price of land? Could this be counted as a benefit of climate change? What if the rise in sea level was 20 meters?

Climate change involves intergenerational discounting to trade off costs and benefits at different points in time. The theory of discounted utility is the most widely used approach to make those intergenerational trade-offs. The commonly used discounting model assumes that total utility can be decomposed into a weighted sum of utility in each period of time (Ramsey 1928):

\[ U_t = \sum_{\tau=0}^{\infty} d(\tau) u_{t+\tau}, \]

where \( U_t \) is the total utility discounted from the perspective of the current time \( t \), \( d(\tau) \) is the discount function, \( u_{t+\tau} \) is the instantaneous utility in period \( t + \tau \), and the flow of utilities is over an infinite horizon for this intergenerational model. Time preferences are represented by the rate at which the discount function declines, \( d'(\tau) < 0 \).

A. The standard discount function is the exponential discount function \( d(\tau) = \delta^\tau \) with \( 0 < \delta < 1 \). In this case the discount factor, \( \delta \), is assumed to be independent of the time horizon. Show that the exponential discount rate satisfies the property of dynamic consistency: that is, preferences at one point in time do not change with the passage of time. To make your point, consider a project that costs \( C \) at period \( t = 2 \) and produces a benefit \( B \) at period \( t = 3 \). Show that if the \( NPV \) of the project is positive at \( t = 0 \), then it will be positive at \( t = 1 \).

B. The exponential discount function is the only discount function that produces dynamically consistent preferences. However, it fails to match some empirical regularities. Can you provide some examples?

In making intertemporal trade-offs, people seem to be more impatient when making short-run trade-offs (today vs. tomorrow) than when making long-run trade-offs (day 100 vs. day 101). To match this empirical regularity, the following "generalized hyperbola" discount function has been proposed:

\[ d(\tau) = (1 + \alpha \tau)^{-\gamma/\alpha}. \]

A. Show that this discount function displays a higher discount rate in the short run than in the long run. For the special case of \( \gamma = \alpha \) (so that \( d(\tau) = \frac{1}{1 + \alpha \tau} \)) plot the discount function and compare it to exponential discounting.

B. Use a similar investment trade-off as in the previous exercise to show that such preferences fail to satisfy dynamic consistency (i.e., the passage of time may change a person’s preferences or change the \( NPV \) of a project from positive to negative).

Re-work the previous exercise with another non-exponential discount function. This is called the quasi-hyperbolic discount function

\[ d(\tau) = \begin{cases} 1 & \text{if } \tau = 0 \\ \beta \delta^\tau & \text{if } \tau = 1, 2, \ldots \end{cases} \]

What is special about the case of \( \beta = 1 \)?
26.13 (Geant, Guesnerie and Lasry, 2012) This exercise explores the role of the elasticity of substitution between goods and environment, \( \mu \), in the determination of the discount factor in the long run. The utility function is given by

\[
u(C_t, y_t) = \left[ C_t^{(\mu-1)/\mu} + \bar{Q}^{(\mu-1)/\mu} \right]^{\mu/(\mu-1)}.
\]

where \( C_t \) is consumption at time \( t \) and \( \bar{Q} \) is the fixed level of environmental quality. Consumption, \( C_t \), is assumed to grow at a constant rate

a. Consider the case of \( \mu > 1 \). Write utility in terms of \( C_t \) and \( \bar{Q} \). What happens to the value of \( \left[ \frac{\bar{Q}}{C_t} \right]^{(\mu-1)/\mu} \) as \( t \) increases? What will happen to the level of utility as \( t \) increases?

b. Now consider \( \mu < 1 \). Write the utility function in terms of \( \bar{Q} \) and \( \frac{\bar{Q}}{C_t} \). What happens to the value of \( \left[ \frac{\bar{Q}}{C_t} \right]^{(1-\mu)/\mu} \) as \( t \) increases? What happens to the value of utility?

c. Explain the policy implications of these different results.

26.14 (Environmental Kuznets curve) The relation between environmental quality and GDP is called the environmental Kuznets curve. Such a curve can be drawn for a general measure of environmental quality and for each pollutant separately.

a. Do you expect there to be a positive or a negative relationship between the gross domestic product (GDP) of a country and environmental quality? Does the answer depend on whether the country is developing or developed?

b. Will all pollutants for a given country have the same environmental Kuznets curves? If so, what will their shapes be? As the country develops, will all pollutants be at the same point on their curves?

c. If increased GDP is associated with improved environmental quality, does this imply that a cleaner environment comes at no cost? If so, how can you explain the frequent objections to the imposition of environmental regulations?

26.15 The idea behind “emission reduction” is that rather than deal with the consequences of emissions, businesses should avoid the creation of emissions in the first place by changing production processes.

a. Firms can use carbon-capture technologies to capture carbon that would otherwise enter the atmosphere. It is valid to argue that emissions represent waste, so should reducing emissions reduce costs?

b. A firm is observed to reduce costs and increase profits through emission reduction. Is the only explanation for this observation that the firm was operating inefficiently before implementing emission reduction?

26.16 An economy has two industries, \( i = 1, 2 \). One industry is competitive, the other industry is a monopoly. The demand functions for the products of the industries are described as

\[ p_i = a - b y_i, \]

where \( y_i \) is the output of industry \( i \). There are constant returns to scale in both industries, with a marginal cost constant at \( c \). Each firm produces emissions equal to its level of output. Each unit of emissions causes a loss in consumer welfare of \( e_ig_i \). In order to produce, a firm must have a quantity of pollution permits equal to the quantity of emissions. The government chooses to make a quantity of permits, \( Z \), available.
a. If the permits are allocated by the government directly to firms with no trading permitted, what is the efficient allocation of licences?
b. If the permits are allocated but can then be traded, what is the efficient allocation and the resulting allocation after trade?
c. If the permits are auctioned, what is the resulting allocation?
d. Which process for allocating permits is most efficient?
Index

Absolute poverty, 482
Accountability, of government, 643–46
Addictive activity, 54–55
“Adult goods,” 463
Ad valorem taxation, 274–77
Adverse selection, 297, 300, 304–305
and antipoverty programs, 302
government intervention for, 305–307, 330
and health care reform, 332
vs. moral hazard, 334, 336
policy recommendations on, 334–35
AFDC (Aid to Families with Dependent Children), 688, 694
Agenda manipulation, 358–61
Agent monotonicity, 548, 549, 555, 559
Agglomeration rents, 722–23
Aggregate demand, 25
Aggregate economic model, 915
Aggregate poverty gap, 484–85
Aggregate preferences. See Arrow’s Impossibility Theorem; Preference aggregation
Aggregate production function, in overlapping generations economy, 755–56
Aggregate supply, 25
Aggregate welfare, 881
Aid to Families with Dependent Children (AFDC), 688, 694
Airline industry, price wars in, 260
Akerlof, George, 232
AK model, 824, 824–25
“Alabama” paradox, 371–72
Allocation
optimal, 611, 617–18, 622
and Pareto-efficiency, 32–33
of public goods, 152, 153, 154
truly feasible, 613
of utility, 423, 426
Allocation mechanism, 607, 609, 610
alternatives among, 607
anonymous, 615
incentive-compatible, 608–609
optimal, 609, 622
for public goods
personalized prices, 159–63
preference revelation mechanisms, 163–70
Samuelson rule, 147, 154–55, 161, 182, 195, 209, 687
voting, 156–59
Allocation problems, and Pareto-efficiency, 438, 439
Altruism, 78
as sustaining social security, 790–91, 796
Ambiguity theory, and tax evasion, 597
Analysis of policy. See Policy analysis
Ancillary condition, 72
Annuity market, 335
Anomalies, 6, 51
downplaying of (Smith), 64
Antipoverty programs, and adverse selection problem, 302
Antitrust legislation, US, 277
Apportionment problem, and “Alabama” paradox, 371–72
Approval voting, 363–64
Arbitrage, intertemporal, 870–72
Arbitrage opportunity, 871
Argentina, democracy and social security in, 798
Arm’s-length prices, 714, 717–18, 719, 720
Arrow–Debreu economy, 22
Arrow’s Impossibility Theorem, 347–50, 381, 441–42, 446
and household utility function, 466
Asian disease problem, 60
Assessment, impact, 860–61
Assumptions
of awareness of own best interests, 51
on competitive economy, 41–43
Cournot (competitors’ prices regarded as fixed), 262
of customers seeing prices as parametric, 14
in evaluation of inequality and poverty, 458
of government efficiency, 133
of identical agents, 123, 173, 323
of inability to affect prices, 255
of independence of demands, 516
for market efficiency, 223
of Median Voter Theorem, 157
of policy analysis, 6
of self-interest, 6
separability, 79
in standard economic analysis, 51
of symmetric information, 14–15, 42, 336
in Two Theorems of Welfare Economics, 39
on voting, 365
Asymmetric information, 281, 297–99, 334
and hazard insurance, 302–305
and hidden knowledge vs. hidden action, 300–302
as moral hazard (hidden action), 323
in insurance, 323–30
and public provision of health care, 330–33
and screening, 307–13
with government intervention, 313–14
and signaling, 314–15
educational, 316–21
policy implications for, 321–22
Athletes, and performance-enhancing drugs, 232–33
Atkinson measure of inequality, 478–79, 480, 496
in OECD study, 481
Attitudes, and intergenerational inequalities, 495
Attribute framing, 61
Auctioneer, Walrasian, 42
Auditing, for detection of tax evaders, 583–86
in compliance game, 590–93
critical mass of compliance crucial in, 600
Australia
climate change effects in, 911
cost–benefit analysis in, 847
emissions level of, 935, 936, 937
equalization grants in, 689
poverty line in, 489
public spending in, 91
social value of life in, 869
tax revenues in, 102
Austria
hidden economy of, 575–76
social value of life in, 869
tax rates of, 691
Autarkic equilibrium, 781–82
Autarky, 736
Automatic enrollment, 69
Autonomy of choice
and conformism bias, 62
and interventionist policies, 52
Bandwagon effect, 235–36
Bangladesh, and climate change, 911
Bank’s set, 361
Bargaining, over externalities, 245–47
Barriers to entry, 277–78
patents as, 278
Barro, Robert, 876
Barter, 42
Baumol’s law, 125–26
Behavioral economics, 6, 51–52
and imperfect decision-making, 53
(β, δ) model of self-control, 55–57
and confidence bias, 60
and confirmation bias, 59
and conformism bias, 62–63
and framing bias, 60–62
and gambler’s fallacy, 58
identity and social norms in, 63–64
present bias, 54–55
and reference-dependence bias, 57–58
in savings example, 53–54
lack of clarity in conception of, 80
and markets, 64–65
complementary mistakes in, 65–66
money pump in, 65
and rationality tug-of-war, 67
and other-regarding preferences, 76
market impact of, 78–80
Index

social preferences, 77–78
in ultimatum game, 76–77
and policy, 67–68
automatic enrollment, 69
behavioral economics as complementary to
substantive policy, 70
and internalities vs. externalities, 68–69
SMArT Plan, 69–70
prospects for, 81
and tax evasion, 594–98
and welfare analysis, 71–72
choice-based analysis, 72–73
and global warming, 74–76
preference refinement and structural
modeling in, 73
Behavioralism, 52
Belgium
governmental decentralization in, 657
party “list system” of voting in, 371
poverty line in, 489
secessionism in, 657, 690
social value of life in, 869
tax evasion data from, 588
tax mimicking in, 694
Bell System telephone company, breakup of, 277
Benefit(s), attaching valuations to, 862
Benefit–cost ratio (BCR), 853–54
Bergson–Samuelson social welfare function, 425
Bertrand behavior, 256
Besley–Coate model, 379
Best-response function, 151–52
Beveridge, William, 461
Bias
confidence, 60
confirmation, 59
conformism, 62–63
framing, 60–62
hindsight, 59
opportunity-cost, 69
present, 54–55, 56, 69
reference-dependence, 57–58
status quo, 70
Birth rate, and pensions crisis, 776, 778
and dependency ratio, 797
Black economy, 574. See also
Hidden economy
Bohm, Peter, 170
Borch’s theorem on mutual insurance, 646
Borda voting, 361–62
and approval voting, 363–64
Brazil
and Copenhagen Accord, 898, 935
democracy and social security in, 798
emissions level of, 936, 937
Breakfast cereal industry, 282
Bribery, 134–35
policing of, 414
and rent-seeking, 389
Buchanan, J., 195, 199
Budget(s), hard and soft, 650–56
Budget constraint, 18, 20, 25, 28, 30
of different governments, 708
in federal system, 650–56
lifetime, 620
in optimal-tax determination, 510, 512
and taxation threshold, 542
and Walras’s law, 21
Budget-setting, and bureaucracy,
132–33
Bunching effect, 258
Bundling
of Internet Explorer with Windows, 278
of Media Player with Windows (Europe), 278
Bureaucracy, 130–32
and budget-setting, 132–33
and expenditure, 128
rent creation by, 407
and rent-seeking, 404
Cambodia, Khmer Rouge equality in, 423
Canada
emissions level of, 935, 936, 937
equalization grants in, 689, 690
and Smoot–Hawley Act, 736
social value of life in, 869
tariffs on processed food in, 738
tax revenues in, 99, 101, 102
cigarette taxes, 676–77
vertical expenditure externalities in,
687–88
Cancun, Mexico, WTO meeting in, 740
Capital
interjurisdictional competition for, 665–66
mobility of
across countries, 709, 741
and fiscal competition, 666–67
and mobility of labor, 685
and taxation system, 710–11
tax on, 822, 823, 829, 838
income tax, 622, 624–25, 666–67, 686, 823
in international environment, 709–11
optimal, 821–22, 823
policy experiments on, 831
and tax exporting, 677–79
Capital asset pricing model (CAPM), 875–76
Capital export neutrality, 710
Capital gains, 459
Capital–labor ratio, optimal, 762–65, 766
Capital market, equilibrium on, 784
Capital ownership neutrality, 710
Capital stock, and social security programs, 776
reform of, 800
Cap-and-trade system, EU ETS as, 936,
See also Emission Trading Scheme, EU
Carbon dioxide (CO₂), 901
accumulation of, 895, 895–96
predictions on effects of, 908, 909
Carbon emissions control, EU Emission Trade
Scheme as, 931
Carbon tax, 916–17, 939
in DICE model, 922
Cardinal full comparability, 445–46, 449
Cardinal unit comparability, 445, 448–49
Cardinal utility, 444–45, 448
Cartel agreement, 260
CBA. See Cost–benefit analysis
Cell phone licenses, allocation of, 399
Cereal industry, 282
Chile, democracy and social security in, 798
China
and Copenhagen Accord, 898, 935
emissions level of, 935–36, 937
and Doha talks, 740
Chlorofluorocarbons (CFCs), 901
Choice-based welfare analysis, 72–73
Cigarettes, taxation of in Canada, 676–77
Circumstances,
and equality of opportunity, 489
Clarke–Groves mechanism, 166–68, 609, 610
Clarke tax, 168–69
Clean Development Mechanism, 935, 939
Climate change (anthropogenic), 895, 940–41
and climate policy in practice, 929, 939–40
EU Emissions Trading Scheme (EU ETS), 929, 931–34, 936, 938–39, 940
international agreements, 934–37
prices or quantities as control, 937–39
pricing and policies, 929–31
consensus opinion on, 895, 895–96
deniers of, 941
and ecological discounting, 924–29
economics integrated with, 910–11, 941
and competing generations, 917–24
and effects of policy, 915–17
and modeling of economic damage, 912–15
and sources of damage, 911–12
local effects of, 908–909
and science of global warming, 900–903
special features of, 897–900
and temperature increase, 902, 904, 906, 907, 908
effects of, 911
Climate change levy (CCL), 929
Climate policy. See Climate change
Climate sensitivity, 896, 907
Clinton, Bill, 331–32

Club(s)
and economy overall, 199–207
for clubs with equal fees, 201–204
for clubs with unequal fees, 204–207
large clubs, 200–207
small clubs, 200
single-product, 193
with fixed utilization, 194–96
with two-part tariff, 197–99
with variable utilization, 196–97
Club good, 149, 191, 192
and congestion, 193–94, 195
and local public goods, 193
and pure public goods, 216
CO\textsubscript{2}. See Carbon dioxide
Coase theorem, 242–47
Cobb–Douglas production function, 767, 822, 913
Cobb–Douglas utility function, 561
Cognitive hierarchy, 66
Collective action, and CO\textsubscript{2} reduction as prisoner’s dilemma, 898
Collusion
detection of, 282–83
tacit, 260, 282, 283
Collusiveness, 260
Command economy, 4, 607
Commitment device, 680
Commodity taxes, 505, 622–24
applications on, 523–24
and Indian survey, 526–28
marginal reforms, 524–26
and cross-border shopping, 672, 677, 726
deadweight loss from, 506–509, 517
efficiency of, 528–30, 531, 704
with equity considerations, 521–23
with income taxes, 622, 627
vs. lump-sum, 428
marginal benefit of (CBA), 858
optimal, 505, 509–12, 513
and production efficiency, 513–14
and redistribution, 526–28
and optimal public sector prices, 530
rules for, 515–21
Competition
and efficiency, 13, 26–27
in exchange economy, 33–37
Pareto-efficiency, 31–33, 436–41 (see also Pareto-efficiency)
and production and exchange economy, 37–39
with single consumer, 27–31
intergovernmental, 693
market vs. government, 665, 694, 695
measurement of, 258–60
model of, 13
non–price, 257
Competition, imperfect. See Imperfect competition
Competitive behavior, assumption of, 41
Competitive economies, 14–15, 751
as allocation mechanism, 609
assumptions on, 41–43
as atemporal, 751
as blind to injustice, 610
exchange economy, 15–22, 33–37
idealized, 768
Competitive economies (cont.)
- inefficiency in, 225–28
- as Pareto-efficient, 884
- private goods in, 148
- production and exchange economy, 22–26
  and Two Theorems of Welfare Economics, 37–39 (see also Two Theorems of Welfare Economics)

Competitive environment
- and altruism, 79–80
- and ultimatum game, 77

Competitive equilibrium
- of exchange economy, 17, 19
- and overlapping generations model, 752, 770
- Pareto-efficient allocation as, 884
- and Pareto-preferred allocation, 762
- production and exchange economy, 26

Competitive market, pecuniary externalities in, 224

Complementarity, between behavioral economics and substantive policies, 70

Complementary mistakes, 65–66

Complete Dissipation Theorem, 393–94, 414
- and lobbying, 405
- and rent creation, 407

Complete ordering, Pareto-preference, 439

Compulsory insurance, 305

Concentration, degree of, 258–59

Concorde project, 940

Condorcet method, 351

Condorcet paradox, 347, 348, 349

Condorcet winner, 351, 352, 353, 360, 373, 380
- and approval voting, 364
- and runoff voting, 364
- and “swing voter,” 377

Confidence bias, 60

Confirmation bias, 59

Conformism bias, 62–63

Congestion
- and clubs, 192, 199, 203

Consumption, efficiency in, 38

Consumption externality, 224

Consumption possibility frontier, 760

Consumption taxes, vs. income tax, 620–24
- on capital, 624

Contestability, 258

Contract, second-best, 327–29

Contract curve, 34
- as Pareto-efficient, 36

Contract laws, 119–20, 120

Contribution campaign, 179–81

Convexity, 247

Copenhagen Accord (2009), 898, 898–99, 935

Corporate debt, 109

Corporate taxation environment, 711–15
  Corruption, 134–35

Cost(s), attaching valuations to, 862

Cost–benefit analysis (CBA), 9, 847–51
- and climate change, 895, 900, 917–24
- difficulties of, 862–65
- and double counting, 861
- and Flood Control Act, 847
- and politics of actual decision-making, 889
  practical application of, 898

Constitution, US, Sixteenth Amendment of, 89

Consumer choice, theory of, 443

Consumers
- decision-making of, 819
- in exchange economy, 15–18
- in overlapping generation economy, 754

Consumption, efficiency in, 38

Consumption externality, 224

Consumption possibility frontier, 760

Consumption taxes, vs. income tax, 620–24
- on capital, 624

Contestability, 258

Contract, second-best, 327–29

Contract curve, 34
- as Pareto-efficient, 36

Contract laws, 119–20, 120

Contribution campaign, 179–81

Convexity, 247

Copenhagen Accord (2009), 898, 898–99, 935

Corporate debt, 109

Corporate taxation environment, 711–15
  Corruption, 134–35

Cost(s), attaching valuations to, 862

Cost–benefit analysis (CBA), 9, 847–51
- and climate change, 895, 900, 917–24
- difficulties of, 862–65
- and double counting, 861
- and Flood Control Act, 847
- and politics of actual decision-making, 889
  practical application of, 898
<table>
<thead>
<tr>
<th>Topic</th>
<th>Page(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index</td>
<td>953–954</td>
</tr>
<tr>
<td>principles of, 856</td>
<td>856–57, 865, 868</td>
</tr>
<tr>
<td>common unit of measurement, 856–57, 865, 868</td>
<td></td>
</tr>
<tr>
<td>decision criteria, 861–62</td>
<td></td>
</tr>
<tr>
<td>impact assessment, 860–61</td>
<td></td>
</tr>
<tr>
<td>revealed preferences, 857</td>
<td></td>
</tr>
<tr>
<td>valuing market goods, 858–59</td>
<td></td>
</tr>
<tr>
<td>valuing nonmarket goods, 859–60</td>
<td></td>
</tr>
<tr>
<td>process of, 851–52</td>
<td></td>
</tr>
<tr>
<td>discounting future values, 853–56, 861, 863</td>
<td></td>
</tr>
<tr>
<td>theoretical foundations of, 876</td>
<td></td>
</tr>
<tr>
<td>distributive weights, 882–83</td>
<td></td>
</tr>
<tr>
<td>market prices and shadow prices, 883–88</td>
<td></td>
</tr>
<tr>
<td>social welfare evaluation, 877–81</td>
<td></td>
</tr>
<tr>
<td>and valuing of future, 869–76</td>
<td></td>
</tr>
<tr>
<td>and valuing of life, 865–69</td>
<td></td>
</tr>
<tr>
<td>Cost diffusion, 137–38</td>
<td></td>
</tr>
<tr>
<td>Cost pessimism or optimism, 855</td>
<td></td>
</tr>
<tr>
<td>Countries, large vs. small, 733</td>
<td></td>
</tr>
<tr>
<td>Cournot assumption, 262</td>
<td></td>
</tr>
<tr>
<td>Cournot behavior, 256, 263</td>
<td></td>
</tr>
<tr>
<td>Cournot equilibrium, for duopoly, 273–74</td>
<td></td>
</tr>
<tr>
<td>“Creative destruction.” 825</td>
<td></td>
</tr>
<tr>
<td>Crime, incentives for in design of punishments, 585–86</td>
<td></td>
</tr>
<tr>
<td>Cross-border shopping, 672, 677, 726</td>
<td></td>
</tr>
<tr>
<td>Deadweight loss</td>
<td></td>
</tr>
<tr>
<td>in commodity taxation, 506–509, 517</td>
<td></td>
</tr>
<tr>
<td>from tax evasion, 593</td>
<td></td>
</tr>
<tr>
<td>Debt, 112</td>
<td></td>
</tr>
<tr>
<td>corporate, 109</td>
<td></td>
</tr>
<tr>
<td>government, 107–12</td>
<td></td>
</tr>
<tr>
<td>and Ricardian equivalence, 797</td>
<td></td>
</tr>
<tr>
<td>household, 109</td>
<td></td>
</tr>
<tr>
<td>Decentralization, 36–37, 424, 426–27, 436</td>
<td></td>
</tr>
<tr>
<td>and allocation, 616</td>
<td></td>
</tr>
<tr>
<td>governmental competition through, 693</td>
<td></td>
</tr>
<tr>
<td>and Second Theorem of Welfare Economics, 39</td>
<td></td>
</tr>
<tr>
<td>Decentralization of government activities, 639, 641, 644</td>
<td></td>
</tr>
<tr>
<td>and accountability, 644</td>
<td></td>
</tr>
<tr>
<td>advantages of, 640</td>
<td></td>
</tr>
<tr>
<td>evidence on by functions, 655</td>
<td></td>
</tr>
<tr>
<td>worldwide, 653–54</td>
<td></td>
</tr>
<tr>
<td>and size of government, 680–81</td>
<td></td>
</tr>
<tr>
<td>socially optimal, 642</td>
<td></td>
</tr>
<tr>
<td>Decision criteria, for cost–benefit analysis, 861–62</td>
<td></td>
</tr>
<tr>
<td>Decision-making in economic models, 4, 14</td>
<td></td>
</tr>
<tr>
<td>imperfect</td>
<td></td>
</tr>
<tr>
<td>and (β, δ) model of self-control, 55–57</td>
<td></td>
</tr>
<tr>
<td>and confidence bias, 60</td>
<td></td>
</tr>
<tr>
<td>and confirmation bias, 59</td>
<td></td>
</tr>
<tr>
<td>and conformism bias, 62–63</td>
<td></td>
</tr>
<tr>
<td>and framing bias, 60–62</td>
<td></td>
</tr>
<tr>
<td>and gambler’s fallacy, 58</td>
<td></td>
</tr>
<tr>
<td>identity and social norms in, 63–64</td>
<td></td>
</tr>
<tr>
<td>present bias, 54–55</td>
<td></td>
</tr>
<tr>
<td>and reference-dependence bias, 57–58</td>
<td></td>
</tr>
<tr>
<td>in savings example, 53–54</td>
<td></td>
</tr>
<tr>
<td>Decision weights, in tax evasion decision, 594</td>
<td></td>
</tr>
<tr>
<td>Default, sovereign debt, 107</td>
<td></td>
</tr>
<tr>
<td>Defense spending, 91</td>
<td></td>
</tr>
<tr>
<td>on national level only, 636, 654</td>
<td></td>
</tr>
<tr>
<td>Deficits, 107</td>
<td></td>
</tr>
<tr>
<td>Defined benefits system, 801–802</td>
<td></td>
</tr>
<tr>
<td>Defined contributions system, 801–802</td>
<td></td>
</tr>
<tr>
<td>Demand</td>
<td></td>
</tr>
<tr>
<td>and allocation mechanism, 610</td>
<td></td>
</tr>
<tr>
<td>compensated, 520</td>
<td></td>
</tr>
<tr>
<td>and equity, 523</td>
<td></td>
</tr>
<tr>
<td>excess, 20–22, 26</td>
<td></td>
</tr>
<tr>
<td>and price, 26</td>
<td></td>
</tr>
<tr>
<td>Demand curve, 858</td>
<td></td>
</tr>
<tr>
<td>Demand elasticities and merger results, 284</td>
<td></td>
</tr>
<tr>
<td>in Wagner’s law, 125</td>
<td></td>
</tr>
<tr>
<td>Demand function homogeneous of degree 0, 20</td>
<td></td>
</tr>
<tr>
<td>in tax rate calculation, 524</td>
<td></td>
</tr>
<tr>
<td>Demand–supply balance. See Supply–demand balance</td>
<td></td>
</tr>
</tbody>
</table>
Democracy. See also Voting and social security, 798
Demographic variables, 460
Denmark
equalization grants in, 689
inequality in, 480–81
social mobility in, 493
social value of life in, 869
Dependency ratio, 779
Destination-based taxation (destination principle), 672, 713, 730
Deterministic rent-seeking game, 391–94
Developing countries
damage from climate change in, 911
and tariffs, 738
and trade agreements, 736, 737–38, 739, 741
Developmental models, 123–24
Devolution, and common resource problem, 137–38
Diamond–Mirrlees Production Efficiency Lemma, 513–14, 708
DICE (Dynamic Integrated model of Climate and Economy), 912–13, 916–17, 920, 922
Differentiation, product. See Product differentiation
Directly unproductive activities, 390
Direct revelation mechanism, 611–12
Disappointment theory, and tax evasion, 596–97
Discounting future values, 853–56
Discount rate, 56
and cost–benefit analysis, 861, 863
in climate policy, 918, 924
ecological, 924–29
and global warming action, 74–75
and Ramsey rule, 929
and UK Green Book, 873–75
Discount rate for produced goods, 925
Discrimination, in price, 263, 264
Distribution. See also Redistribution and cost–benefit analysis, 862–63, 879, 882
equity in, 119
and identical-consumer assumption, 123
and Pareto-efficiency, 32–33
and Second Theorem of Welfare Economics, 423
and social decision rules, 882
and social welfare function, 440
and taxation (Indian commodity taxes), 526–27
Distributive arguments for multi-level government, 640
Distributive politics, 377
Doha Round of trade talks, 737, 740, 741
Downs, Anthony, 372
Downsian model of political competition, 372–73, 379
Dupuit, Jules, 848
Duverger law, 379
Dvorak’s keyboard, 235–36
Dynamically efficient states, 764
Dynamically inefficient states, 766
Dynamic aspects of economic activity, 267, 399
Dynamic consistency problem, self-control problem as, 55
Dynamic efficiency, 769, 770
Dynamic inconsistency, 56
Dynamic Integrated model of Climate and Economy (DICE), 912–13, 916–17, 920, 922
Dynamic systems, with positive feedback, 495
Ecological discounting, 924–29
Economic allocation, 609
Economic forecasting, 915
Economic growth, 809, 837–38
endogenous growth, 823–26, 838
and government expenditure, 826–29
and taxation, 833
exogenous growth, 810, 838
and constant saving rate, 810–16
optimal taxation in, 817–23
and taxation, 809, 810
and empirical evidence, 833–37
reform of, 829–33
Economic incidence of tax, 268–74
Economic models. See Models
Economics
behavioral, 6, 51–52, 70, 80, 81 (see also Behavioral economics)
and climate change, 910–11, 941
and competing generations, 917–24
and effects of policy, 915–17
and modeling of economic damage, 912–15
and sources of damage, 911–12
and psychology, 70
and public economics, 3
as study of scarcity, 26
Edgeworth box, 16, 18, 20, 36, 425, 426
contract curve in, 34
and feasible allocations of resources, 32
and Two Theorems of Welfare Economics, 33
Economic environment, 609
Economies of scale, external, 722
Economists, and need to understand actual decisions, 889
Economy(ies). See Competitive economies; Hidden economy; Overlapping generations economy
Education
as human capital investment, 824
and intergenerational inequality, 494–95
as measurable, 693
as redistribution, 434, 436
as signal, 316–21
as sorting process, 298
as transfer program, 626
Educational finance policy, 491–92
Effective average tax rate (EATR), 724
Efficiency, 119, 704
for club goods, 194, 199
and club membership, 207
and competition, 13, 26–27
in exchange economy, 33–37
Pareto-efficiency, 31–33, 436–41 (see also Pareto-efficiency)
and production and exchange economy, 37–39
with single consumer, 27–31
in consumption, 38
across countries, 704–709
vs. equity, 52, 122, 123
and commodity taxes, 522–23
and income tax, 538
and objective function, 607
Pareto-efficiency, 438–39
and socially optimal allocations, 427
and first- vs. second-best outcome, 27
and fiscal federalism, 636, 637, 640–43, 646, 653
as decentralization of government activities, 656, 657
in health care, 330–32
of income tax, 537
and intergovernmental transfers, 686
and local public goods, 209–11
Tiebout hypothesis on, 212–14, 216–17, 216, 639
of market
in international environment, 706–708
presumptions behind, 223
of overlapping generations model (testing of), 769
and prices, 14, 30
and private goods, 148, 150
in private vs. public sector, 665
in production, 38, 513–14, 524
and public goods, 154–56
private purchase of, 170
for pure public goods, 193
and repeated behavior, 178
and taxes
commodity taxes, 528–30, 531
and Ramsey rule, 521
and Ramsey rule modified, 521–23
tax harmonization, 725–26
Efficient production frontier, in single-consumer economy, 30–31
Efficient tax competition, 679
Effort
and equality of opportunity, 489, 491
as source of inequality, 608
Index

Emission Trading Scheme, EU, 929, 931–34, 936, 938–39, 940
Endogenous growth, 823–26, 838
and government expenditure, 826–29, 836, 837
and taxation, 833
Endowment(s), 15–16
and comparative advantage, 736
as lifetime labor supply, 41
as purchasing power, 20
Engel, Ernst, 462
England. See also United Kingdom
enclosure in (16th century), 149
gross revenue minus investment in, 769
Enrollment, automatic, 69
Entry deterrence, in oligopoly, 256
Entry into market
barriers to, 277–78
patents as, 278
threat of, 258
Environmental damage, as discouraged by high discount rate, 863
Envy, 78
Equality. See also Inequality
of opportunity, 489–91
measurement of, 491
as policy, 491–92
pursuit of, 423
Equality of Opportunity (Roemer), 489
Equalization formula, 640
Equalization grants, 689–90
Equally distributed equivalent income, 478–79
Equal sacrifice approach, 539
Equilibrium
asymmetric (rent-seeking game), 397
autarkic, 30
on the capital market, 784
in competitive economy (competitive equilibrium), 6, 13, 34–35, 424, 751
exchange economy, 17, 19
and overlapping generations model, 752, 770
Pareto-efficient allocation as, 884
and Pareto-preferred allocation, 762
production and exchange economy, 26
focus on, 42–43
general, 14
in model construction, 14
Nash, 66, 76, 152, 153, 175, 177 (see also Nash equilibrium)
in overlapping generations economy, 758
intertemporal equilibrium, 758, 759
steady-state equilibrium, 759–62, 766–67
partial, 14, 401
and policy analysis, 5
refinements of, 320
in single-consumer economy, 28, 29, 31
and social welfare, 18
subgame-perfect, 76, 77
and supply–demand equality, 21
Walrasian, 79
Equilibrium analysis, partial, 14, 401
Equilibrium prices, 17
Equity, 119, 122
and decentralized decision-making, 640
vs. efficiency, 52, 122, 123
and commodity taxes, 522–23
and income tax, 538
and objective function, 607
Pareto-efficiency, 438–39
and socially optimal allocations, 427
and fiscal federalism, 636
and inverse elasticity rule, 517
practical tools for, 457
redistribution required by, 607
and social welfare functions, 425, 457
in taxation (commodity), 521–23
Equivalence scales, for income, 458, 460–66
Equivalent variation, 880, 882
Euro countries
currency crisis in, 111
and debt, 107
and financial crisis, 108
Europe
climate change effects in, 911
public ownership in, 281
and social security transition model, 799–800
Index

European Commission
and cost–benefit analysis, 847
on formula apportionment, 715–16
and mergers, 283
European Steel and Coal Community, 731
European Union
agricultural subsidies in, 740
apportionment tax policy of, 725
cost–benefit analysis, 847
constraints on, 703
and Copenhagen Accord, 898, 925
Emission Trading Scheme (ETS) of, 929, 931–34, 936, 938–39, 940
Everything But Arms programme of, 740
European Union evolution of, 703
and fiscal federalism, 635
and Microsoft, 278
and origin system of taxation, 732
tariffs on processed food in, 738
as transfer scheme, 708–709
underprovision of social insurance in, 686
Excess capacity, as barrier to entry, 279
Excess demand, 20–22, 26
Excess government, 828–29
Excessive decentralization, 643
Exchange economy, 15–22
efficiency in, 33–37
Exogenous growth, 810, 838
and constant saving rate, 810–16
optimal taxation in, 817–23
Expected utility function, and tax evasion, 577, 579, 593, 594–95
Expected utility theory, 594
Expenditure-matching grants, 687
Explicit collusion, 260
Exponential discounting, 871
External economies of scale, 722
Externalities, 68, 223–25, 248
attribution, 791
capital outflows across jurisdictions as, 687
from climate change, 897
carbon emissions as, 937
Coase theorem on, 242–47
consumption, 224
in cost–benefit analysis, 851
in country’s tax increase, 672
and endogenous growth model, 825–26
examples of, 228
bandwagon effect, 235–36
pecuniary, 230–31
rat race problem, 232–33
river pollution, 228–29
traffic jams, 229–30
tragedy of the commons, 233–34
fiscal, 675, 694–95
vertical, 676
in inflating of insurance premiums, 329
vs. internalities, 68–69
internalization of, 241–42
and level of government, 637
licenses for, 239–41, 249
as market failure, 260
market inefficiency from, 225–28
network, 235
and nonconvexity, 247–48
pecuniary, 224, 230–31
price or quantity as instruments for, 937
production, 224
in terms of trade, 727
uncertainty about costs and benefits of, 240
Fairness, 78. See also Equity
False overstatement, and preference revelation, 165–66
False understatement, and preference revelation, 163–64
Family Expenditure Survey, US, 575
Farm (agricultural) subsidies
in developed countries, 738–39
and WTO agreement, 740
Feasible allocations of resources, 32–33
Feasible plan, 16
Federal countries, 94
Federalism, and common resource problem, 137–38. See also Fiscal federalism
Federal Old-Age and Survivors Insurance and Disability Insurance Trust Funds (OASDI), 780
Feedback, positive (in segregated groups), 495
Financial crisis, global, 107
Finland
social value of life in, 869
tax rate of, 691
Firms, 22
production sets of, 22–23
First-best outcome, 27
First-order stochastic dominance (FSD), 490
First Theorem of Welfare Economics, 35–37, 436, 609–10
assumptions in, 39
and overlapping generations economy, 762, 766
and single-consumer economy, 37
Fiscal competition, 694–95
and capital mobility, 666–67
and efficiency, 665
evidence for, 691–94
and “flypaper effect,” 690
with imperfect mobility, 682–85
with perfect mobility, 681–82
and race to the bottom, 672, 685–86, 691
and intergovernmental transfers and efficiency, 686–89
and redistribution, 689–90
tax competition, 665–67
and capital mobility, 666–67
efficient, 679–81
and public input provision, 674–76
and race to the top, 693
and size, 673–74
strategic behavior in, 667–72
tax exporting, 677–79
tax overlap, 676–77
Fiscal equalization, 689–90
Fiscal federalism, 635, 657
accountability in, 643–46
and competition for capital, 672
and evidence on decentralization determinants for, 656–57
by functions, 655
worldwide, 653–54
hard and soft budgets in, 650–56
and multi-level arguments, 656–37
and costs of uniformity, 637–39
distributive arguments for, 640
and Tiebout hypothesis, 639
optimal structure for, 640–43
and race to the bottom, 672
and risk-sharing, 646
and insurance vs. redistribution, 649–50
and secessionist regions, 657, 690
voluntary, 646–49
and tax overlap, 676–77
Fiscal policy, and growth, 836
Flat tax, voting over, 563
Flood Control Act (1936), 847
Flypaper effect, 690
Food Stamps, 688
Forecasting, economic, 915
Foreign direct investment (FDI), 736
and taxation, 724
Formal (legal) incidence of tax, 268
Formula apportionment, 715–16
Foster–Greer–Thorbecke (FGT) poverty measure, 487, 488
Framing bias, 60–62
Framing effect, 54
France
in Concorde project, 940
debt repudiation by (16th century), 107
emissions level of, 936, 937
government debt in, 109
gross revenue minus investment in, 769
health insurance in, 335
public provision of, 330
median-voter age increase in, 801
public spending in, 90–94
social value of life in, 869
Free-market principle, 864. See also Market(s)
Free-riding, 150, 174
and allocation mechanisms, 159
in fund-raising campaign, 178
and public goods, 174
in game, 175–76
with pure public goods, 212
Free trade, as Pareto-efficient, 736
and market failure, 736
Full comparability, 450
Fully funded system of social security, 777–78, 783, 797
and social security transition, 798, 800–801
Fund-raising campaigns, 178
and bargaining, 245
contribution campaign, 179–81
subscription campaign, 181–82
Future values, discounting of, 853–56, 863, 869–76
Gabaix–Laibson model, 67
Galileo, 896
Gambler’s fallacy, 58–59
Game theory, 791. See also Nash equilibrium; Prisoner’s dilemma game
Gender differences
in labor force participation rate, 545, 546
in tax evasion, 588
General Agreement on Tariffs and Trade (GATT), 737, 738, 740
General equilibrium analysis, 14
Generalized choice sets, 72
Geography of trade, 191
Germany
composition of general spending in, 95–97, 98
corporate tax rate in, 829
emissions level of, 936, 937
equalization grants in, 689
as federal country, 94
and financial crisis, 108, 109
gross revenue minus investment in, 769
median-voter age increase in, 801
poverty line in, 489
public provision of health insurance in, 330
public spending in, 90–94
tax rate of, 691
tax revenues in, 102, 103
Gevers, Louis, 228
Gini measure of inequality, 472–74, 476–77
in OECD study, 480, 481
in Sen measure of poverty, 486
Gini social welfare function, 477
Global financial crisis, 107
Globalization, 703, 709, 717, 741
and capital taxation, 710–11 (see also International taxation)
Global warming, 74–76
as externality, 223
and Kyoto Protocol on climate change, 80
science of, 900–903
controversies over, 904–907
on greenhouse gases and temperatures, 900–903
and method of reconstruction, 902, 906
predictions, 907–10
Golden Rule capital–labor ratio, 752, 762–65, 765, 769, 815–16
and social security, 776, 778, 782, 783, 784–86, 787, 787–88, 789
reform of, 800
Gorman, William M. “Terence,” 460
Government. See also Public sector
definition of, 113, 114
distortionary policy (rent-seeking) of, 400–401
as employment competitor, 138–39
function of, 4
as “leviathan,” 680
and local public goods, 193
monopoly position created by, 399–400
and public economics, 3 (see also Public economics)
rent-seeking model of, 404–408
Government debt, 107–12
and Ricardian equivalence, 797
Government expenditure, and endogenous growth model, 826–29, 836–37. See also Public expenditures
Government intervention. See also Taxation
for adverse selection, 304–307
and asymmetric information, 299–300
and Coase theorem, 242
for equity, 122–23
extent of influence of, 114
for externalities, 239
and health care, 330–33, 336
for imperfect competition, 288
regulation of monopoly, 277–82
regulation of oligopoly, 282–84
Government intervention (cont.)
and irrational people, 53
against moral hazard, 329–30
and market failure, 121–22
as motivated by mistakes, 52
non–tax distributions, 434–36
and overlapping generations economy, 762
and Pareto improvement, 610
and paternalism, 51
and public goods, 147, 154, 155–56, 182
and public sector pricing, 530–31
in screening (insurance), 313–14
and signaling, 321–22
on standard economic approach, 67, 68–69
for technological change on climate control, 940
universal programs necessary, 434, 626
Government optimization problem, 820–21
Grading, by relative standards, 232
Grants
expenditure-matching, 687
lump-sum, 690
revenue-matching, 687
Great Depression, Smoot–Hawley in, 736
Greece
debt–GDP ratio, 111
democracy and social security in, 798
and financial crisis, 108, 109
social value of life in, 869
Green Book, UK, 873–75
Greenhouse gases (GHGs), 895, 898, 900–901.
See also Climate change
in modeling of warming process, 904
and temperatures, 900–903
Gross domestic product, nominal, 113
Growth. See Economic growth
Guessing game, 65–66
Gulf Stream, and UK temperature, 912
Haig–Simons approach to taxation, 620, 625
Hamilton, Alexander, 371
Hamilton apportionment, 372
Hard-budget constraint, 689
Harmonization of taxes, 725
international, 725–29
Hazard insurance, 302–305
Hazard rate (Pareto distribution), 561
Head-count ratio, 484
Health care
public provision of, 330–33
as redistribution, 434–35
Health maintenance organization, 332
Hearth tax, 505
Hedonic prices, 859–60
Herfindahl index, 259
Hicks, John R., 459
Hicks’s test, 878
Hidden action, 300–302
Hidden economy, 574–76
Hidden knowledge, 300–302
Hindsight bias, 59
Historical development of public sector, 89–94
Hobbes, Thomas, 120
Hockey stick plot, 903, 906–907
Holding company of multinational, and taxation, 714–15
Honesty, and tax evasion, 580, 588–90
Horizontal product differentiation, 257
Hotelling, Harold, 372
Hotelling model, 372–73, 375
Hotelling’s principle of minimal differentiation, 353–54
Household debt, 109
Household utility function, 464–65, 466, 478, 479
and equivalence scales, 496
Household utility of income function, 478
Housing, low-income, 627
Human capital
and credit constraint, 494
as input, 824
production process for, 831, 838
Human life, value of (cost–benefit analysis), 863, 865–69
IAM (Integrated Assessment Model), 912–13, 924
Identity, imperfect decision-making from, 63–64
Identity economics, 63
IMF, Government Finance Statistics of, 655
Immigration policy, 193
Impact assessment, 860–61
Imperfect competition, 255–57, 260
inefficiency from, 260–63
through incomplete information, 263–64
measures of welfare loss from, 264–67
and market structure, 257–60
policy response to, 277, 278
and tax incidence, 268–74
specific vs. ad valorem taxation, 274–77
and tax rules, 528–30
and unions, 286–88
Imperfect information. See Asymmetric information, 584–85
Imperfect information equilibrium, 310–13
Implementation problem, 610
Impossibility Theorem. See Arrow’s Impossibility Theorem
Incentive-compatible allocations, 616–17
Incentive-compatible allocation mechanisms, and truthful revelation, 608–609
Incentive compatible taxes, 429–30
Incentive incompatible taxes, 430, 555–56, 558, 608
lump-sum taxes as, 433
Incentives in contracts, 323
and health care behavior, 334–35
and tax structure, 538
Income change in vs. levels of, 73
from interest, 622, 818, 822
measurement of, 458–60, 496
equivalence scales for, 460–66
Income distribution invariance, 172
Income distribution across jurisdictions, 681 and “flypaper effect,” 690
with imperfect mobility, 682–85
with perfect mobility, 681–82
and race to the bottom, 685–86
evidence on, 691
Income effect, 508, 541, 544, 545
and lump-sum transfers, 40
and wage increase, 58
Income gap ratio, 485
Income level, and public expenditures, 125
Income taxes, 89, 537. See also Taxation
capital, 622, 624–25, 666–67, 823
and mobility of labor, 685
vs. consumption or commodity tax, 620–24, 627
and direct vs. indirect revelation mechanisms, 612
and incentives, 538
and interjurisdictional competition, 665
and labor supply, 539–44
empirical evidence on, 544–46
vs. lump-sum, 428
non-uniform marginal rates for, 413
optimal, 547–54, 563
numerical analysis of, 561–63
quasi-linearity in, 555–58
Rawlsian 559–61
and progressivity, 554, 836
and redistribution, 609
and savings, 621
as second-best solution, 434, 547, 614
for selected countries, 101
in United Kingdom (1799), 537
voting on (flat tax), 563–65
Income transfers. See Redistribution; Transfers
Incomplete information, 263–64
India
and Copenhagen Accord, 898, 935
and Doha talks, 740
emissions level of, 935–36, 937
Indian National Sample Survey, and optimal tax rate, 526–28
Indirect revelation mechanism, 612
Inefficiency from commodity taxation, 505
of imperfect competition, 260–63
through incomplete information, 263–64
measures of welfare loss from, 264–67
from monopoly, 261, 265–67, 398
of rent-seeking, 387
Index

Inequality
acceptable vs. unacceptable, 489
effort vs. luck as cause of, 608
and equity, 457
intergenerational transmission of, 492–93
causal mechanisms of, 493–96
and interregional risk-sharing, 649
measurement of, 457, 458–59, 466
in applications, 480–81
income data in, 457–58
intuitive vs. specific, 496
and social welfare function, 475–80
statistical measures of, 467–75 (see also Gini
measure of inequality; Lorenz curve)
and welfare assumptions, 458
Inflation, from printing money, 107
Information (knowledge)
and Arrow’s Impossibility Theorem, 446
asymmetric, 281, 297–99, 334 (see also
Asymmetric information)
in bargaining, 245–46
behavior as conveying, 297, 298–99
on climate change, 939
and government intervention for market
failure, 122
incomplete, 263–64
irreversibility of, 59
lack of, 52
and excessive government, 135
and licenses, 239–40
and Lindahl equilibrium, 147, 238
and lump-sum taxation, 434
and modeling of bureaucratic efficiency, 131
and monopoly, 263
and optimal allocation mechanism, 609
and Pigouian taxation, 238, 239, 240
private, 429
and public ownership, 281
and redistribution, 610
and social welfare function, 450
symmetric, 14–15, 42
and taxation, 614
Informational rent, 136–37
Information equilibrium, perfect, 308–309
Information about preferences, revelation of.
See at Preference revelation
Informative lobbying, 408–12
Infrastructure expenditure, 124
In-kind transfers, 435–36, 627
Innovation
and monopoly, 267
and patents, 278
Inquisition, and Galileo, 896
Inspection effect, 129
Insurance
behavior and information in market for, 298
compulsory, 305
hazard, 302–305
and hidden knowledge, 300
moral hazard in, 323–24, 330
with effort observable, 325–26
with effort unobservable, 326–27
government intervention for, 329–30
and second-best contract, 327–29
vs. redistribution (interregional), 649
and screening, 307–13
and social security, 775
Integrated Assessment Model (IAM), 912–13,
924
Intellectual property rights, 740
Interest income, tax treatment of, 622,
818, 822
Interest rate
in discounting future, 863 (see also Discount
rate)
in intertemporal arbitrage, 872
and overlapping generations economy, 755
and Golden Rule, 764, 769
Intergenerational equity, and discounting, 874
Intergenerational inequality, 492–93
causal mechanisms of, 493–96
Intergenerational transfer of resources 778
social security program as, 782–83
Intergovernmental transfers
and efficiency, 686–89
and redistribution, 689–90
Intergovernment Panel on Climate Change
(IPCC), 916
Intermediate goods and free trade, 736–37
and taxation, 513–14, 531
Internalities, 52
vs. externalities, 68–69
Internalization, 241–42
Internal rate of return (IRR), 854–55
Internal Revenue Service, studies of taxpayer compliance by, 575
International agreements, on climate change, 934
International taxation, 703–704, 938
and capital taxes, 709–11
corporate, 711–15
and efficiency, 704–709
and formula apportionment, 715–16
and harmonization of taxes, 725–29
and location choice, 720–25
and agglomeration rents, 722–23
and mobility, 703, 707, 709, 710–711, 714–15, 717, 841
principles of, 729–32
and tariff policy, 732–33
and welfare costs of tariffs, 733–36
and trade agreements, 736–41
and transfer pricing, 703–704, 717–20
International Trade Organization, 737
Interpersonal comparability, 443–46
Intertemporal arbitrage. 870–72
Intertemporal economy, 751. See also Overlapping generations economy
Intertemporal equilibrium 758, 759
Intertemporal preferences, 74, 888
Intrinsic motivation, 64
Inverse elasticity pricing rule, 262, 288, 516–17, 521, 725
Investment, in education, 316
Investment, and monopoly, 267
Invisible hand, 13
Ireland
corporate tax rate in, 829
debt–GDP ratio in, 111
and financial crisis, 108, 109
reduced corporation taxes in, 686
social value of life in, 869
tax rates of, 691
Irrational behavior, and policy interventions, 53
Isoprofit curves, 23, 24, 31
Italy
criminal activity in national accounts of, 574
democracy and social security in, 798
emissions level of, 936, 937
and financial crisis, 108, 109
governmental decentralization in, 657
government debt in, 109
gross revenue minus investment in, 769
inequality in, 480
median-voter age increase in, 801
public provision of health insurance in, 330
secessionism in, 657, 690
social value of life in, 869
Item-by-item voting, 357
Japan
emission level of, 936, 937
government debt in, 109
gross revenue minus investment in, 769
hidden economy of, 575–76
inequality in, 480–81
poverty line in, 489
public spending in, 90–94
social value of life in, 869
tariffs on processed food in, 738
tax revenues in, 99–100, 103
Jefferson, Thomas, 371
Job training, as in-kind redistribution, 436
Kahneman, Daniel, 64
Kaldor approach to taxation, 620
Kaldor’s test, 878, 881
Keyboard, arrangement of keys on, 235
Kleenex, in merger, 284
Korea
emissions level of, 936, 937
public sector expenditure in, 92–93
tax revenues in, 102
Kyoto gases, 901
Kyoto Protocol on climate change (1997), 80, 898, 988–99, 935, 939
Labor markets
asymmetry of information in, 42
Labor markets (cont.)
equilibrium on, 405
monopsony in, 286–88
and valuation of life, 867
Labor supply, and taxation, 538,
539–44
differential commodity taxes, 622
empirical evidence on, 544–46
Laissez-faire policy, 5
Leisure, and income, 539–44
Lerner index, 262
for breakfast cereals, 282
and oligopoly, 263
Level comparability, 447, 448, 449, 450
ordinal, 445, 447
Levels of government, 635. See also Fiscal federalism
Libertarians, 51
Licenses, for externalities, 239–41, 249
Life, valuing of (cost-benefit analysis), 863, 865–69
Life-cycle model
and income–consumption equivalence, 620
and overlapping generations model, 754
Lindahl equilibrium, 147, 160, 182, 213, 238, 610
Lindahl reaction function, 160–61, 162
Lobbying and lobbyists, 387,
404–406
and equilibrium effects of rent-seeking, 401
informative, 408–12
in rent creation, 407
as rent-seeking, 387
in rent-seeking game, 390
Local public goods, 191, 192–93, 208–11, 216
and club goods, 193
Tiebout hypothesis on, 192, 212–14
empirical tests of, 214–16
Location choice, and international taxation,
720–25
and agglomeration rents, 722–23
Lorenz curve, 471–72, 496
and Atkinson theorem, 476
and Gini measure, 474
Loss aversion, 58, 73

Lump-sum transfers, 39–41, 628

Majority voting, 350, 441. See also Voting agenda manipulation in, 358–61
and Condorcet winner, 351
and May’s theorem, 350–51
and Median Voter Theorems, 351–56
multidimensional, 356–58
Maldives, and climate change, 911
Mandates, enforcement of at private level, 330
Marginal rate of substitution (MRS), 29
and club goods, 195
and production efficiency, 513–14
and public goods, 154
and Median Voter Theorem, 159
Marginal rate of transformation, 29
and club goods, 195
for private goods, 155
Marginal reforms, of tax rates, 524–26
Marginal revenue benefit, of taxation, 525
Marginal utilities of income, in compensation tests, 881
Marginal valuations
and public vs. private goods, 148
quality of, 34
Market(s)
and allocation of individuals to professions, 231
behavioral, 64–65
complementary mistakes in, 65–66
money pump in, 65
and rationality tug-of-war, 67
definition of, 257–58
and efficiency (in international environment), 706–708
vs. government, 665, 694, 695

Luck, as source of inequality, 608
Lump-sum subsidies or grants, 688, 690
Lump-sum taxes, 156, 424, 426, 427–30, 506
and commodity tax, 508–509, 512
and competitive market, 610
faults of, 428
optimal, 614
practical impossibility of, 430–34
and tax on capital income, 823

Maldives, and climate change, 911
and other-regarding behavior, 78–80
and problem of multiple stakeholders, 139
and Tiebout hypothesis, 212, 214
Market capture
by public sector, 134
from pure public goods, 212
Market failure (inefficiency), 42, 121–22, 225–28
and climate control, 940
and free trade, 736
imperfect competition as, 260 (see also Imperfect competition)
from moral hazard, 329–30
from public goods, 147
and public intervention, 51
Market power, 256
attempts to increase, 255
and breakfast cereal industry, 282–83
Market prices, vs. shadow prices, 883–88
Market share, vs. market size, 258
Market structure, and imperfect competition, 257–60
Market unraveling, 304, 307
Markup, oligopoly, 263
Marshall, Alfred, 848
Marshallian surplus, 858–59
Matching grant programs, 687–88
May’s theorem, 350–51
Meade’s formula for cost–benefit analysis, 883
Measurement
backward- vs. forward-looking (income), 458, 460
in climate change controversies, 905
common unit of needed (CBA), 856–57, 865, 868
of competition, 258–60
of education, 693
of equality of opportunity, 491
of income, 458–60, 496
equivalence scales for, 460–66
of inequality, 457, 458–59, 466–67
in applications, 480–81
income data in, 457–58
intuitive vs. specific, 496
and social welfare function, 475–80
statistical measures, 467–75 (see also Gini measure of inequality; Lorenz curve)
and welfare assumptions, 458
of poverty, 457, 458–59, 483–87, 497
applications of, 488–89
required properties of, 485–86
of public expenditures, 112–14
of social costs and benefits (calculation), 851–52
of social welfare, 5–6
of tax evasion, 574–75
and utility, 436–37, 443
of welfare loss from monopoly (imperfect competition), 264–67
Mechanism, allocation. See Allocation mechanism
Mechanism design, 611, 612
Median voter, 128
Median Voter Theorems, 157–59, 351–56
and flat tax voting, 564–65
and multidimensional voting, 356–57
Medicaid, 332, 688
Medical savings accounts (MSA), 332
Mental accounts, 65
Merger, determining price effect of, 283–84
Merger policy, 283–84
Microsoft, knowledge-sharing forced on, 278
Migration, and local public goods, 210–11
Mill, John Stuart, 3, 539
Minimal state, 119–21
Minimum needs, 460–66
and poverty line, 483
Mistakes complementary, 65–66
and preferences, 71
MITI (Ministry of International Trade and Industries), Japan, 278
Mixed economy, 4, 607
Mixed strategies, 391
Mobility of capital
across countries, 709, 710–11, 717, 741
and fiscal competition, 666–67
and mobility of labor, 685
of consumers of public services, 695
imperfect, 682–85
perfect, 681–82
of firms (and rent extraction), 724–25
of labor, 685, 741
of multinational holding company,
714–15
social, 492
of tax base (international taxation), 703, 707,
709, 710–11, 714–15, 717, 741
Model(s), 3–4, 13–14
and behaviorist vision, 53
of competition, 13
and spurious explanation, 129
Modeling
of economic damage, 912–15
structural, 73
Money, as unit of account, 20
Money pump, 65
Monopolist, 256, 264
Monopolistic competition, 257
Monopoly, 255
inefficiency of, 261, 265–67, 398
international, 708
natural, 279–82
product differentiation under, 262
regulation of, 277–82
social cost of, 398–403
tax incidence in, 270–72
and specific vs. ad valorem taxes,
275
Monopoly power, 256
of public sector, 133–34
Monopsonist, 256
Monopsony, 286
in labor markets, 255, 286–88
Monopsony power, 256
Montreal Protocol, 901
Moral hazard, 299, 301, 323–24
vs. adverse selection, 334, 336
and antipoverty programs,
302
in fiscal federalism, 653
in insurance, 323–24, 330
with effort observable, 325–26
with effort unobservable,
326–27
government intervention for,
329–30
for health care, 330, 331, 332
policy recommendations on,
334–35
and second-best contract, 327–29
and Volvo drivers, 301–302
Mortgage obligations, securitized, 107
Motivation, intrinsic, 64
Multi-level government, 636–37. See also
Fiscal federalism
and costs of uniformity, 637–39
distributive arguments for, 640
and Tiebout hypothesis, 639
Naïve individual, 56, 67
Nash assumption, 669
Nash bargaining solution, 245
Nash behavior, and private purchase of public
goods, 176
Nash equilibrium, 66, 76, 152, 391,
791–92
in preference revelation, 165
and public goods, 150–51, 153, 175, 177
private purchase of, 176, 177
in rent-seeking game, 391, 394–95, 396
and social security transfers, 791, 793
in tax competition, 670, 671
and truth, 614
Nash reaction function, 151–52
Nash strategies, 792
National defense, 120
as public good, 147
NATO, as club, 191
Natural monopoly, 279–82
Nature-nurture relationship, and inequality,
493–94
Net benefit, 166
Net consumption, 28
Net present value (NPV), 853, 855, 856, 861,
870, 883, 886
Network externality, 235
Index

Neutrality, policy, 173
Neutrality properties of taxes, 710
New Economic Geography, 722
Newtonian mechanics, as superseded, 896
Noncompatability, 445, 448, 449
Nonconvexity, and externalities, 247–48
Nondiscriminatory policies, 413
Nondiscriminatory motives, 53
Nonexcludability, of public goods, 148
Non–fully funded social security system, 778
Nonmarket goods, valuing of (CBA), 859–60
Nonpecuniary factors, and intergenerational transmission of inequality, 495
Non–price competition, 257
Nonrivalry, of public goods, 148, 191
Non-tax redistribution, 626–28
Nordhaus, William, 913, 920, 924
Nordic countries. See also Denmark; Sweden
capital and labor income taxes in, 625
cost–benefit analysis in, 847
social mobility in, 492, 494–95
Normative policy analysis, 5
Nuclear power station, cost–benefit analysis of, 863–64

Oates, Wallace E., 214–15
Objective function, 281, 607
OECD
corporate tax rate in, 829
cost–benefit analysis in, 847
equality study of, 480
farm subsidies in, 739
and income tax–GDP relationship, 835–36
insurance for health care in, 330
poverty study of, 488–89
profit-shifting in, 723
and tariffs on imports from developing countries, 739
tax competition report of, 691
Oligopoly, 262
regulation of, 282–84
and tax incidence, 272–74
Oligopoly markets, and club membership, 207
Oligopoly markup, 263
Opportunity, equality of, 489–91
measurement of, 491
as policy, 491–92
Opportunity arbitrage, 871
Opportunity-cost bias, 69
Optimal allocation, 611, 617–18, 622
Optimal allocation mechanism, 609, 622
Optimal capital–labor ratio, 762–65, 766. See also Golden Rule capital–labor ratio
Optimality, social, 424–27
Optimal policy analysis, 4
Optimal redistribution, 617
Optimal risk-sharing arrangements, 647
Optimal structure of federal system, 640–43
Optimal taxation
commodity, 505, 509–12, 513
and production efficiency, 513–14
and redistribution, 526–28
discrimination in, 413
in exogenous growth model, 817–23
income tax, 547–54, 563
numerical analysis of, 561–63
quasi-linearity in, 555–58
Rawlsian, 559–61
lump-sum, 428, 614
Optimism over cost, 855
Ordinal level comparability, 445, 447
Ordinal utility, 443, 444
Origin principle, of commodity taxation, 672, 730
Orshansky, Mollie, 461, 462
Osborne–Slivinski model, 379
Other-regarding preferences, 76
market impact of, 78–80
social preferences, 77–78
in ultimatum game, 76–77
Overconfidence, 60
Overlapping generations economy, 752, 753–54, 770
and birth rate, 776, 778, 797
consumers in, 754–55
Overlapping generations economy (cont.)
equilibrium in, 758
   intertemporal equilibrium, 759
   steady-state equilibrium, 759–62, 766–67
and Golden Rule capital–labor ratio, 762–65, 765, 766
and Pareto-efficiency, 752, 765–69
production in, 755–58
and social security for retired, 775
testing efficiency of, 769
Overoptimism, 60
Overshifting, tax, 271, 529, 530
Ozone layer destruction, as externality, 223

Panel on Climate Change Third Assessment Report, 902
Pareto, Vilfredo, 71, 436
Pareto criterion, 348, 441
   for project evaluation, 877
Pareto distribution of income, 561
Pareto-efficiency, 31–33, 436–41
deficiency of, 423, 438–39
and First Theorem of Welfare Economics, 33, 35–57, 609–10
and overlapping generations economy, 752, 765–69
and policy-making, 424
social comparison in, 444
and social welfare function, 457
and taxation of goods, 704
and Two Theorems of Welfare Economics, 884
Pareto-efficient steady states, 752
Pareto improvement, 437
in social security system, 782–87
Pareto preference, 437
Pareto-preferred location, 33–34
Partial Dissipation Theorem, 396
Partial equilibrium analysis, 14, 401
Partial revelation of information, by lobbyist, 411
Participation decision (labor supply), 543, 544
Patents, as barrier to entry, 278
Paternalism, 51, 72
Payoff, from preference revelation mechanism, 164, 165
Payoff functions, in tax-evasion decision, 594
Payroll taxes, 102
Pay-as-you-go social security program, 776–77, 783, 790, 795
and dependency ratio, 797
and population growth, 789
and social security transition, 798
Peak-load pricing, 280–81
Pecuniary externality, 224, 230–31
Pension(s), 775
increased expenditures on, 91–92
   as redistribution, 434
   and time-consistency, 436
tax deductions for, 625
and workers’ inadequate provision for retirement, 797
Pension crisis, 775, 778–81
Perfect information equilibrium, 308–309
Perfect price discrimination, 263, 264
Performance-enhancing drugs, 232–37
Personalized prices
   and externalities, 238
   for public goods, 159–63, 213
Personalized pricing model, 609
Peru, democracy and social security in, 798
Pessimism over cost, 855
Pigou–Dalton Principle of Transfers, 469–71, 472, 474, 475
Pinochet, Augusto, 798
PISA (Program for International Student Assessment), 63
Pivotal player, 168
Pivotal-voter theory, 370
Plurality voting, 363
Polar ice cap, melting of, 909
Policy analysis, 5
   assumptions of, 6
   behavioral, 67–68
   automatic enrollment, 69
   as complementary to substantive policy, 70
and internalities vs. externalities, 68–69
SMaRT Plan, 69–70
and climate change, 899–900, 907, 915–17, 929–40, 941
and future generations, 917–24, 926
normative, 5
optimal, 4
positive, 5
and Second Theorem of Welfare Economics, 424
Policy divergence, 373–75
Policy neutrality, 173
Political competition, 372. See also Voting citizen-candidate model of, 379–80
Downsian model of, 372–73
manipulation of debt figures in, 113
multidimensional, 376–77
and policy divergence, 373–75
and “swing voter” politics, 377
Political structures, and expenditure level, 94
Politicians
rent creation by, 406–407
and rent-seeking, 404
Pollution, river, 228–29
Pooling equilibria, 299, 313, 320, 321
Population growth, and social security system, 787–90
Population size, and public goods, 173–74
Portugal, democracy and social security in, 798
Positional social welfare function, 447
Positive feedback, and segregated groups, 495
Positive policy analysis, 5
Positive responsiveness, 364–65
Poverty, 481–82
absolute, 482
and equity, 457
intuitive vs. specific measures of, 496
measurement of, 457, 458–59, 483–87, 497
applications of, 488–89
proposed properties of, 485–86
and poverty line, 482–83, 497
relative, 482
Poverty gap, aggregate, 484–85
Poverty scale, US, 461–62
Power equalization, 689
Precautionary principle, 864
Precommitment, 55, 75
Prediction
on climate change, 907–10
“out-of-sample,” 899
Preference aggregation. See also Arrow’s Impossibility Theorem
determination of, 27
and social welfare function, 450
Preference revelation, and local public goods, 212
Preference revelation axioms, 71
Preference revelation mechanisms, 163–66, 169–70, 182
Clarke–Groves mechanism, 166–68
Clarke tax, 168–69
Preferences, 22
and allocation mechanism, 610
over allocations of consumption and leisure, 539
collective, 347
over commodities, 622, 623
in competitive equilibrium, 762
and design of lump-sum taxes, 429
and identical-consumers assumption, 123
intertemporal, 74, 888
interpersonal similarities in, 444
intransitivities in, 65
of lobbyist vs. policy maker, 412
of median voter, 157–59
and mistakes, 71
in optimal tax determination, 510
other-regarding, 76
market impact of, 78–80
social preferences, 77–78
in ultimatum game, 76–77
and public-goods prices, 162–63
revealed, 857
in cost–benefit analysis, 857
revealed vs. “true,” 52
of single consumer, 27
single-peaked, 352, 353, 356, 373
social, 441
and risk, 864
Preferences (cont.)
and socially optimal allocation, 426
of unions, 285
and voting agenda, 358
and welfare effect, 71
Preference refinement, 73
Preference sorting effect, 656
Present bias, 54–55, 56, 69
Price control, on greenhouse gas emissions, 937–39
Price discrimination, perfect, 263, 264
Price normalization, 20
Prices
and competitive behavior, 42
in competitive economies, 14
in exchange economy, 15, 17–20, 22
in production and exchange economy, 26
and demand curve, 858
efficiency through, 14, 30
equilibrium prices, 17
and imperfect competition, 255
and “invalid” choices, 68
and isoprofit curves, 24
market vs. shadow, 883–88
and mergers, 283–84
under monopoly, 261–62
and oligopoly markets, 256
personalized (public goods), 159–63, 213, 238, 609
public sector, 530–31
and taxation, 268–70
transfer pricing (between divisions within firm), 703–704, 713–14, 717–20
Price wars, 260
Principles of Economics (Marshall), 848
Principles of Political Economy (Mill), 529
Prisoner’s dilemma game
emission reduction process as, 898–99
rat race problem as, 232
and social security model, 792
Private good, 148, 149
efficiency in consumption of, 154
in preference revelation, 169–70
Private information, and lump-sum taxes, 429
Private ownership economy, 25
Probabilistic rent-seeking game, 394–96
Probability density functions (pdfs), 907
Product differentiation, 256–57
for breakfast cereal, 282–83
by monopoly, 262
Production
in overlapping generations economy, 755–58
and social security, 783–87
Production efficiency, 38, 513–14, 524
Production and exchange economy, 22–26
and Two Theories of Welfare Economics, 37–39
Production externality, 224
Production function
Cobb–Douglas, 767, 822, 913
in endogenous growth model, 824, 825, 826
in exogenous growth model, 810, 811, 812, 813, 814, 816
and Golden Rule condition, 762, 815
and overlapping generations model, 757, 758
aggregate, 755–55
and Pareto-efficiency, 765
Production set, 22
Profit, with ad valorem vs. specific taxes, 276
Profit maximization
in production and exchange economy, 23–24
in Robinson Crusoe economy, 27
in single-consumer economy, 30
Profit-seeking, 388–389
Profit-shifting, international, 723
Progressive income tax, 554, 834
Project
defining of, 851
scope of, 847
Property rights, 119–20
and Coase theorem, 243–47
in commons, 149
Property taxes, 102, 106, 694
and Tiebout hypothesis, 214–15
Prospect theory, 61
and tax evasion, 595–96
Psychology, and economics, 70
and behavioral economics, 51
Ptolemaic system, as superseded consensus, 896
Public economics, 3
as how and what (what should be done), 3
institutional setting for, 4
models in, 3–4
and time, 751, 857 (see also Overlapping generations economy; Time consistency)
tradition of, 3
Public expenditures
for five developed economies, 89–94
and growth, 826–29, 836–37
and income level, 125
Public goods, 147, 148–49
allocation mechanisms for
personalized prices, 159–63
preference-manipulation of, 162–63
preference revelation mechanisms for, 163–70
Samuelson rule, 154–55
voting, 156–59
congestion in, 148, 191, 195
efficiency of
and majority voting, 158–59
through Samuelson rule, 154–55
and government intervention, 155–56
impure, 148
as inputs, 825, 826
and level of government, 636–37, 638, 639, 640, 641, 642–43
local, 191, 192–93, 208–11, 216 (see also Local public goods)
nonrivalry of, 148, 191
in political model of public sector growth, 126–28
private-purchase model of, 150–54
experimental evidence on, 174–76
and fund-raising campaigns, 178–82
modification of, 176–77
and population size, 173–74
and redistribution, 170–73
pure, 148, 192–93, 195, 216
spillovers of, 665
and tax competition, 674
Public intervention. See Government intervention
Public ownership, 280–81
Public sector. See also at Government composition of expenditures in, 94–97, 98–100
and corruption, 134–35
as excessive, 129–30, 138–39
and bureaucracy, 130–32, 138
and bureaucratic budget-setting, 132–33
from cost diffusion, 137–38
and lack of information by voters, 135–37
from monopoly power, 133–34
government debt, 107–12
growth of, 123
Baumol’s law on, 125–26
developmental models of, 123–24
political model of, 126–28
ratchet effect in, 128–29
Wagner’s law on, 124–25
historical development of, 89–94
justification for, 119
for equity, 122–23
from market failure, 121–22
and minimal state, 119–21
measurement of expenditure of, 112–14
revenue of, 97–106
in social-market vs. free-market economies, 89
Public sector pricing, 530–31
Punishment
and incentives for crimes, 585
tax evaders, 583–86
“Punishment strategy,” for social security game, 793–94
Pure public good, 148, 192–93, 195
and club goods, 216
Pure strategies, 391
Quality-adjusted life year (QALY), 867, 869
Quasi-hyperbolic utility, 56
Quasi-linearity, of utility function, 555–58
Qwerty keyboard, 235
Race to the bottom, 672, 685–86
evidence on, 691–91
Race to the top, 693
Radiative forcing, 901
Ramsey growth model, 817, 818, 919
Ramsey price, 280–81, 531
Ramsey problem of efficient taxation, 505
Ramsey rule, 517–21, 622
and equity considerations, 521–23
Ramsey rule for intertemporal consumption, 920, 929
Rand Health Insurance Experiment, 334
Rank-dependent expected utility, 594
Ratchet effect, 128–29
Rational choice models, inapplicability of, 54
Rationality, limited or imperfect, 51, 52, 53–54
Rationality tug-of-war, 67
Rat race problem, 232–33
Rawlsian philosophy, 440–41
Rawlsian (maxi-min) social welfare function, 425–26, 447, 538, 559
Rawlsian taxation, 559–61
Redistribution, 128
choice of allocation mechanisms for, 607–608
and decentralization, 654
and health care, 332–33
incentive constraints on, 432
and income distribution across jurisdictions, 682, 684–85
through income tax, 537
on capital, 625
vs. insurance (interregional), 649
and intergovernmental transfers, 689
limits of, 27, 616, 626, 628
and private information, 617
and loss aversion, 73
and luck vs. effort, 608
lump-sum, 610
non–tax, 434–36, 626–28
optimal, 617
and political influence of wealthy, 628–29
and private purchase of public good, 170–73
and public intervention, 51
and Second Theorem of Welfare Economics, 37
and social welfare function, 457
success of (OECD study), 480–81
and taxation, 609, 619–20
commodity taxes, 527, 528
income taxes, 561, 609
lump-sum taxes, 39–41, 424, 428, 430, 432–34
OECD study on, 480–81
retributive vs. incentive effects of, 608
vs. wealth incentives, 608
Reference-dependence bias, 57–58
Reform
of social security system, 797–802
of tax rates, 524–26
Regulation
corrupt evasion of, 134–35
merger policy, 283–84
of monopoly, 277–82
of oligopoly, 282–84
collusion detection, 282–83
Relative index, 468
Relative mean deviation, 468–69
Relative poverty, 482
Renewable Obligation (RO), 929
Rent(s)
agglomeration, 722–23
informational, 136–37
of politicians, 644–45
Rent-creation, 278, 406–408, 413
Rent-seeking, 7, 134, 387–90, 414
and climate science, 941
control of, 413
games of, 390–98
with free entry, 396–97
and risk aversion, 397–98
by government, 404–408
monopoly as, 398–403 (see also Monopoly)
vs. profit-seeking, 389
Residence-based taxation (corporate), 713
Residence-based taxation (personal), 713
Retirement, as certainty, 775
Revealed preference relation, 73
Revealed preferences, in cost–benefit analysis, 857
Revelation of information, by lobbyist, 411
Revelation mechanism, 613–14
direct, 611–12
indirect, 612
Revenue
and minimal state, 120–21
for public sector, 97–106 (see also Taxation)
revenue-matching grants, 687
Reversal paradox, Scitovsky, 879
Ricardian equivalence, 776, 778, 794–97
Ricardo, David, 3
Rietz, Thomas, 876
Risk
in cost–benefit analysis, 864–65, 875–76
in private investment funds for social security, 802
Risk aversion
and cost–benefit analysis, 864
in valuation of life, 868
in rent-seeking game, 397–98
Risk framing, 60
Risk-sharing, interregional (mutual insurance), 646
and insurance vs. redistribution, 649–50
and secessionist regions, 657
voluntary, 646–49
River pollution, 228–29
Road Transport Fuel Obligation (RTFO), 930
Robinson Crusoe economy, 27–28, 31, 37, 38, 509
Roemer, John, 489
Romer, Thomas, 563
Rowntree, Seebohm, 460, 461, 484
Rubinstein, Ariel, 245
Runoff voting, 364–65, 372
Sales tax, 624. See also Consumption taxes
Samaritan’s dilemma, 331
Samuelson rule, 147, 154–55, 161, 182, 195, 209, 687
Saudi Arabia, emission level of, 935, 936, 937
Saving rate, constant, and exogenous growth, 810–16
Savings
and income tax, 621
on capital, 624–25
and limited rationality, 53–54
and social security reform, 800
SMarT Plan for, 69–70
Scarcity, economics as study of, 26
Schelling, T., 228
Scholes, Christopher, 235
School systems, and intergenerational inequality, 494–95
Schumer, Chuck, 282
Schumpeter, Joseph Alois, 267, 825
Science, of global warming, 900–903
controversies over, 904–907
on greenhouse gases and temperatures, 900–903
and method of reconstruction, 902, 906
predictions, 907–10
Scitovsky reversal paradox, 879
Scitovsky test, 878
Screening, 298, 307–308
with imperfect information equilibrium, 310–13
with perfect information equilibrium, 308–309
SCV (squared coefficient of variation), 480–81
Sea levels, rise of, 909
Secession, threat of, 656–57
Second-best contract, 327–29
Second-best outcome, 27
Second-order stochastic dominance, 490
Second Theorem of Welfare Economics, 36, 37, 423, 426, 434, 436, 610
assumptions in, 39
and lump-sum taxes or transfers, 39, 427, 428, 433–34
and policy analysis, 424
and single-consumer economy, 37
and social welfare function, 451
Securitized mortgage obligations, 107
Segregation, and group inequality, 495–96
Self-control, (β, δ) model of, 55–57
Self-control problem, 55
Self-discipline, lack of, 52
Self-interest
assumption of, 6
deviations from, 76
limited, 51
Selfishness, 77
Social welfare function(s), 426, 440, 457
and Arrow’s Impossibility Theorem, 442
Bergson–Samuelson, 425
consistency of, 446
and equity, 457
Gini, 477
and inequality, 475–80
for optimal income tax, 561
positional, 447
Rawlsian (mini-max), 425–26, 447, 538, 559
and tax evasion, 584–85
and tax policy, 522
utilitarian, 425–26, 479, 538
and weighted welfare function, 557–58
Social welfare weights, and optimal tax rule, 528
Solow growth model, 810
Sophisticated agent or consumer, 56–57, 67
Sophisticated outcomes (voting), 358
Source-based taxation, 713
South Africa
and Copenhagen Accord, 898, 935
emissions level of, 936, 937
Sovereign debt default, 107
Soviet Union, and incentives, 432
Spain
debt–GDP ratio in, 111
democracy and social security in, 798
emissions level of, 936, 937
government debt in, 109
median-voter age increase in, 801
social value of life in, 869
Specialists,
and market capture, 134
Specific taxation, 274–77
Squared coefficient of variation (SCV), 480–81
Stability, and fiscal federalism, 640–43
Stability and Growth Pact (euro area), 111
Stabilization, of GHGs, 907–908
Stakeholders, in cost–benefit analysis, 862
Standard industry classification, 258
Standard Oil Company, monopoly action against, 277
State of nature, Hobbesian, 120
Static analysis, 267, 399
Statistical measures, of inequality, 467–75
Status quo bias, 70
Steady-state capital–labor ratio, 812–13. See also Golden Rule capital–labor ratio
Steady-state equilibrium, 759–62, 766–67, 767
Golden Rule, 752, 762–65, 765
Steel industry, US, as lobbying for tariffs, 404–405
Stern, Nicholas H., 920–21, 924
Stern Review, 921
Stock market, bubbles in, 66
Strategic behavior, in tax competition, 667–72
Strategic environment, 76
Strategic interaction and public goods, 150–51, 154
tax compliance from, 592–93
Strategic voting, 358, 363
Structural modeling, 73
Subgroup consistent poverty measures, 486
Subscription campaign, 181–82
Subsidies (negative taxes)
on capital income, 624
and externalities, 236, 239
lump-sum, 688
wasteful, 679–80
Substitution, marginal rate of, 29
Substitution effect, 506, 508–509
and income taxes, 541, 544, 545
and lump-sum transfers, 40
Summers, Larry, 704
Supply
and price, 26
and wage increase, 58
Supply–demand balance
in competitive equilibrium, 13
in equilibrium of exchange economy, 17
market equilibrium from, 133
in model construction, 14
and production-and-exchange economy, 26
Surplus, attempts to increase, 255
Sweden
equalization grants in, 689
inequality in, 480–81
poverty line in, 489
public provision of health insurance in, 330
social value of life in, 869
tax rate of, 691
“Swing voter” politics, 377–78
Switching cost, in keyboard adoption, 235
Switching values analysis, 855–56
Switzerland
  effective tax rates in, 711
  equalization grants in, 689
  social value of life in, 869
Symmetry of information, assumption of, 14–15, 42
Tacit collusion, 260, 282, 283
Tariff(s)
  international
    efficiency lost through, 736
    lobbying for, 404
    Smoot–Hawley Act, 736
    and trade agreements, 737, 738–39, 740
    welfare cost of, 733–36
  two-part (domestic)
    for clubs, 197–99
    for regulated firms, 281
Tariffication, 738
Tariff policy, 732–33
  and welfare cost of tariffs, 733–36
Taxation
  ad valorem, 274–77
  on capital, 822, 823, 829, 838
  in international environment, 709–11
    optimal, 821–22, 823
    policy experiments on, 831
  on capital income, 622, 823
  in climate policy, 937–39
  and international tax policy, 938
  club fees as, 208
  commodity, 428, 505 (see also Commodity taxes)
  and consumption growth, 828
  consumption–income distinction in, 817
  as country-specific, 9
  distortions in choice from, 121
  and economic growth, 809, 810
  empirical evidence on, 833–37
  reform of, 829–33
  vs. emission permits, 933–34
  incidence of, 268–74
  income taxes, 89 (see also Income taxes)
  lump-sum, 39–41, 156, 424, 426, 427–30, 506
    (see also Lump-sum taxes)
  mixed-economy model of, 607
  optimal, 413, 428, 509–12, 547–54, 817–23
    (see also Optimal taxation)
  payroll, 102
  and problem of determining marginal rate, 835, 837
  property taxes, 102
    for public goods, 156
    local, 213
  race to the bottom in, 691
  and redistribution, 609, 619–20
    commodity taxes, 527, 528
    income taxes, 561, 609
    OECD study on, 480–81
    redistributive vs. incentive effects, 608
  in rent creation, 408
  rising level of, 809, 833
  and social security, 781–82
  specific, 274–77
  and tax overlap, 676–77
  and tragedy of the commons, 234
  and unions, 285–86
  value-added (VAT), 102, 624, 704, 729–30
  voting on (flat tax), 563–65
Taxation Principle, 614–20, 626, 628
Tax avoidance, 574
Tax competition, 665–66
  and capital mobility, 666–67
  among countries, 709, 741
  from destination tax principle, 732
  efficient, 679–81
  and public input provision, 674–76
  and size, 673–74
  strategic behavior in, 667–72
  and tax exporting, 677–79
  and transfer pricing, 720 (see also Transfer pricing)
Tax compliance game, 590–93
Tax exporting, 677–79
Tax evasion, 573
  auditing for punishment of, 583–86
  in compliance game, 590–93
### Index

- **Tax evasion (cont.)**
  - critical mass of compliance crucial in, 600
  - and decision on evading, 576–83
  - behavioral models of, 594–98
  - social aspects in, 588
  - standard model at odds with empirical evidence on, 593–94
  - evidence on, 586–88, 589
  - extent of, 574–75
  - and honesty, 588–90
  - and social context, 599–600
  - vs. tax avoidance, 574

- **Tax harmonization**, 687

- **Tax havens**, 711

- **Tax mimicking**, 693–94

- **Tax mix**, 609

- **Tax overlap**, 676–77

- **Taxpayer Compliance Measurement Program**, 580

- **Tax principle**, 609

- **Tax Reform Act, US (1986)**, 724

- **Technological (technical) progress**
  - for control of global warming, 940
  - and endogenous growth model, 825
  - and growth, 809, 814
  - and human capital, 824
  - Temptation good, 56

- **Terms of trade**, 726–27, 734–35

- **Terms-of-trade effect**, 679

- **Thaler, Richard H.**, 53, 54

- **Theil entropy measure**, 475

- **Through the Looking Glass**, on use of words, 283

- **Tiebout hypothesis**, 192, 212–14, 216–17, 681, 682
  - empirical tests of, 214–16
  - and fiscal federalism, 639
  - and levels of government, 637

- **Time**
  - and economic activity, 751 (see also Overlapping generations model)
  - value of, 857

- **Time consistency**, 55, 627–28
  - for in-kind transfers, 435–36

- **Tinbergen, Jan**, 731

- **Tobacco advertisements**, limitations on, 279

- **Top cycle**, 358–60

- **Top-level condition**, 38

- **Trade, geography of**, 191

- **Trade agreements**, 736–41

- **Traffic jams**, 229–30

- **Tragedy of the commons**, 233–34

- **Transaction costs, and local public goods**, 639

- **Transactions costs hypothesis**, 214

- **Transfer pricing** (between divisions of firm), 703–704, 713–14, 717–20

- **Transfers**
  - in-kind, 435–36, 627
  - interjurisdictional
    - interregional, 647–48
    - between levels of government, 635
  - lump-sum, 39–41, 628 (see also Lump-sum taxes)
  - vs. Nash equilibrium, 791–92
  - and poverty, 487
  - and redistribution (OECD study), 480–81
  - social security as, 776
  - in “swing voter” politics, 377

- **Transformation, marginal rate of**, 29, 155, 195

- **Transition matrices**, 493

- **Transitivity requirement**, 349

- **and quasi-transitivity**, 442

- **Transparency**
  - in control of rent-seeking, 413–14
  - and tax compliance, 598

- **Travel costs**, 191

- **Turkey**, tax revenues in, 100, 102

- **Two Theorems of Welfare Economics**, 33, 884.
  - See also First Theorem of Welfare Economics; Second Theorem of Welfare Economics
  - and prices, 255
  - and production, 37–39
  - and public goods, 147
  - and Tiebout hypothesis, 212

- **Ultimatum game**, 76–77

- **Uncertainty**
  - about climate change, 895, 897
  - and policy, 899–900, 907, 922, 937–38
  - and price control, 939
Uncertainty (cont.)
and income measurement, 457, 458, 459
and social security, 775
in valuing of future, 870, 874
Uncovered set, 360
Undershifting, tax, 271, 529
Unions
with monopoly power, 255, 285–86
tax incidence with, 286
Unit of account, 20
Unitary country, 94
United Kingdom
climate change effects in, 912
climate policies in, 929–31
composition of general spending in, 95–97, 98
in Concorde project, 940
corporate tax rate in, 829
cost–benefit analysis in, 847
emissions level of, 936, 937
Family Expenditure Survey in, 575
and financial crisis, 109
and fully funded pensions, 797
Green Book of, 873–75
income tax in, 609
introduction of (1799), 537
levels of government in, 635
median-voter age increase in, 801
minimum-needs calculation in, 461
“mis-selling” of pension plans in, 802
National Insurance in, 649
pay-as-you-go social security in, 777
Poll Tax of, 428
poverty line in, 483
public ownership in, 281
public provision of health insurance in, 330
public spending in, 90–94
regulatory schemes in, 281
and residency of company headquarters, 714
and social-security transition model, 799–800
social value of life in, 869
taxation-growth relation in, 834
tax evasion penalties in, 578
tax rates of, 691
effects of, 545
tax revenues in, 99, 100, 102, 103, 106
social mobility in, 492
state pension value in, 780, 781
tie-breaking in, 366
as unitary country, 94
window and hearth taxes in, 505
United Nations Framework Convention on
Climate Change, 935
United States
agricultural subsidies in, 740
climate change effects in, 911, 911–12
composition of general spending in, 95–97, 98, 99, 100
and Copenhagen Accord, 898, 935
corporate tax rate in, 829
cost–benefit analysis in, 847
cotton subsidy in, 740
emissions level of, 935, 936, 937
equalization grants in, 689
as federal country, 94
and financial crisis policy, 109
and global financial crisis, 107–108, 109
“flypaper effect” in, 690
GATT not ratified by, 737
gross revenue minus investment in, 769
income tax in, 89, 609
on capital and on labor, 627
inequality in, 480–81
and Kyoto Protocol, 898
levels of government in, 635
mean income in, 868
median-voter age increase in, 801
pay-as-you-go social security in, 777
poverty line in, 489
public provision of health insurance in, 330
public spending in, 90–94
social mobility in, 492
social security program of, 780
reform of, 797
social value of life in, 869
tax evasion extent in, 580
tax evasion penalties in, 580
tax rates of, 691
Tax Reform Act (1986) in, 724
tax revenues in, 100, 102, 103, 104
and VAT, 704
welfare expenditures division, in, 688
United States Consumer Expenditure Survey, 53
Unmeasured economy, 574
UN System of National Accounts, 574
Index

Uruguay, democracy and social security in, 798
Uruguay Round of trade talks, 739
Utilitarian philosophy, 440–41, 443
Utilitarian social welfare function, 425–26, 479, 538
and weighted welfare function, 557–58
Utilities, as natural monopolies, 280
Utility cardinal, 444–45, 448
measurability of, 443
measurement of rejected, 436–37
ordinal, 443, 444
Utility function, 16
aggregation of, 442
and altruism, 791
cardinal and ordinal, 444–45
and climate policy, 924, 926–28
and intertemporal preferences, 919
and clubs, 196, 204
Cobb–Douglas, 561
and comparability, 445
concavity of, 478
for conformism, 62
for consumer preferences, 754, 767, 822–23, 827
and distributive weights, 882
in endogenous growth model, 831
in exogenous growth model, 20
expected (and tax evasion), 577, 579, 593, 594–94
and externalities, 224–25
for honesty, 589
household, 464–65, 466, 478, 479
and equivalence scales, 496
individual, 442
under centralization and decentralization, 642
and insurance, 308
for preferences, 324
as nonseparable, 623
and public goods, 150, 151, 173
local, 209, 210
quasi-linear, 555–58, 563, 651, 859
and risk aversion, 582
as separable, 622
and taxation, 524, 528, 540
and labor supply, 539
transformation of, 443
for union, 285
and voting, 374, 378, 563
Utility possibility frontier, 425
Value-added taxation (VAT), 102, 624, 704, 729–30 (see also Consumption taxes)
Value of statistical life (VSL), 866–69
Value of statistical life year (VOLY), 866–67
Vertical product differentiation, 257
Volvo drivers, careless driving of, 301–302
Voting, 345
and “Alabama” paradox, 371
and allocation, 424
alternatives to majority rule for, 361
approval voting, 363–64
Borda voting, 361–62
plurality voting, 363
runoff voting, 364–65
over flat tax, 563–65
and Impossibility Theorem, 347–50, 381
item-by-item, 357
majority rule in, 350
agenda manipulation in, 358–61
and Condorcet winner, 351
and decentralized, 641, 642–43
and May’s theorem, 350–51
and Median Voter Theorems, 157–59, 351–57, 564–65
multidimensional, 356–58
paradox of, 365–70
and political competition, 372–80
and politicians’ accountability, 644
for public goods, 156–59
and social security, 798
stability in, 345–46

Wagner, Adolph, 124
Wagner’s law, 124–25, 837
Walrasian auctioneer, 42
Walrasian equilibrium, 79
Walras’s law, 21
Warm glow effect, 177
Wartime, expenditure increase in, 129
Wasteful subsidies, 679–80
Weitzman rule, 938
Welfare. See Social welfare
Welfare analysis, 71
behavioral, 71–72
choice-based analysis, 72–73
and global warming, 74–76
bounds on, 72–73
Welfare economics, revolution in, 80
Welfare economics, two theorems of. See Two
Theorems of Welfare Economics
Welfare loss, from monopoly, 264–67
Welfare shopping, 686
Welfarism, 51–52, 491, 629
Willingness to accept (WTA) project, 880–81,
881, 882
Willingness to pay (WTP) for project, 880–81,
881, 882
Will power, limited, 51
Window tax, 505
World Bank, and cost–benefit analysis, 847
World trade, increase in, 703 (see also Free
trade; Globalism)
World Trade Organization, 737, 740

Zero-profit equilibrium, 30–31